EMERGING TECH CONFERENCE 2023
Edge Intelligence

Proceedings & Highlights of the 2nd Emerging Tech Conference Edge Intelligence (ETCEI 2023)[1]

HETiA
EXPANDING THE DIGITAL FRONTIERS

ARISTOTLE UNIVERSITY OF THESSALONIKI

Volume 02 | 2023

DOI: 10.63438/ELYQ4188

[1] https://conference.hetia.org/emerging-tech-conference-edge-intelligence-2023/

Publication Information

Published by
Hellenic Emerging Technologies Industry Association (HETiA)

Volume DOI: 10.63438/ELYQ4188

This volume has been assigned a Digital Object Identifier (DOI) to ensure persistent identification and long-term accessibility.

Individual full papers associated with ETCEI 2023 have been assigned separate Digital Object Identifiers (DOIs) to support persistent citation and academic indexing. Poster abstracts are published without individual DOIs.

Preface

This volume of the Emerging Tech Conference – Edge Intelligence 2023 (ETCEI 2023) Proceedings includes **28 peer-reviewed full papers and 17 accepted posters** presented at the 2nd Emerging Tech Conference – Edge Intelligence, held on October 19–20, 2023, in Thessaloniki, Greece, at the Research Dissemination Center (KEDEA). The conference was co-organized by the Hellenic Emerging Technologies Industry Association (HETiA) and the Aristotle University of Thessaloniki. The volume has been assigned a Digital Object Identifier (DOI: 10.63438/ELYQ4188), while individual full papers have been assigned separate DOIs to ensure persistent identification and citation.

ETCEI 2023 was organized under the scientific leadership of the Conference Co-Chairs, Prof. Spyridon Nikolaidis (Aristotle University of Thessaloniki) and Prof. Dimitris Mourtzis (University of Patras), who coordinated a program aimed at strengthening the interface between academic research, technological innovation and industrial application.

The conference took place in a context shaped by rapid advances in artificial intelligence, increasing adoption of Internet-of-Things (IoT) systems and growing demand for intelligent, connected and energy-efficient digital infrastructures.

All contributions underwent peer review by the Technical Program Committee, ensuring the scientific quality and relevance of the presented work, while poster presentations supported early-stage research dissemination and interdisciplinary exchange.

The scientific program covered a broad range of topics, including artificial intelligence and machine learning, IoT systems, edge computing architectures, hardware and circuit design, embedded and cyber-physical systems, energy systems and storage technologies, and digital transformation approaches.

By facilitating interaction among researchers, industry representatives, startups and policymakers, ETCEI 2023 reinforced its role as a platform for knowledge exchange and innovation in emerging technologies.

The Editors acknowledge the Organizing Committee of ETCEI 2023, the Local Organizing Committee of the Aristotle University of Thessaloniki, and the reviewers for their contribution to the successful realization of the conference. Special thanks are extended to Konstantina Tsiapali for her coordination and contribution to the organization of the conference, as well as to the volunteers and supporting staff.

October 19-20, 2023
Thessaloniki

Edge Intelligence 2023 Proceedings

On behalf of HETiA
Dr. Emmanouil Zervakis

Scientific Editor
Prof. Spyridon Nikolaidis

ETCEI 2023:
A Flagship Industry–Academia Platform on Emerging Technologies

The 2nd Emerging Tech Conference – Edge Intelligence (ETCEI 2023) was held on October 19–20, 2023, in Thessaloniki, Greece, at the Research Dissemination Center (KEDEA). The conference was co-organized by the Hellenic Emerging Technologies Industry Association (HETiA) and the Aristotle University of Thessaloniki.

ETCEI 2023 brought together researchers, industry representatives, technology developers, startups and policymakers to present and discuss developments in emerging technologies with emphasis on Edge Intelligence, artificial intelligence (AI), Internet of Things (IoT) and advanced electronic systems. The program was structured to promote interaction between academia and industry through scientific presentations, keynote speeches, special sessions and applied technology discussions.

The conference included scientific presentations, keynote addresses, invited talks, technical sessions and industry-oriented discussions. Topics addressed encompassed smart and AI-driven applications, Internet-of-Things (IoT) systems, edge computing architectures, hardware and circuit design, embedded and cyber-physical systems, energy systems and storage technologies, as well as digital transformation frameworks and industrial innovation approaches.

Aims & Objectives

ETCEI 2023 aimed to foster structured interaction between academia, research institutions and industry through peer-reviewed research contributions and validated technological developments in emerging digital technologies and intelligent systems.

Particular attention was given to technological maturity, system-level integration, and alignment with contemporary European and international developments in artificial intelligence, Internet-of-Things (IoT), embedded systems and advanced digital infrastructures.

The conference agenda included:

- √ Keynote speeches by invited experts
- √ Presentations of peer-reviewed full papers and accepted posters
- √ Thematic technical sessions and panel discussions
- √ Industry-oriented presentations and discussions
- √ Networking opportunities supporting research–industry collaboration

Day 1: Keynote Addresses, Technical Sessions and Industry Dialogue

The first day of the 2nd Emerging Tech Conference (ETCEI 2023) brought together academic researchers, industry representatives and public sector stakeholders, establishing the thematic focus of the conference on Artificial Intelligence, smart systems and emerging digital technologies.

The opening session included a conference overview delivered by Prof. Spyridon Nikolaidis and welcome addresses by representatives of HETiA and the co-organizing institution.

Figure 1. Dr. Manolis Zervakis (President of HETiA), Athanasios Kyriazis (Secretary General for Research and Innovation), Prof. Eustratios Stylianidis (Vice Rector for Research and Development, Aristotle Univ. of Thessaloniki), and Prof. Spyridon Nikolaidis (Conference Chair) during the opening session of ETCEI 2023.

Keynote Address

The day commenced with the keynote speech entitled: **"Trusted Value Chains and Digital Business Ecosystems from CHIPS to Edge"**, delivered by Tom Katsioulas (Member of the IoT Advisory Board for the U.S. Department of Commerce and President & CEO of Archon Design Solutions, Inc.).

The keynote addressed the development of trusted digital value chains and the integration of semiconductor technologies with edge computing and digital business ecosystems, highlighting emerging architectures and ecosystem-level challenges.

Figure 2. Tom Katsioulas delivering the keynote address.

Special Session – RENESAS

A special session was delivered by Renesas, platinum sponsor of the conference, entitled: "**The Powerful Convergence of AI & IoT**".

The session included:

- √ A keynote presentation by **Kaushal Vora** (Sr. Director of Business Acceleration & Ecosystem, Renesas), focusing on AI–IoT ecosystem development and business acceleration strategies
- √ A technical presentation by **Georgios Flamis** (Sr Staff Engineer, AIoT, Renesas), entitled: "**AI IoT applications are here!**", presenting real-world AIoT applications and implementation perspectives

Figure 3. Kaushal Vora and Georgios Flamis during the RENESAS special session at ETCEI 2023.

Technical Sessions and Poster Presentations

A series of technical sessions (S1.1–S1.4) were conducted throughout the day, covering:

- √ Smart Applications
- √ Novel Applications
- √ Hardware Design
- √ AI – Edge AI and Machine Learning

The sessions included peer-reviewed paper presentations addressing topics such as artificial intelligence and machine learning applications, IoT systems, embedded and hardware design, energy-aware systems and data-driven digital infrastructures.

Figure 4. Paper presentations during the technical sessions.

A parallel poster session provided an interactive platform for the presentation of emerging research results and early-stage developments.

Figure 5. Poster Presentation of ETCEI 2023.

Keynote Address – Digital Twins

In the afternoon, a keynote speech entitled:
"Cognitive Digital Twins for Sustainable Value Chains", was delivered by Dimitrios Kyritsis (EPFL & University of Oslo).

The presentation addressed the role of cognitive digital twins in supporting sustainable value chains, focusing on data integration, system-level optimization and intelligent industrial processes.

Figure 6. Dimitrios Kyritsis delivering the keynote address.

Networking Reception

The first day concluded with a networking cocktail held at the Thessaloniki Concert Hall, offering participants the opportunity to engage in informal discussions, exchange perspectives and strengthen connections between academia, industry and the public sector.

Day 2: Advanced Systems, Industry Engagement and Career Development

The second day of ETCEI 2023 focused on advanced electronic systems, semiconductor technologies, industrial applications and industry–academia interaction, with particular emphasis on career development and innovation-oriented activities.

Technical Session – Circuits and Systems

The day commenced with a technical session (S1.5 – Circuits and Systems), including peer-reviewed paper presentations addressing topics such as analog and digital circuit design, model order reduction techniques, electromagnetic analysis, hardware accelerators and reliability in electronic systems.

The session reflected current research activity in semiconductor technologies, circuit design and advanced electronic system architectures.

Figure 7. Paper presentations during Session S1.5 – Circuits and Systems.

Keynote Address – Automotive Radar Technologies

A keynote speech entitled: **"The pathway to higher resolution in mm-wave automotive radar"**, was delivered by Prof. Kostas Doris (NXP Semiconductors, The Netherlands).

The presentation addressed technological advancements in automotive radar systems, focusing on high-resolution sensing, mm-wave design challenges and next-generation semiconductor solutions for automotive applications.

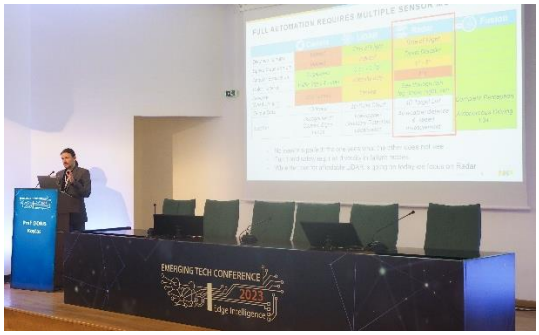


Figure 8. Prof. Kostas Doris delivering the keynote address.

Career Day and Industry Presentations

A dedicated Career Day session was organized, bringing together representatives from academia and industry, facilitating interaction between students, early-stage researchers and technology-oriented enterprises.

The session included an invited talk by **Katerina Papakota** (Career Counselor, Career Office of Aristotle University of Thessaloniki), entitled: **"The Career Office of Aristotle University of Thessaloniki"**, presenting available support mechanisms and career development opportunities for students and graduates.

This was followed by a series of presentations from industry representatives, showcasing technological capabilities and innovation activities, innovation perspectives and career

opportunities across sectors including semiconductors, embedded systems, energy technologies and digital transformation.

Figure 9. Career & Industry Session at ETCEI 2023.

Keynote Address – Digital Transformation and Industry 4.0

A keynote speech entitled: **"Industry 4.0: Unlocking the Future of Digital Transformation"**, was delivered by Vasileios Chatzikos (CEO, Siemens A.E., Greece).

The presentation addressed the role of digital technologies in industrial transformation, focusing on automation, smart infrastructures and the integration of advanced digital systems in industrial environments.

Figure 10. Vasileios Chatzikos delivering the keynote address.

Invited Talks – Innovation and Energy Transition

The program continued with invited talks addressing innovation and sustainability:

- **Byron Chrysovergis** (Technology Transfer and Innovation Manager, Aristotle University of Thessaloniki) presented: **"Technology Transfer and Innovation at Aristotle University of**

Thessaloniki”, focusing on mechanisms for research commercialization and innovation support

- **Spyros Kiartzis** (HELLENiQ ENERGY Holdings S.A.) delivered a presentation entitled: **“Applying emerging technologies to address the challenges of energy transition”**, addressing the role of digital and emerging technologies in sustainable energy systems

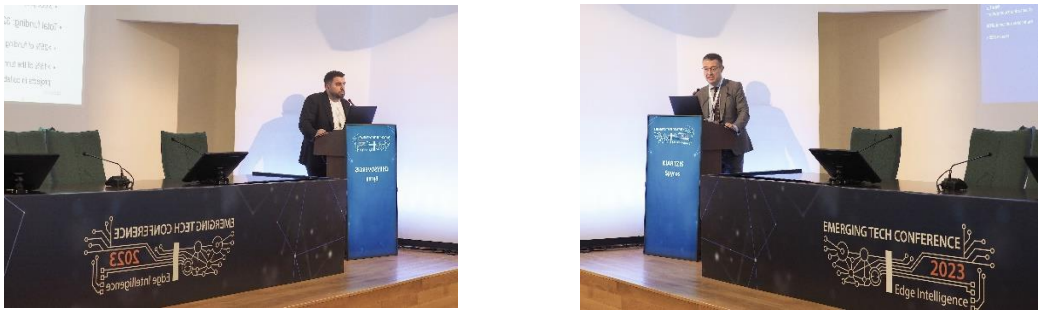


Figure 11. Byron Chrysovergis and Spyros Kiartzis, during their invited talks at ETCEI 2023

Round Table – Industry–Academia Collaboration

A round table discussion entitled: **“Connecting industry to academia”** brought together representatives from industry and academia to discuss collaboration models, knowledge transfer mechanisms and ecosystem development.

The discussion focused on strengthening synergies between research institutions and industry, addressing both challenges and opportunities in bridging innovation and real-world implementation.

Figure 12. Round table discussion at ETCEI 2023

Exhibition Area

Throughout the conference, a dedicated Exhibition Area operated in parallel with the technical sessions, hosting participating companies, technology providers and interactive presentations.

The exhibition space featured industry-oriented activities and provided an opportunity to showcase technological capabilities, ongoing developments and application-oriented solutions related to emerging technologies.

The exhibition area facilitated direct interaction between researchers, students, industry representatives and visitors, supporting knowledge exchange reinforcing the practical and application-oriented dimension of the conference.

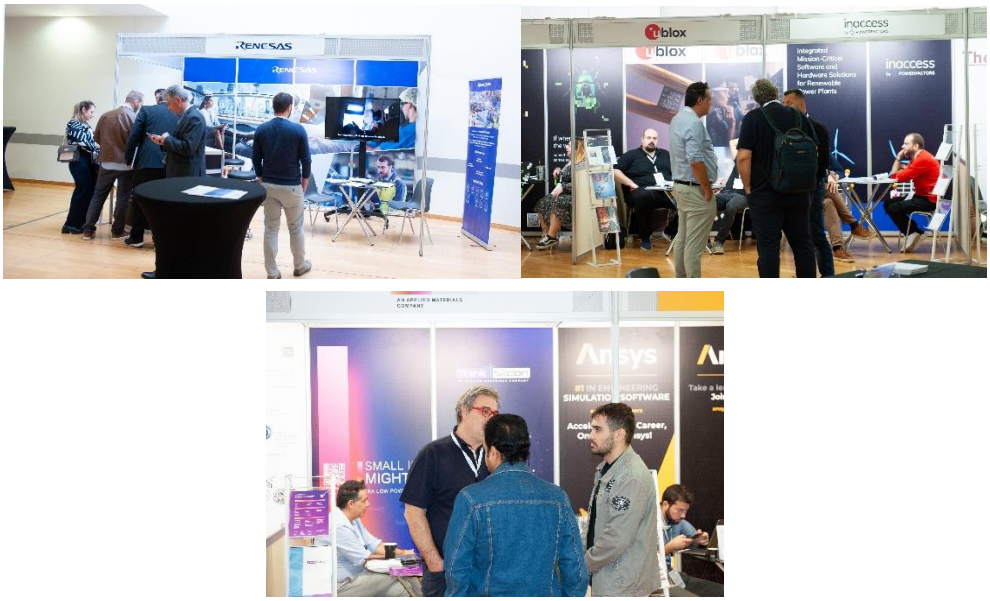


Figure 13. Exhibition Area at ETCEI 2023

Closing and Outlook

The conference concluded with closing remarks by the organizing committee, summarizing the key outcomes of ETCEI 2023 and highlighting the strong participation from academia, industry and the public sector.

The discussions and presentations throughout the two-day event demonstrated the growing importance of emerging technologies in shaping digital transformation, industrial innovation and research collaboration.

The announcement of the forthcoming **ETCEI 2024 (Edge Intelligence)** event marked the continuation of the conference as a structured platform for knowledge exchange, innovation and ecosystem development in emerging technologies.

Figure 14. Closing session of ETCEI 2023.

Conference Co-Chairs

Spyridon NIKOLAIDIS	Professor, Department of Physics Aristotle University of Thessaloniki
Dimitris MOURTZIS	Professor, Department of Mechanical Engineering and Aeronautics, University of Patras

Program Committee

Christos ANDROULIDAKIS	Member of HETiA Board of Directors, Co-Founder at Alma Technologies
Thrasos AXIOTIS	Co-Founder and Director of Engineering – Thess IC
Gregory DOUMENIS	Assistant Professor, Autonomous Systems Laboratory, Department of Informatics & Telecommunications, University of Ioannina
Vassilios KOSTOPOULOS	Professor, President of the Department of Mechanical Engineering and Aeronautics, University of Patras
Alkiviadis HATZOPOULOS	Professor, School of Electrical & Computer Engineering, Aristotle University of Thessaloniki
John KIKIDIS	Member of HETiA Board of Directors, Program Manager at Renesas Electronics Corp
Lena MARAVELIA	Investment Project Manager, Olympia Electronics S.A
Dimitris MOURTZIS	Professor, Vice President of the Committee for Research, University of Patras
Tilemachos MATIAKIS	Co-Founder and Technical Director at Kenotom – Embedded Engineering Excellence
Spyridon NIKOLAIDIS	Professor, School of Physics, Aristotle University of Thessaloniki
Thomas NOULIS	Assistant Professor, School of Physics, Aristotle University of Thessaloniki
Thomas NOULIS	Assistant Professor, School of Physics, Aristotle University of Thessaloniki
Stylianos SISKOS	Professor, School of Physics, Aristotle University of Thessaloniki
Lena KOUGIOUMTZIDOU	Researcher, PhD, University of Thessaly
Konstantina TSIAPALI	Member of HETiA Board of Directors, Managing Director at Global Digital Technologies

Scientific Committee

Bakalis Dimitris	Professor, Department of Physics University of Patras
BIRBAS Michael	Professor, Department of Electrical and Computer Engineering, University of Patras

Achilles BOURSIANIS	EDIP, Department of Physics Aristotle University of Thessaloniki
BUCHER Matthias	Professor, School of Electrical and Computer Engineering, Technical University of Crete
Minas DASYGENIS	Professor, Department of Electrical and Computer Engineering University of Western Macedonia
Giorgos DIMITRAKOPOULOS	Professor, Department of Electrical and Computer Engineering Democritus University of Thrace
Georgios DIMITRIOU	Professor, Department of Computer Science University of Thessaly
Maria DRAKAKI	Professor, Department of Science and Technology International Hellenic University
Nestor EVMORFOPOULOS	Conference Co-Chair, Associate Professor, Electrical and Computer Engineering Department, University of Thessaly
Vasiliki GOGOLOU	Department of Physics Aristotle University of Thessaloniki
Sotirios GOUDOS	Professor, Department of Physics Aristotle University of Thessaloniki
Alkiviadis HATZOPOULOS	Professor, School of Electrical and Computer Engineering Aristotle University of Thessaloniki
Constantinos HILAS	Professor, Department of Computer, Informatics, and Telecommunications Engineering International Hellenic University
Athanasios KAKAROUNTAS	Professor, Department of Computer Science and Biomedical Informatics University of Thessaly
Vasileios KALENTERIDIS	PhD, Senior Analog IC Designer, Thess IC
John KALOMIROS	Professor, Department of Computer, Informatics, and Telecommunications Engineering International Hellenic University
Michail KIZIROGLOU	Professor, Department of Electrical and Electronic Engineering International Hellenic University
Nikos KONOFAOS	Professor, School of Informatics Aristotle University of Thessaloniki
Vasileios KONSTANTAKOS	Professor, Department of Physics Aristotle University of Thessaloniki
Stavros KOULOURIDIS	Professor, Department of Electrical and Computer Engineering University of Patras
Georgios KOUSIOPOULOS	Department of Physics Aristotle University of Thessaloniki
Eftichios KOUTROULIS	Professor, School of Electrical and Computer Engineering Technical University of Crete
Dimitris MOURTZIS	Professor, Department of Mechanical Engineering and Aeronautics, University of Patras
Spyridon NIKOLAIDIS	Professor, Department of Physics Aristotle University of Thessaloniki

HECTOR E. Nistazakis	Professor, Department of Physics National and Kapodistrian University of Athens
THOMAS Noulis	Professor, Department of Physics Aristotle University of Thessaloniki
Maria PAPADOPOULOU	Professor, Department of Information and Electronic Engineering International Hellenic University
Ioannis PAPAEFSTATHIOU	Professor, School of Electrical and Computer Engineering Aristotle University of Thessaloniki
Dimitrios PAPAKOSTAS	Professor, Department of Information and Electronic Engineering International Hellenic University
Vasileios PAVLIDIS	Professor, School of Electrical and Computer Engineering Aristotle University of Thessaloniki
Nikos PETRELIS	Professor, Department of Electrical and Computer Engineering University of Peloponeese
Dionysios REISIS	Professor, Department of Physics – Section of Electronic Physics and Systems, National & Kapodistrian University of Athens
Theodoros SAMARAS	Professor, Department of Physics Aristotle University of Thessaloniki
Georgios SIRAKOULIS	Ch. Professor, Department of Electrical and Computer Engineering Democritus University of Thrace
Stilianos SISKOS	Professor, Department of Physics, Aristotle University of Thessaloniki
Nicolas SKLAVOS	Professor, Department of Computer Engineering and Informatics University of Patras
Dimitrios SOUDRIS	Professor, School of Electrical and Computer Engineering National Technical University of Athens
Georgios STAMOULIS	Professor, Department of Electrical and Computer Engineering University of Thessaly
Vasileios TENENTES	Professor, Department of Computer Science and Engineering University of Ioannina
George THEODORIDIS	Professor, Department of Electrical and Computer Engineering University of Patras
Yiorgos TSIATOUHAS	Professor, Department of Computer Science and Engineering University of Ioannina
Spyridon VLASSIS	Professor, Department of Physics University of Patras
Christos VOLOS	Professor, Department of Physics Aristotle University of Thessaloniki
Sotirios XYDIS	Professor, School of Electrical and Computer Engineering National Technical University of Athens

Traianos YIOULTSIS	Professor, School of Electrical and Computer Engineering Aristotle University of Thessaloniki
Kyriakos ZOIROS	Professor, Department of Electrical and Computer Engineering Democritus University of Thrace

Sponsors

PLATINUM SPONSOR

RENESAS

GOLD SPONSOR

KENOTOM
EMBEDDED ENGINEERING EXCELLENCE

SILVER SPONSOR

Ansys Deloitte. Thess IC inaccess by POWERFACTORS

Think Silicon AN APPLIED MATERIALS COMPANY ublox

BRONZE SPONSOR

olympia electronics CHIP GLOBE lighting the chip world ES SYSTEMS

ARISTOTLE UNIVERSITY RESEARCH DISSEMINATION CENTER
OCTOBER 19 - 20

ARISTOTLE UNIVERSITY OF THESSALONIKI HETiA EXPANDING THE DIGITAL FRONTIERS

Conference Program

Thursday 19/10/2023 – DAY 1				
09:00 – 09:30	Registrations + Welcome coffee			
09:30 – 10:00	Opening | Main HALL (II)			
	Dr. Manolis ZERVAKIS | President, Hellenic Emerging Technologies Association			
	Athanasios KIRIAZIS | General Secretary for Research and Innovation			
	Prof. Eustratios STILIANIDIS | Vice Rector for Research & Development, University of Thessaloniki			
	Prof. Spyridon NIKOLAIDIS | Conference Overview			
10:00 – 10:45	Keynote Speech | Main HALL (II) **Tom KATSIOULAS** | Member of the IoT Advisory Board for the U.S. Department of Commerce | President and CEO Archon Design Solutions, Inc. Title: **"Trusted Value Chains and Digital Business Ecosystems from CHIPS to Edge"** Moderator: **John KIKIDIS** | Global Ecosystem Lead, Renesas			
10:45 – 11:45	Special Session | Main HALL (II) RENESAS: **"The Powerful Convergence of AI & IoT"**			
	10:45 – 11:30	**Keynote by Kaushal Vora** | Sr. Director of Business Acceleration & Ecosystem | Renesas Q&A		
	10: 30 – 11:45	**Georgios FLAMIS** | Sr Staff Engineer, AIoT | Renesas Title: **"AI IoT applications are here!"** Moderator: **John KIKIDIS** | Global Ecosystem Lead, Renesas		
11:45 – 12:00	Coffee Break			
12:00 - 14:30	Parallel Sessions			
	Main HALL (II)		Foyer	
12:00 – 13:15	Paper Presentations S1.1 | Smart applications | Session Chair: Minas DASYGENIS		Poster Presentations	
12:00 – 12:15	3360	**Autonomous Unmanned Ground Vehicle (UGV) In Smart Farming With A* Algorithm** Antonios Chatzisavvas, Malamati Louta and Minas Dasygenis	1864	**The Smart Fridge Project** Aikaterini Griva, Vasileios Rekkas, Kyriakos Koritsoglou, Sotirios Sotiroudis, Achilles Boursianis, Maria Papadopoulou and Sotirios Goudos
12:15 – 12:30	8374	**Comparing the Performance of Cameras and SW packets for the Cloud Coverage Process in Photovoltaic (PV) Parks** Dionysios Reisis, Christoforos Vasilakis, Georgios Venitourakis, Alexandros Tsagkaropoulos, Panagiotis Golemis, Georgios Konstantoulakis and Panagiotis Tzouma Amrou	2051	**Ground-based cloud observation using wide-view optical and thermal representations.** Dimitrios Tsourounis, Panagiotis Tzoumanikas, Alexios Kotronis, Andreas Kazantzidis, George Economou, Orestis Panagopoulos and Christos Theocharatos
12:30 – 12:45	1160	**Advanced Monitoring System of Industrial Refrigeration (ADMOSIR)** Theodoros Athanasopoulos, George Mavropoulos, Antonios Depastas, Panagiotis Vlagopoulos, Katerina Spyropoulou and Emmanouil Zervakis	2309	**An OTA-based power sensing integrated circuit for MPPT in photovoltaic energy harvesting applications.** Zoi Agorastou, Vasileios Konstantakos, Konstantinos Siozios and Stylianos Siskos
12:45 – 13:00	6373	**Implementation of a Beehive Health Monitoring System Based on Sound** Dimitrios Kampelopoulos, Giannis Sofiannidis, Vasileios Konstantakos, Spyridon Nikolaidis, and Kostas Siozios	2550	**Prediction of Analog Circuit Sizing Using an Artificial Neural Network** Shubham Agarwal, Benjamin Prautsch and Uwe Hatnik

13:00 – 13:15	9471	**An Audio Fingerprinting Approach based on the 2DFT for Byzantine Hymn Recognition** Dimitrios Kampelopoulos, Lazaros Moysis, Konstantinos Karasavvidis, Achilles D. Boursianis, Sotirios K. Goudos and Spyridon Nikolaidis	2735	**Classification of Fetal Images using Deep Learning Methodologies: The Smart Embryo Project** Lazaros Alexios Iliadis, George Vergos, Paraskevi Kritopoulou, Achilleas Papatheodorou, Sotirios Sotiroudis, Achilles Boursianis, Kostas Kokkinidis, Maria Papadopoulou and Sotirios Goudos
13:15 – 14:30	Paper Presentations S1.2 | Novel applications | Session Chair: Nikolaos KONOFAOS		2870	**Estimation of pedestrian trajectory using LSTM network architecture** Christos Theocharatos, Dimitris Kastaniotis and Vassilis Tsagaris
13:15 – 13:30	1139	**IDEAN Neurosciences Education Platform** Aggelos Kostas, Anastasios Vittas, Eirini Georgia Dimitriou, Konstantinos Kalafatakis, Kalliopi Basiakou, Nikolaos Giannakeas, Alexandros Tzallas, Nikolaos Katertsidis and Markos G. Tsipouras	3392	**A deep learning approach for detecting defects in melt pool images in DED-AM process.** Thanassis Polizogopoulos, Christos Theocharatos, Konstantinos Tzimanis, Nikolas Porevopoulos, Panagiotis Stavropoulos, Konstantinos Ntalas and Albano Memlika
13:30 – 13:45	3726	**SmartGlove Cloud Platform for Parkinson's Patients Data Collection & Analysis** Eirini Georgia Dimitriou, Vasiliki Fiska, Konstantinos Kalafatakis, Nikolaos Giannakeas, Alexandros Tzallas, Nikolaos Katertsidis and Markos G. Tsipouras	3584	**Machine learning methods for the discrimination of refrigerant gases** Nikolaos Argirusis, Petros Karvelis, Georgia Sourkouni, John Konstantaras, Alice Baroncelli, Peter Segers and Christos Argirusis
13:45 – 14:00	4929	**Rapid virtual prototyping of battery storage systems** Sotirios Athanasiou and Georgios Koukos	4149	**Design of an Autonomous, Multi-functional Stress Assessment Sensor for Naval Applications: The AMSA project** Gregory Doumenis, Vasiliki Naskari, Evangelos Hristoforou, Polychronis Pattakos, Georgia Stamou, Christos Papakis, Ioannis Masklavanos
14:00– 14:15	9053	**S2W-Med Health Status Monitoring** N. Sionas, G. Pomportsis. P. Christodoulou, I. Tzamtzis	4363	**A unified framework for object and person 3D-pose estimation with application in ancient drama** Andreas Makedonas, Ioannis Papakonstantinopoulos, Panteleimon Alkinoos Peftikoglou and Christos Theocharatos
14:15– 14:30	8723	**Fish Shape Alignment Based on Deformable Shape Tracking Suite of Tools** Nikos Petrelis, Georgios Keramidas, Christos Antonopoulos and Nikolaos Voros	4625	**Deep Learning Architectures for Greek Orthodox Church Hymns Recognition: The Ymnodos Project** Lazaros Alexios Iliadis, Nikolaos Tsakatanis, Sotirios Sotiroudis, Achilles Boursianis, Kostas Kokkinidis, Georgios Patronas, Pavlos Serafeim, Maria Papadopoulou and Sotirios Goudos
			4922	**An IoT System for Innovative Cultural Experience** Christos Sad, Aggeliki Ziaka and Kostas Siozios
			4933	**An Improved Algorithm for Equalization and Energy Support for Lithium-ion Battery Storage Systems in Electric Motor Drive Applications** Nikolaos Jabbour, Evangelos Tsioumas, Dimitrios Papagiannis and Christos Mademlis
			6288	**Design of an Autonomous Wireless Electric Field sensor for maritime applications: the EFOS project** Ioannis Masklavanos, Vasiliki Naskari, Christos Koutsos, Fotios Vartziotis, Gregory Doumenis, Stylianos Siskos, Achilleas Bardakas, Apostolos Segkos, Christos Tsamis, Christos Papakis and George Koukas
			6895	**Area Allocation for Coverage Path Planning Using Affinity Propagation Clustering** Nikolaos Baras, Antonios Chatzisavvas, Irene Tabakis and Minas Dasygenis
			9101	**Crosstalk exploration between PA and LNA inductors** Dimitrios Samaras, Alkis Hatzopoulos, Vasilis Pavlidis, Athanasios Stefanou, Georgios Chararas and Rafaela Themeli
			9465	**Auditory Scene Profile Adaptation for ANC Headphones** Dennis Tsoukalas and Fotios Kontomichos
14:30 – 15:15	Lunch Break			
15:15 – 16:00	Keynote Speech | Main HALL (II) **Dimitrios Kyritsis** | EPFL & University of Oslo Title: **"Cognitive Digital Twins for Sustainable Value Chains"** Moderator: **Asst. Prof. Gregory DOUMENIS** | Informatics & Telecommunications, University of Ioannina			

16:00 – 17:45	Paper Presentations | S1.3 | Hardware design | Session Chair: Gregory DOUMENIS	
16:00 – 16:15	7941	**Design and Implementation of Bone Wax Application Device** Michail Tsilikas and Dimitrios Papakostas
16:15 – 16:30	6794	**Towards a modular IoT device design and prototyping for the Sports domain** Evripidis Chondromatidis, Emmanouil Tsardoulias, Konstantinos Panayiotou and Andreas Symeonidis
16:30 – 16:45	9301	**High-Performance Design of SRT IP Blocks** Aristotelis Tsekouras, Giorgos Stagakis, Anastasis Avgoustidis, Grigoris Kokkonis, Konstantinos Gkekas, Vasilis Pavlidis, Thomas Noulis and Giorgos Keramidas
17:00 – 17:15	1120	**Fractional-N Phase Locked Loop for Wi-Fi 6/6E With 135 mW Power Consumption and Spur Reduction Techniques** Savvas Sgourenas, Christos Andriakopoulos, Stefanos Pokamisas, Charis Basetas, Chrysa Vassou, Vasilis Tsamis, Kostas Retsinas, Vasilis Kolios, Nektarios Sgourenas and Giorgos Kasapoglou
17:15 – 17:30	5815	**Enhanced Safety Architecture with Fault Preventive Mechanisms for Automotive Li-ion Battery Management Systems** Apostolos Delizonas, Christos Mademlis, Evangelos Tsioumas, Dimitrios Papagiannis, Nikolaos Jabbour, Christos Sansaridis and Tilemachos Matiakis
17:30 – 17:45	838	**Energy Efficient ML Accelerator using a Decoupled Vector Engine and a Systolic Array attached to a Low-Cost RISC-V Scalar Core** Giorgos Dimitrakopoulos, Vasileios Titopoulos, Christodoulos Peltekis, Dionysios Filippas and Kosmas Alexandridis
17:45 – 19:00	Paper Presentations | S1.4 | AI - Edge AI and ML | Session Chair: Kostas Siozos	
17:45 – 18:00	8081	**Light Bulb control with DNN based voice user interface – the journey to design it** Georgios Flamis, George Chardalias, Vin D'Agostino and Suad Jusuf
18:00 – 18:15	9736	**Enabling Grid Resilience and Efficiency - EDGE Device for Power Grid analysis and Asset Monitoring** Sami Hammal, Vagelis Alifragkis, Pavlos Psimadas and Nikolaos-Antonios Livanos
18:15 – 18:30	9653	**Federated transformers for non-intrusive load monitoring in heat pumps** Stylianos Kanoutas, Athanasios Bachoumis, Michael Birbas and Alexios Birbas
18:30 – 18:45	5311	**AI-assisted Serious Games: Dialogue Management with Generative AI** Eleni Panopoulou, Davide Aversa and Stavros Vassos
18:45 – 19:00	4607	**Transforming the Path Towards Automation of Monitoring and Management for Edge Computing** Georgios Samaras, Marinela Mertiri, Maria-Evgenia Xezonaki, Nikos Psaromanolakis, Vasileios Theodorou and Theodoros Bozios
19:00 – 19:15	102	**Intelligence Functions Placement in B5G / 6G wireless networks** Vasiliki Lamprousi, Sokratis Barmpounakis, Vera Stavroulaki and Panagiotis Demestichas
20:00 – 21:30	**Networking Cocktail at Allegro | Thessaloniki Concert Hall** Only for Paper Presenters - Sponsors - Speakers	

Friday 20/10/2023 – DAY 2		
09:00 – 09:30	Welcome coffee	
09:30 – 11:00	Paper Presentations | S1.5 | Circuits and Systems | Session Chair: Vasilis PAVLIDIS	
09:30 – 09:45	8700	**Analysis of SQRL strengths and weaknesses compared to other authentication mechanisms** Dimitrios Simeonidis
09:45 – 10:00	1994	**Fault Detection in Analog Circuits by utilizing the Current Supply Transients** Vassilis Vassios, Argirios Hatzopoulos, Dimitrios Papakostas and Ioannis Intzes
10:00 – 10:15	4126	**Aging alleviation technique for 8T IMC SRAMs** Helen-Maria Dounavi and Yiorgos Tsiatouhas
10:15 – 10:30	3841	**Reduction of large-scale RLCk models via low-rank balanced truncation** Christos Giamouzis, Dimitrios Garyfallou, Anastasis Vagenas and Nestor Evmorfopoulos
10:30 – 10:45	7520	**MORCIC: Model Order Reduction Techniques for Electromagnetic Models of Integrated Circuits** Dimitrios Garyfallou, Athanasios Stefanou, Christos Giamouzis, Moschos Antoniadis, Georgios Chararas, Konstantinos Chatzis, Dimitris Samaras, Rafaela Themeli, Anastasios Michailidis, Vasiliki Gogolou, Nikos Zachos, Nestor Evmorfopoulos, Thomas Noulis, Vasilis F. Pavlidis, Alkiviadis Hatzopoulos, Elpida Chatzineofytou and Yiannis Moisiadis
10:45 – 11:00	4978	**Design and implementation of a compact RISC-V based Machine Learning accelerator on Low End FPGA** Manolis Galetakis, Stavros Kalapothas, Georgios Flamis, Paris Kitsos and Fotis Plessas
11:00 – 11:40	Keynote Speech | Main HALL (II) Prof. Kostas DORIS | NXP Semiconductors, The Netherlands Title: *"The pathway to higher resolution in mm-wave automotive radar"* Moderator: **Prof.Stelios SISKOS** | Electronics Lab. Section of Electronics and Computers Physics Dept. | AUTH	
Career Day		
11:40 - 14:00	Companies Presentations	
	11:40 - 11:50	Invited Speaker: **Katerina PAPAKOTA** | Career Counselor, Career Office of AUTH Title: **"The Career Office of Aristotle University of Thessaloniki"**
	11:50 - 14:00	RENESAS | Stella TSAMPOUKOU
		KENOTOM | Stefanos GIANNOPOULOS
		ANSYS | Manolis FOTAKIS & Ioannis GEORGAKIS
		DELOITTE | Pierrina MONASTIRIOTI & Eirini PAPAGIANNIDOU
		INACCESS, by POWER FACTORS | Ioannis SAPOUNADELIS
		THESS IC | Thrassos AXIOTIS
		THINK SILICON, an APPLIED MATERIALS Company | Georgios KERAMIDAS
		CHIPGLOBE | Max MERGANTHALER & Lia ELEFTHERIADOU

		ES SYSTEMS | Theodoros ATHANASOPOULOS
		OLYMPIA ELECTRONICS | Sophia CHATZIFOTI
		U-BLOX | Panos PAGRATIS
		MEICSI | Savvas SGOURENAS
		Moderator: **Prof. Alkis CHATZOPOULOS** | Dept. of Electrical and Computer Engineering | AUTH
14:00 – 14:10	Snack coffee break	
14:10 – 14:50	Keynote Speech | Main HALL (II) **Vasileios CHATZIKOS** | CEO Siemens A.E., Greece Title: **"Industry 4.0: Unlocking the Future of Digital Transformation"** Moderator: **Tilemachos MATIAKIS** | KENOTOM Technical director	
14:50 – 16:00	Invited Speakers	
	14:50 – 15:00	Invited Speaker: **Byron CHRYSOVERGIS** | Technology Transfer and Innovation | Manager at Aristotle University of Thessaloniki Title: **"Technology Transfer and Innovation at Aristotle University of Thessaloniki"**
	15:00 – 15:15	Invited Speaker: **Spyros KIARTZIS** | HELLENiQ ENERGY Holdings S.A. Title: **"Applying emerging technologies to address the challenges of energy transition"** Moderator: **Prof. Spyridon NIKOLAIDIS** | Physics Department, Aristotle University of Thessaloniki
	15:15 – 15:40	**Round table: "Connecting industry to academia"** Participants: **RENESAS** |George FLAMIS, **KENOTOM** |Tilemachos MATIAKIS, **AUTH** | Prof. St. SISKOS), **ACCADEMIA** | Georgios SAVVIDIS, **PROGRESSIVE ROBOTICS** | Marios KIATOS Moderator: **Prof. Alkis CHATZOPOULOS** | Dept. of Electrical and Computer Engineering | AUTH
	15:40 – 16:00	**Closing Day 2 remarks** Prof. Spyridon **NIKOLAIDIS** | Prof. Dimitris **MOURZTIS** **Edge Intelligence 2024 announcement:** Dr. Manolis **ZERVAKIS** | Prof. George **STAMOULIS**
16:00– 17:00	Late Lunch	

Contents

EMERGING TECH CONFERENCE 2023
Edge Intelligence

Papers

Session 1.1 | Smart applications

Session Chairs: Minas DASYGENIS

Autonomous Unmanned Ground Vehicle (UGV) In Smart Farming With A* Algorithm
Antonios Chatzisavvas, Malamati Louta and Minas Dasygenis

Comparing the Performance of Cameras and SW packets for the Cloud Coverage Process in Photovoltaic (PV) Parks
Dionysios Reisis, Christoforos Vasilakis, Georgios Venitourakis, Alexandros Tsagkaropoulos, Panagiotis Golemis, Georgios Konstantoulakis and Panagiotis Tzouma Amrou

Advanced Monitoring System of Industrial Refrigeration (ADMOSIR)
Theodoros Athanasopoulos, George Mavropoulos, Antonios Depastas, Panagiotis Vlagopoulos, Katerina Spyropoulou and Emmanouil Zervakis

Implementation of a Beehive Health Monitoring System Based on Sound
Dimitrios Kampelopoulos, Giannis Sofiannidis, Vasileios Konstantakos, Spyridon Nikolaidis, and Kostas Siozios

An Audio Fingerprinting Approach based on the 2DFT for Byzantine Hymn Recognition
Dimitrios Kampelopoulos, Lazaros Moysis, Konstantinos Karasavvidis, Achilles D. Boursianis, Sotirios K. Goudos and Spyridon Nikolaidis

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Pages 1 – 7

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Autonomous Unmanned Ground Vehicle (UGV) In Smart Farming With A* Algorithm

Antonios Chatzisavvas[1]*, Malamati Louta[1], Markos Tsipouras and Minas Dasygenis[1]

[1] *University of Western Macedonia, Kozani, Greece.*
achatzisavvas@uowm.gr, louta@uowm.gr, mtsipouras@uowm.gr,
mdasyg@ieee.org

Abstract

Smart farming is a popular agricultural concept that uses modern technology to optimize agricultural operations, increase efficiency, and enhance crop yields. Finding the most effective and ideal approach for different agricultural chores such as planting, irrigation, and harvesting using algorithms and embedded systems is one of the core issues in smart farming. In robotics, the A* algorithm is a popular pathfinding technique. To effectively discover the shortest route, the A* is well-suited for tackling the pathfinding issue in smart farming methods. Here, we focus our research on improving the well know A*. Specifically, we investigate how the A* algorithm might be modified to solve particular agricultural difficulties, such as travelling over unevenly shaped fields and avoiding obstacles. The research also looks at selecting appropriate heuristic functions for the A* algorithm, which may provide more efficient routes for various agricultural tasks. This paper demonstrates the effectiveness and benefits of using the proposed A* algorithm in smart farming applications through extensive simulations and real-world case studies. The results show significantly improved overall efficiency compared to the traditional A* algorithm.

1 Introduction

Smart farming, a cutting-edge agricultural concept, harnesses the strength of advanced technologies to revolutionize traditional farming approaches. By optimizing agricultural processes, enhancing efficiency, and elevating crop yields, smart farming promises to address the challenges posed by a growing global population and the necessity for sustainable food production. At the heart of this transformative approach lies the critical task of finding the most efficient and optimal paths for various agricultural activities, including planting, irrigation, and harvesting [1].

The A* algorithm emerges as a robust computer science and robotics solution. Renowned for its effectiveness in pathfinding, the A* algorithm offers a balanced approach that combines the benefits of Dijkstra's algorithm for guaranteeing the shortest path and the heuristic-based optimization of greedy search. As such, the A* algorithm is ideal for tackling the pathfinding complexities prevalent in smart farming scenarios [2-3].

This study incorporates the A* algorithm into smart farming practices and investigates its adaption to

agriculture's particular difficulties. The A* algorithm and autonomous Unmanned Ground Vehicles (UGVs) enhance their capabilities, enabling them to navigate and perform agricultural tasks with remarkable accuracy and speed. The selection of appropriate heuristic functions is critical to the effectiveness of the A* algorithm in smart farming. This research examines the A* algorithm using the Euclidean distance heuristic function [4]. The Euclidean distance heuristic offers a simple but effective metric for estimating the straight-line length from the present place to the target, allowing for more efficient pathfinding in the agricultural environment.

Compared to conventional farming practices, this adaptive strategy enables smart farming systems to function with greater accuracy, resource conservation, and overall agricultural efficiency. Extensive simulations and real-world case studies demonstrate the usefulness of the A* algorithm in smart farming, offering actual proof of its benefits over conventional pathfinding approaches. The findings show that combining the A* algorithm with the Euclidean distance considerably improves overall agricultural efficiency.

The work offers many vital advances to smart farming and pathfinding algorithms:

- The study presents a novel A* algorithm by integrating the Euclidean distance heuristic. This heuristic supplement the original A* method by estimating the straight-line distance between the present position and the destination more efficiently and effectively. The new A* algorithm may make more informed judgements during pathfinding in smart farming situations by using this improved heuristic, resulting in optimised paths for different agricultural operations such as planting, irrigation, and harvesting.
- The results show that the upgraded A* algorithm with the Euclidean distance heuristic beats the standard A* algorithm in terms of performance. The algorithm's performance has substantially improved, allowing it to locate shorter courses and traverse unevenly shaped fields more successfully. Furthermore, the improved algorithm is more resilient in dealing with impediments and constantly changing environmental circumstances, making it well-suited for real-world smart farming applications.

The remaining sections of this work are structured as follows: Section II presents an outline of similar studies. Section III describes the newly suggested A* algorithm with the Euclidean distance heuristic. In Section IV, we give the experimental outcomes. Finally, in Section V, we summarise the study's key findings and discuss the broader impacts of the proposed enhanced A* algorithm in the context of smart farming.

2 Relative Work

The A* approach is able to create the route on a map that has the shortest distance between two places, but before it can do so, it must first traverse around the nodes of the path and choose the way that has the path cost that is the least expensive feasible. Consequently, the method executes a significant number of calculations and necessitates a considerable amount of time spent doing computations. In addition, the algorithm's effectiveness will diminish as the size of the map rises [5]. Moreover, the approach requires the availability of a significant quantity of random-access memory (RAM). The A* algorithm is a graph search algorithm that employs rasterized maps as a form of map representation. Consequently, its route could be smoother and is dependent on an excessive number of orthogonal twists, which in turn produces a decline in its level of dependability [6].

The development of the A* algorithm has been the issue of expansive research over the years. Researchers have primarily focused on enhancing its capabilities, particularly in dealing with obstacles, optimizing various implementation scenarios, and improving its overall efficiency. As a result, a wealth of knowledge and information has been accumulated regarding the A* algorithm, making it a fundamental tool in the field of route planning and graph traversal. A* is a widely used approach in route planning, particularly in scenarios where finding the shortest path is crucial. It's often likened to Dijkstra's algorithm, and in some ways, it can be viewed as a modified version of Dijkstra's. This similarity stems from the fact that both algorithms aim to discover the optimal path from a source to a terminus [7]. However, A* introduces a key feature that sets it apart: using a heuristic function to guide its search. The heuristic function in A* directs the algorithm towards states with the shortest path, enabling A* to discover the shortest achievable route efficiently. This heuristic-driven approach significantly boosts A*'s efficiency when compared to Dijkstra's algorithm, as it minimizes the exploration of unnecessary states, leading to faster pathfinding. As the A* algorithm has evolved, it has found applications in various domains, including planning routes for unmanned surface vehicles (USVs), robot route planning and automated guided vehicles (AGVs). Its adaptability and efficiency have made it a cornerstone in solving complex pathfinding problems in diverse fields. This research and development have enriched the algorithm's capabilities and expanded its applicability in agricultural scenarios [8].

The A* algorithm has proven highly versatile and widely applied in various sectors, including the transportation industry [9]. One of the main reasons behind its popularity is that it is more straightforward to comprehend than other strategies for route planning, and it uses a smaller number of search nodes. So, the A* algorithm has a reduction in search nodes, resulting in less computational overhead, making the algorithm well-suited for environments with resource constraints [6].

In contrast to similar works by other writers, this paper makes a significant advancement in the field of pathfinding algorithms by presenting an improved version of the A* algorithm that leverages the Euclidean distance heuristic to optimize the average path length. While previous studies have explored various heuristic functions to enhance the A* algorithm's performance, our work focuses explicitly on incorporating the Euclidean distance as the heuristic measure, which has proven to be an effective and efficient choice in smart farming scenarios [10]. The demonstrated improvements in path quality and time efficiency have the potential to impact a wide range of industries, fostering more intelligent and optimized decision-making processes.

3 Implementation

The A* algorithm is a widely employed pathfinding technique that efficiently discovers the quickest path between two points on a grid. It is especially well-suited for scenarios where finding the optimal path is essential, such as robotics and smart farming. The A* algorithm guarantees the quickest path while using heuristics to optimize the examination procedure. The A* algorithm combines the principles of Dijkstra's algorithm (for finding the shortest path) and greedy search (for heuristic-based optimization). It uses a priority queue to explore nodes in order of their calculated cost from the start to the goal route. The formula for calculating the estimated cost for each node, known as the "f-score" is as follows:

$$f(a) = g(a) + h(a)$$

where *f(a)* is the calculated entire cost of the path from the starting route to the goal route passing via route a, *g(a)* is the path's true cost (distance) from the starting to route a, and *h(a)* is the heuristic function's estimated cost from route a to the destination route.

The A* algorithm iteratively evaluates nodes in the priority queue based on their f-scores, selecting the node with the lowest f-score for exploration. The algorithm continues this procedure until the destination node is reached or the priority queue becomes empty, meaning no path exists from the start to the destination node. During exploration, the A* algorithm updates the g-score and f-score for each node it encounters based on the paths it evaluates. If a shorter path to a node is discovered, the algorithm replaces the node's g-score and updates its parent to improve the path's quality.

Integrating the Euclidean distance heuristic into the A* algorithm makes it even more effective in navigating complex conditions. The Euclidean distance heuristic is a method that estimates the straight line distance (also known as the Euclidean distance) between two points, which serves as a heuristic for the algorithm. The heuristic provides an optimistic estimate of the distance to the goal from any given node, assuming no obstacles or restrictions in the path. The Euclidean distance formula estimates the straight-line length between two points in a two-dimensional Cartesian coordinate system (b, c). The Euclidean distance formula between two points, P1(b1, c1) and P2(b2, c2), is given by:

$$\text{Euclidean Distance} = \sqrt{(b_2 - b_1)^2 + (c_2 - c_1)^2}$$

In this formula, (b_1, c_1) and $((b_2, c_2)$ are the coordinates of the two points we want to calculate the distance, $(b_2 - b_1)$ represents the horizontal distance (Δx) between the two points, $((y_2 - y_1))$ represents the vertical distance (Δy) between the two points and $[(b_2 - b_1)^2$ and $(c_2 - c_1)^2$ represent the squares of the horizontal and vertical distances, respectively. The sum of the squares $[(b_2 - b_1)^2 + (c_2 - c_1)^2]$ calculates the square of the straight-line distance between the two points. Taking the square root of the sum gives us the Euclidean distance, which represents the actual straight-line distance between the two points. The Euclidean distance formula can be extended to higher dimensions for calculating distances in three dimensional or multi-dimensional spaces. In pathfinding algorithms like the A* algorithm, the Euclidean distance heuristic estimates the shortest distance between two points, guiding the algorithm towards the goal efficiently.

The A* algorithm with the Euclidean distance heuristic is optimal (guaranteed to find the shortest path). The Euclidean distance is a heuristic that never overvalues the actual cost to reach the terminus. This attribute is instrumental in the A* algorithm, where the heuristic function directs the examination by evaluating the remaining cost to gain the goal from the current node to make decisions. Theoretical benefits associated with the utilization of the Euclidean distance heuristic are substantial. Notably, it enhances the convergence of the A* algorithm towards the solution of optimal. By accurately approximating the distance to the goal, the Euclidean distance heuristic encourages A* to focus on exploring paths that are more likely to lead to the shortest route. Consequently, A* tends to traverse fewer unnecessary nodes and paths during its search, ultimately resulting in faster and more efficient pathfinding. Its ability to incorporate real-world knowledge through the heuristic function makes it a powerful tool for pathfinding in smart farming applications, where navigating irregularly shaped fields and optimizing resource usage are critical requirements.

4 Experimental Findings

We utilized the VELOS UGV for our experimental studies to evaluate our research objectives. The VELOS UGV is a mobile robot designed explicitly for autonomous navigation in agricultural fields, capable of collecting images for disease diagnosis and monitoring purposes. To conduct our pathfinding experiments, we employed the A* algorithm as the basis for 5 different path-planning scenarios, where each experiment's start and end points remained consistent. The A* algorithm is a well-known and commonly used pathfinding technique that guarantees the shortest route. We presented an improved version of the A* algorithm using the Euclidean distance heuristic as the heuristic function.

On an Intel NUC i5 with 16GB of RAM, we implemented the standard A* and A* algorithms with the Euclidean distance heuristic. The algorithms were written in Python, a flexible and widely used programming language. We used critical measures such as path length and execution time to compare and assess the efficiency of the two methods. Table I summarizes the experimental findings, which include execution durations and route lengths. In order to provide that our comparison with the A* algorithm is accurate, we do each experiment using the exact identical implementation details, route planning scenario, and equipment. This level of consistency in the experimental setup is instrumental for achieving a fair and meaningful evaluation of their improved algorithm against the standard A* algorithm.

Path planning scenario	**Standard A* algorithm**		**A* algorithm with the Euclidean distance**	
	Path Length (m)	**Time (s)**	**Path Length (m)**	**Time (s)**
1	17.47	157.71	16.02	143.42
2	16.26	149.44	15.14	137.15
3	14.18	136.17	12.49	124.31
4	12.31	127.46	11.04	116.43
5	11.58	118.52	10.13	109.54

Table I: Standard A* and A* algorithm experimental findings with the Euclidean distance.

Table I indicates that the A* algorithm with the Euclidean distance heuristic consistently provided shorter pathways than the Standard A* method in each trial. The route length is an important parameter since it represents the distance the UGV travels while traversing from the start to the destination point. The execution time measurements in the Time (s) column reflect the computational efficiency of each method. In all five experiments, the A* algorithm with the Euclidean distance heuristic outperformed the Standard A* algorithm, exhibiting shorter execution times. The A* algorithm with the Euclidean distance provides shorter pathways in less time, making it a better alternative for path planning in agricultural areas and other applications needing efficient route optimization. The A* algorithm with the Euclidean distance in all five scenarios outperforms the Standard A* algorithm with quicker execution times. This implies that the improved A* method with the heuristic function may discover the best route more rapidly and efficiently, which is especially useful for real-time applications and resource-constrained systems. The performance comparison of the two algorithms in Table I demonstrates the benefits of introducing the Euclidean distance

heuristic into the A* algorithm. The A* method with the Euclidean distance creates shorter pathways with less computing time, making it a preferable alternative for path planning in agricultural areas and other applications needing efficient route optimization. Finally, the length of each route and the amount of time it takes are shown in Figure 1 (a) and Figure 2 (b) for the five different path-planning situations using the standard and improved version of the A* algorithm.

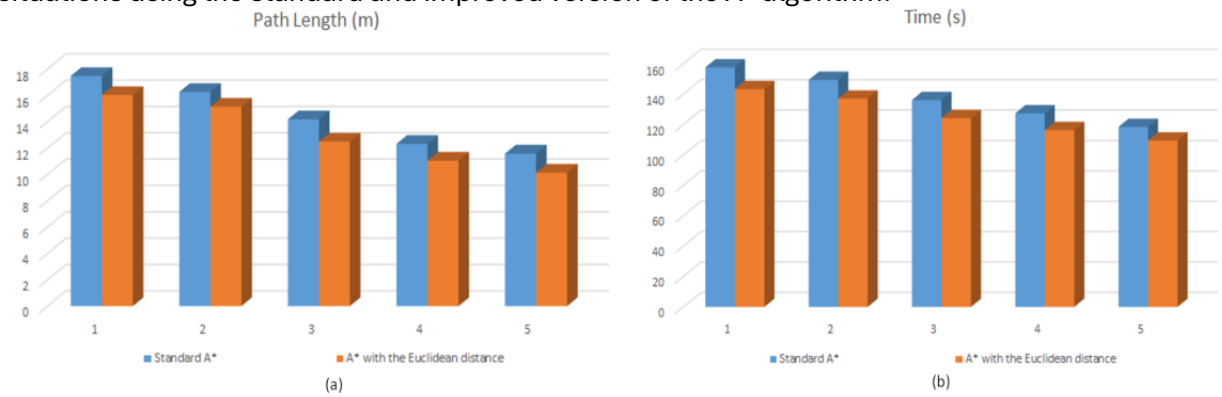


Figure 1. (a) A comparison of the lengths of the routes taken under the five different planning scenarios between the traditional A* and our improved A* with the Euclidean distance. (**b)** A comparison of the amounts of time required for each of the five different route planning scenarios between the traditional A* and our improved A* with the Euclidean distance.

5 Conclusions

In this work, the A* algorithm with the Euclidean distance heuristic represents a significant improvement in pathfinding, providing an effective and efficient technique for finding optimum routes. Extensive testing and assessment have shown that the upgraded A* algorithm with the Euclidean distance heuristic surpasses the standard A* method regarding path length and execution time. Consequently, the algorithm is well-suited for navigating unevenly shaped fields, avoiding obstacles, and optimizing numerous agricultural activities in smart farming applications. Future study in this area concentrate on improving the A* algorithm with new heuristics or investigating hybrid systems that combine the capabilities of different pathfinding algorithms.

6 Acknowledgment

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation (project code MIS 5047196).

7 References

[1] De Alwis, Sandya, et al. "A survey on smart farming data, applications and techniques."*Computers in Industry* 138 (2022): 103624.

[2] Pak, J., Kim, J., Park, Y., & Son, H. I. (2022). Field evaluation of path-planning algorithms for autonomous mobile robot in smart farms. *IEEE Access, 10*, 60253-60266

[3] Charania, I., & Li, X. (2020). Smart farming: Agriculture's shift from a labor intensive to technology native industry. *Internet of Things*, 9, 100142.

[4] Danielsson, P. E. (1980). Euclidean distance mapping. *Computer Graphics and image processing*, 14(3), 227-248.

[5] Erke, S.; Bin, D.; Yiming, N.; Qi, Z.; Liang, X.; Dawei, Z. An improved A-Star based path planning algorithm for autonomous land vehicles. Int. J. Adv. Robot. Syst. 2020, 17.

[6] A. Candra, M. A. Budiman and K. Hartanto, "Dijkstra's and A-Star in Finding the Shortest Path: a Tutorial," 2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics, 2020, pp. 28-32

[7] Kim, H.; Park, B.; Myung, H. Curvature path planning with high resolution graph for unmanned surface vehicle. In Robot Intelligence Technology and Applications (RiTA)

[8] Thirunavukkarasu K, Signh A S, Rai P, Gupta S 2018, December. Classification of iris dataset using classification based kNN algorithm in supervised learning. In 2018 4th Int. Conf on Computing Communication and Automation, pp 1-4.

[9] Liu, Q.; Zhao, L.; Tan, Z.; Chen, W. Global path planning for autonomous vehicles in off-road environment via an A-star algorithm. Int. J. Veh. Auton. Syst. 2017, 13, 330–339.

[10] Shi Y, Yang J, Bu S, Zhu L. Intelligent vehicle path planning algorithm based on improved RRT [J]. Computing Technology and Automation, 2019, 38 (04): 81-86.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 8 – 15

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Comparing the Performance of Cameras and SW packets for the Cloud Coverage Process in Photovoltaic (PV) Parks

C. Vasilakis[a], G. Venitourakis[a], A. Tsagkaropoulos[a], P. Tz. Amrou[b], G. Konstantoulakis[b], P. Golemis[b], D. Reisis[a]*

[a] *Electronics Lab, Physics Dpt, National & Kapodistrian University of Athens, Greece*
[b] *InAccess Networks, 12 Sorou Str, 15125, Athens, Greece.*
* *Corresponding author.*
dreisis@phys.uoa.gr

Abstract

The clouds' position in the sky, their creation, movements and dissipation have a significant impact on the energy produced in photovoltaic (PV) parks. To improve the power supply and the grid balance, the sky cameras will provide the images of the sky and edge computing devices will estimate the clouds' coverage in the sky and consequently compute the upper amount of energy that can be produced. For the execution of these processes the PV plants depend on systems of considerable cost that include the cameras and the supporting software most often executed on edge computing devices. The requirements for extreme weather conditions of these application specific cameras and their supporting hardware and software lead to high-cost figures. These facts decrease the efficiency defined as the results accuracy over the system's cost. Aiming at efficient solutions in the course of developing a PV park controller [1], the current study presents and compares the results of applying a cloud coverage process on systems with cameras of different cost and specifications. Moreover, it proposes an effective image segmentation technique for the cloud coverage executed on edge devices and it compares its accuracy to that of commercially available software. The comparison analyzes the differences in the accuracy results of three system configurations using each of two different cameras and either the proposed cloud segmentation algorithm or the commercial software. The real time performance of the proposed system algorithmic techniques on the edge-oriented NVIDIA Jetson Nano [2] validates the techniques.

Keywords: Image Segmentation, Cloud Coverage, Edge Computing

1 Introduction

The photovoltaic (PV) power plants have been integrated in the power grid and they continuously enhance their contribution to the energy produced. Their power production strongly depends on the cloud variability and on the seasonal conditions of their geographical location. How the clouds are created, how the parks are obscured by the clouds and how they disappear, has a significant effect on the power production especially in the large-scale power plants. Therefore, the estimation of the PV park's power production by the PV park controller in the time window of 5 minutes is crucial for improving the production as well as for the grid balance. This estimation becomes necessary in order to prevent the abrupt loss of a major fraction of the electricity production and possible blackouts in

the park's neighboring industrial, business and residential areas. To calculate the curve of the physical limits of the PV power, the park controller has to estimate the cloud coverage based on the current sky images. For this purpose, it includes a system based on a camera featuring an 180° "fisheye" lens to take all-sky pictures. These cameras are called sky imagers and since they are placed in the open most often, they have specifications for extreme weather conditions. Such complete systems are commercially available at a relatively high cost (in the class of 10-30 K€). Moreover, such systems' vendors provide software for image segmentation more often (specifically for cloud coverage). For a large number of smaller size parks though, the weather conditions are mild over the entire year and so, the high cost of such a complete system including a sky imager and the software becomes a challenge.

Targeting a cost-efficient solution to the cloud coverage problem the current paper shows the results of the research focusing on improving the cloud coverage estimation while employing systems of reduced cost, especially in geographical locations and places where these installations are feasible. The most significant contribution of this work is an image segmentation algorithm for cloud coverage, which can be executed on edge computing devices, and it was developed in the Archon project [1]. The proposed algorithm utilizes three techniques for: a) locating the sun in the sky image, b) removing from the image all the objects that irrelevant to the clouds and they are located close to the image's margin, such as buildings, antennas, etc. and c) classifications of all the pixels as cloudy or cloudless. The results of the comparison of the performance between the proposed algorithm and commercially available software [3] are equally important. The cameras used in the comparison are [3] and a low-cost camera equipped with fisheye lens [4]. The comparison involves three configurations of a complete system. The first consists of the sky imager ASI-16/52 system and this system's commercial software computing the cloud coverage [3]. The second includes the sky imager ASI-16/52 and the proposed image segmentation technique for cloud coverage. The third system uses the low-cost camera Vivotek [4] with the proposed Archon image segmentation software. The motivation for this work was the cost improvement of the PV park controller systems while keeping the performance with respect to the accuracy of the results and achieve the real time execution requirements on an edge computing device.

It is worth mentioning that reducing the cost of the cameras and their supporting hardware is realistic only in geographical locations of temperate climate and mild weather conditions. For a complete study and comparison there were tested various configurations, with each configuration consisting of a mixture of commercially available hardware and software solutions as well as software techniques developed in the project. This paper reports the three noteworthy combinations' results.

2 Related Work

There is a plethora of research results related to cloud coverage solutions, which target the accuracy improvement. Most of these reports consider satellite images while the number of studies in the category of techniques using images taken from the ground constantly increases [5]. For the latter category, the first process in the flow is locating the sun in a sky image, which is essential for cloud segmentation algorithms [6]. The authors of [6] introduce a sun tracking algorithm, where the sun's trajectory is estimated by correlating the position of the sun in a sky image with the timestamp of the image for a set of clear sky images with short exposure. The sun's location is revealed by thresholding clear sky images with short exposure based on pixel intensity. This method is camera and location agnostic but requires a dataset in order to identify the sun's trajectory. Simultaneously, results from [7] suggest that sun localization data improve the Forecast Skill (FS) of several CNN (Convolutional

Neural Network) models up to 13.75% for the task of irradiance forecasting. The second process isolates and removes from the image the ground objects [8] with mostly straightforward techniques such as removing the pixels close to the image margin. An alternative - automated - way to remove the obstacles and the noise from the images is the Machine Learning (ML) approach proposed in [9]. The authors use supervised machine learning for context awareness in addition to PCA (Principal Component Analysis) and HCA (Hierarchical Cluster Analysis) to reduce the data dimensionality. Focusing on clouds in images [10] and plant management, D. K Danso et al. conclude that ground cameras are more effective than satellite imagery. In [11], a multi-color criterion is proposed for the estimation of total cloud coverage of a whole-sky image system (commercial camera) by classifying the digital images in seven cloud classes and achieving 83% and 94% accuracy. The [12] study proposes a modification of the cloud detection Red-Blue Ratio (RBR) threshold by incorporating the difference between the two channels instead of their division. The authors in [13] compare the performance of six algorithms for cloud coverage taken with sky imagers; the five algorithms are based on thresholding criteria and the sixth utilizes neural network-based cloud segmentation to process the images; their conclusion is that the Hybrid Thresholding Algorithm (HYTA+) and the neural based are producing the best results.

3 Approach

This section presents the design of the techniques for the image segmentation cloud coverage. The proposed Near Infinite Conditions Image Segmentation Algorithm (NICISA) design specifies as input: a) the digital image taken by the sky imager, b) a filtering frame (mask) that is utilized in the process of eliminating the objects at the image margin that are irrelevant to the algorithm such as the buildings, antennas, mountains, etc. and c) the sun location in the image, which is computed based the latitude, longitude, altitude, time, angular offset, camera's focal length, distortion from the camera effects. The following subsections describe first the computation of the sun location in the image, next the filtering process for the unrelated ground objects in the image margin and finally, the image pixels classification as cloudy or cloudless. We note here that the resolution of the low-cost Vivotek camera is 2048x2048 pixels while that of the ASI-16/52 is 1536x1536 pixels and both cameras use 24 bits/pixel (8 bits/channel). Also, we note that the commercial software of the ASI-16/52 system performs defisheye process as the first stage of the cloud cover process, which cannot be deactivated, while all the processes of the NICISA are optimized for the image (both Vivotek and ASI-16/52 provide these files) directly taken from the fisheye lens (without involving any defisheye process).

Using either camera, the development flow for the NICISA software is as follows. The input is the camera image, the time and the geographical location of the camera (latitude, longitude and altitude). Then to accomplish the sun location process, the technique uses the pvlib-python. All these are loaded to the processing unit (either CPU or GPU) by using Python and the PyTorch package to result in the segmented image. The commercially available software that can be used in a complete system with the ASI-16/52 camera is based on the algorithm Schreder and can be executed only in a WindowsR operating system and only on a CPU (not on any GPU).

3.1. Sun Localization

The sun localization problem can be partitioned into two geometric problems with known solutions; a) locating the sun in the sky dome based on the time and location of the observer on Earth and b) projecting a point of a dome in a flat surface, which in this work it is the image frame. The proposed

system solves the first part by using astronomy equations that match time and location coordinates with solar azimuth and elevation. Those equations are implemented in the Pvlib python package [14], a package dedicated to solar power performance simulations. The sun's coordinates on the sky dome can be projected to a flat surface using fisheye equations. The second part of the sun localization problem depends on the projection characteristics of the camera and thus, an ideal fisheye projection is not possible; calibration may be needed in order to find accurate intrinsic coefficients. The number of parameters in this case are 11, which include the orientation of the camera, the sensor's physical dimensions and the 3x3 coefficient matrix of the focal length of the camera. Our solution defines a set of well-known fisheye mapping functions listed in [15]. By defining these mapping functions, we achieve the reduction of the number of hyperparameters to only three: a) the projection function, b) the orientation of the camera and c) a parameter dependent on the focal length of the camera, that is the distance between the lens' centre and the photosensitive chip that captures the image. Finally, the output is a single pixel in the input frame that represents the sun [7].

3.2. Sun Localization

For this process, the proposed solution initially creates a filtering frame of size equal to the input image frame. The PV park engineers decide which pixels will be filtered. Upon the camera installation at the PV park, the engineers check in the input frames the pixels that they will permanently contain ground objects such as trees, antennas, mountains, etc. They check these pixels manually and they create the filtering frame by using the vector graphics software *Affinity Designer* [16]. In this work, we choose to mark the pixels to be filtered as "black" and the remaining in the image as "white."

The processing of the image with the support of the filtering frame first considers whether the image pixel maps to a black or white pixel of the filtering frame. In case that the image pixel corresponds to a black pixel, then this pixel will be ignored for the remaining steps of the image segmentation. If this image pixel corresponds to a white pixel of the filtering frame, then it keeps its initial color and takes part in the segmentation process.

3.3. Image Segmentation Algorithm

For the cloud coverage process we developed the Near Infinite Conditions Image Segmentation Algorithm (NICISA), which is based on the algorithm presented in [8]. The algorithm utilizes for its calculations a very large number (therefore we call it "near infinite"), yet simple conditions. The NICISA is proposed because it is effective with the images taken with either high or low-cost cameras. Moreover, its computational complexity requirements suffice for execution on edge computing devices. The algorithm's output is the classification of the image's pixels either as cloudy or cloudless. The algorithm operates in the following two phases:

1. ***First Phase:*** it divides the image's pixels into these five color sets: green, red, yellow, white and blue. It uses RGB (Red, Green, Blue) and HSV (Hue, Saturation, Value) color spaces for the classification of the pixels into the five aforementioned sets. It also utilizes the distance of each pixel to the pixel that belongs to the sun and the sun altitude. The following steps highlight the sequence of the computations that distinguish the pixels into the five sets:
    a. First, it identifies the green pixels. These are the very bright pixels located close to the sun.
    b. Then it computes the red pixels set, which is a subset of the green set, but all of its pixels are (rather) cloudless.

c. It determines the yellow set: these pixels are less bright than the green and still brighter than the remaining in the image. These pixels are also located close to the sun (they are part of the solar halo). They may represent either sky or parts of clouds that are bright.

d. It classifies as white pixels those that represent definitely cloudy pixels.

e. It also classifies as blue pixels those that represent definitely cloudless pixels.

2. ***Second Phase:*** it realizes the final classification based again on the solar altitude, the distance between the pixel of the sun and the number of all the other pixels in each of the above sets.

   a. For all those pixels, which were not classified during the first phase in any of the five sets, it will decide which out of the two sets (blue or white) they belong to.

   b. All the green, red and yellow pixels are processed and they are classified as either blue or white.

Figure 1 and Figure 2 show the execution of each of the above steps of the proposed segmentation in the two images taken by the ASI-16/52 and the Vivotek cameras located next to each other. The two images were taken at exactly the same time. Note that in advance of the segmentation process, we have applied the sun localization and the removal of the ground objects processes in both images. For both Figures: in each Figure the first two images are the RGB and HSV color spaces and the following seven images present the results of each of the described above NICISA seven steps. A notable detail in the results is the small crack that appears in all the images (produced by both cameras), at the top of the horizon and which was produced by the process removing ground objects (here it was an antenna) from the input frames.

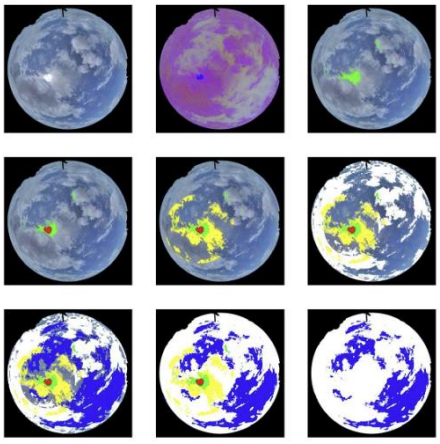


Figure 1: Image Segmentation on the ASI-16/52 [3] input frame.

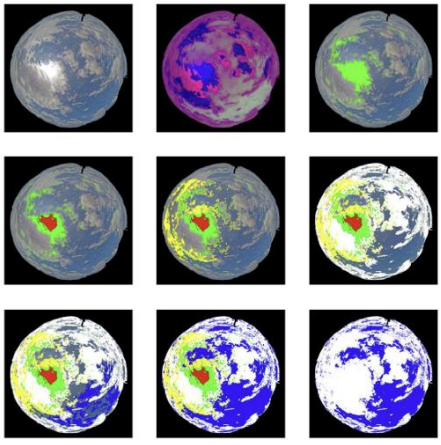


Figure 2: Image Segmentation on the Vivotek [4] input frame.

4 Performance Results

In order to evaluate the performance of the three configurations we tested a variety of weather conditions and cloud coverage. Given the fact that there were not available any datasets evaluated by meteorologists that provide the ground truth, the research team of "Archon" compared the results of the proposed NICISA with the well-established commercial software [3] that supports the ASI-16/52 camera.

Figure 3 presents the results in three indicative cloud coverage cases. We opted for these cases: the

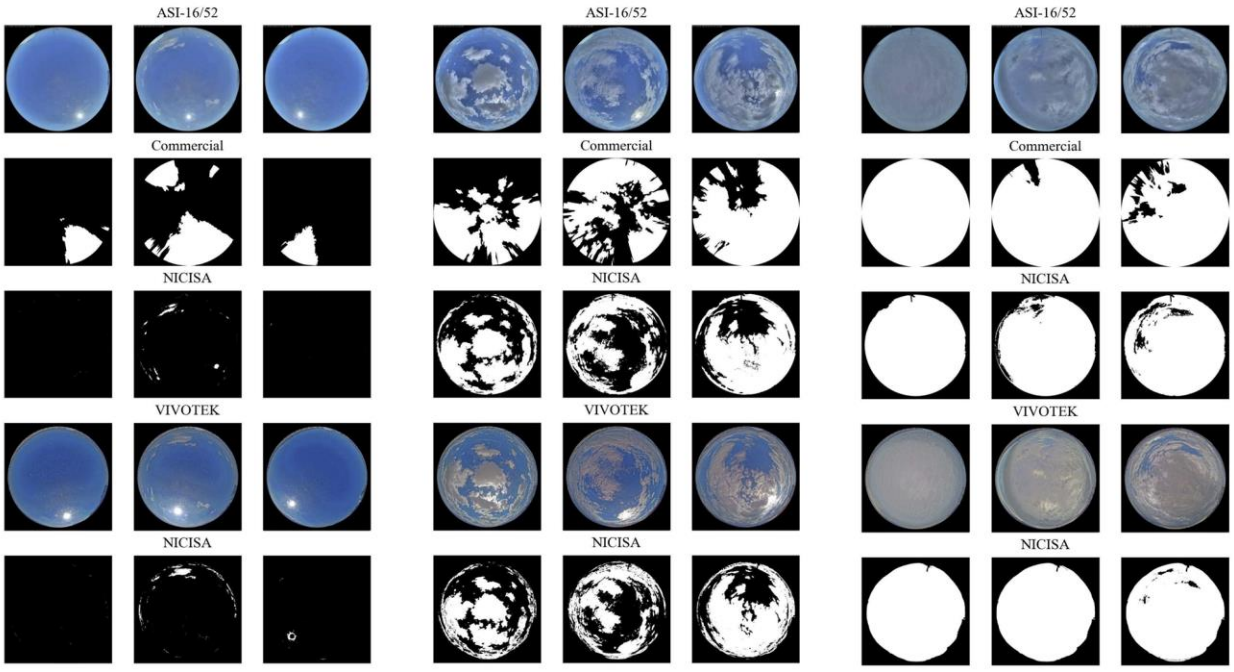
clear sky (**Figure 3a**), the partly cloudy sky (**Figure 3b**) and overcast (**Figure 3c**). For each of these three cases there are depicted 3 images at different times taken by both cameras. The first line in all subfigures shows the image frame taken with the ASI-16/52 camera. The second line shows the result of the image segmentation of the ASI-16/52 input performed with its commercial software. The third line shows the result of the image segmentation of the ASI-16/52 input performed with the Archon developed NICISA software. The fourth line shows the image frame taken with the Vivotek camera. The fifth line shows the result of the image segmentation of the Vivotek input performed with its NICISA software.

With respect to the real time performance, we note here that according to the controller specifications the PV park cameras provide an input frame every 60 seconds. The following table gives the execution time on the NVIDIA Jetson Nano [2], a widely known edge computing device. The performance is within the time limits of the PV park controller requirements. We also note that the device cannot execute the commercial software.

Table 1. Comparison of execution times between different algorithms.

Algorithm	Camera	Resolution (pixels)	NVIDIA Jetson Nano [2]
Commercial [3]	ASI 16/52 [3]	1536 ×1536	–*
NICISA	ASI 16/52 [3]	1536 ×1536	2.45 s
NICISA	Vivotek [4]	2048 ×2048	4.09 s

***cannot be executed on NVIDIA Jetson Nano**



a. Clear Sky **b. Partially Cloudy Sky** **c. Overcast Sky**

Figure 3: Comparison of Image Segmentation Algorithms for Cloud Cover. (From top to bottom) Row 1: Original image from ASI-16/52 sky imager. Row 2: The result of the commercial algorithm on the undistorted image from ASI-16/52. Row 3: The result of the novel NICISA on the image from ASI-16/52. Row 4: Original image from Vivotek camera. Row 5: The result of the commercial algorithm on the image from Vivotek. White pixels depict cloudy pixels. Foreign objects are masked out.

5 Conclusions and Future Work

This work focused on the comparison of image segmentation solutions for the cloud coverage problem and their utilization in PV parks. The comparison showed that the proposed NICISA set of techniques prevails with respect to the results accuracy and the capability of the execution on an edge computing device. Moreover, it achieves real-time performance. The comparison also showed that low-cost cameras can also substitute sky imagers of considerable cost in geographical locations with mild weather conditions.

Our future work will focus on further improvement of the techniques that the proposed NICISA uses. A possible approach to follow for a fully automated "removing ground objects" will be the correlation among several input frames and detection of objects that remain still over a long period of time. Another significant problem is the cloud coverage prediction and sun irradiance prediction for the next 10-minute window. These processes are of key importance to the PV park and grid management.

6 References

[1] "Archon," Inaccess Networks, 2022. [Online]. Available: http://archonproject.eu/.

[2] "Jetson Nano," NVIDIA, [Online]. Available: https://developer.nvidia.com/embedded/jetson-nano. [Accessed 26 06 2023].

[3] "ASI-16/52," [Online]. Available: https://www.eko-instruments.com/media/s01jrcy0/asi-16-52-cms-schreder eko-brochure.pdf.

[4] "Vivotek," [Online]. Available: https://download.vivotek.com/downloadfile/downloads/datasheets/fe9382-ehvdatasheet_en.pdf.

[5] S. Park, Y. Kim, N. Ferrier, S. Collis, R. Sankaran and P. Beckman, "Prediction of Solar Irradiance and Photovoltaic Solar Energy Product Based on Cloud Coverage Estimation Using Machine Learning Methods," Atmosphere, vol. 12, no. 3, p. 395, 2021

[6] P. Quentin and L. Joan, A Temporally Consistent Image-based Sun Tracking Algorithm for Solar Energy Forecasting Applications, arXiv, 2022.

[7] E. Papatheofanous, V. Kalekis, G. Venitourakis, F. Tziolos and D. Reisis, "Deep Learning-Based Image Regression for Short-Term Solar Irradiance Forecasting on the Edge," Electronics, vol. 11, no. 3794, 2022.

[8] J. Alonso-Montesinos, "Real-Time Automatic Cloud Detection Using a Low-Cost Sky Camera," MDPI - Remote Sensing, vol. 12, no. Special Issue Assessment of Renewable Energy Resources with Remote Sensing, 2020.

[9] L. Athanasopoulou, A. Papacharalampopoulos and P. Stavropoulos, "Context awareness system in the use phase of a smart mobility platform: A vision system for a light-weight approach," Procedia CIRP, vol. 88, 2020.

[10] D. K. Danso, S. Anquetin, A. Diedhiou and R. Adamou, "Cloudiness information services for solar energy management in west africa," Atmosphere, MDPI, 2020.

[11] A. Kazantzidis, P. Tzoumanikas, A. Bais, S. Fotopoulos and G. Economou, "Cloud detection and classification with the use of whole-sky ground-based images," Atmospheric Research, vol. 113, pp. 80-88, 2012.

[12] T. Rogiros and C. Alexandros, "Monitoring Cloud Motion in Cyprus for Solar Irradiance Prediction," in Conference Papers in Medicine, 2013.

[13] M. Hasenbalg, P. Kuhn, S. Wilbert, B. Nouri and A. Kazantzidis, "Benchmarking of six cloud segmentation algorithms for ground-based all-sky imagers," Solar Energy, vol. 201, pp. 596-614, 2020.

[14] W. F. Holmgren, C. W. Hansen and M. A. Mikofski, "pvlib python: a python package for modeling solar energy systems," Journal of Open Source Software, vol. 3, no. 29, p. 884, 2018.

[15] B. Felix, "Imaging: Fisheye lenses," WGN, Journal of the International Meteor Organization, vol. 33, no. 1, pp. 9-14, 2005.

[16] "Affinity Designer," Serif, 2022. [Online]. Available: https://affinity.serif.com/en-us/designer/.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 16 – 20

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Advanced Monitoring System of Industrial Refrigeration (ADMOSIR)

T. Athanasopoulos[1], G. Mavropoulos[1], A. Depastas[1], P. Vlagopoulos[1], K. Spyropoulou[1], E. Zervakis[1]

[1] *European Sensor Systems, 62 I. Metaxa str, Koropi-Athens,19441, Greece*

Abstract

The food supply chain industry often uses large scale industrial refrigeration systems with high installation and maintenance cost. In the same time, there are no permanent installations of systems that will monitor the parameters of the refrigeration systems and predict a malfunction (e.g. cooling fluid leak) or even improve the efficiency in order to reduce its power consumption. For example, a refrigeration system that contains 1500kg of cooling fluid will need 7500€ more electrical power if its efficiency drops by only 10% due to fluid leak.

1 Introduction

The European Union, towards the containment of the greenhouse effect has placed strict regulations (EE 517/2014) for the management of HFCs, fact that increased their cost and made leak detection management a necessity towards reducing the impact of cost.

Until now, most of the large industrial refrigeration installations are trying to solve the problem by having technicians performing preventive maintenance tasks. These procedures do not prevent leaks of HFCs and are essentially repair activities performed after the problem becomes apparent. The existing solutions of leak prevention are very complex and far too expensive to be used. ES Systems has implemented an innovating way of monitoring the refrigeration installation with wireless sensors and with the aid of Edge Machine Learning which leads to leak detection at the earliest possible stage and rate.

2 System Description

In order to effectively monitor the refrigeration system, a series of wired and wireless sensors must be installed in strategic places depending on their functionality. The system is composed of the following:

- A main processing unit that collects all the data wirelessly as well as wired.
- Wireless Temperature sensor.
- Wireless Pressure sensor.
- Wired power meter.

- A cloud server that will run the Edge Machine Learning
- Dashboard for depicting the results.

The main processor collects the data from the sensors and performs a pre-process algorithm in order to assist the Edge ML. When the pre-process has generated metadata, these are sent to the cloud server to perform the Edge ML and generate the algorithm. Data from there are displayed on the Thingsboard platform for the user/operator to fully understand the current situation but also see if there is a fault trend in the system. A complete overview of the system is shown in Figure 1.

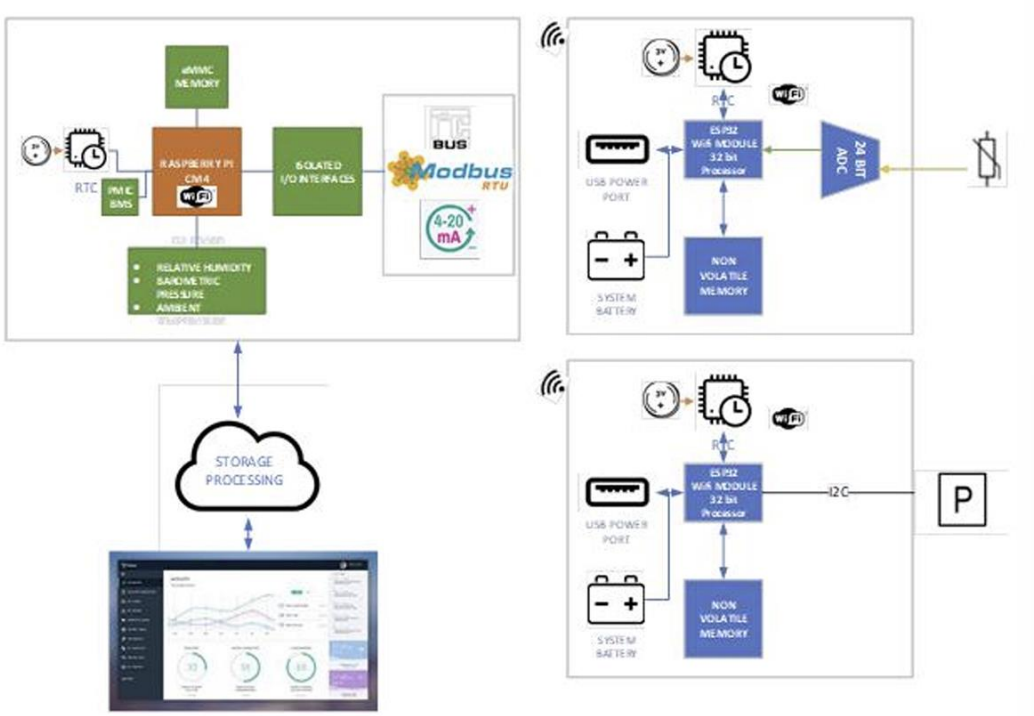


Figure 1. ADMOSIR System Description.

The sensors measuring the required physical properties send their data wirelessly at regular intervals to the "local" data collection and processing system. The current measurement data in relation to the data of the previous measurements compose the time profiles of the operating parameters of the refrigeration system. These profiles are processed by the intelligent algorithms which have been developed and "trained" in the stage of machine learning. During the training stage, the algorithms are implemented on the edge devices as well as on the IoT platform that provides this function where it is fed with sensor data through the cloud. When the results of the algorithms are considered reliable in relation to the expected results template data, these are transferred and implemented in the built-in local processor software. The innovation of the implementation lies in the following factors:

- The measurements, their processing and the alert generation / reporting is taking place at almost real time. The use of local processing with intelligence contributes to this algorithms so as to prevent failures (leaks etc.) before they happen or in their very early stages. Similar low-cost, real time system does not exist.
- The developed solution is a predictive maintenance based solution
- The sensors selected are modified accordingly (wireless, stand-alone) so that they can be

installed easily and at low cost in existing facilities where they were not foreseen.

The solution briefly includes:

- Design and development of a conversion package for conventional or industrial use, of sensors in wireless and wired sensors
- Design and development of data collection and processing node with the use of smart algorithms and ability to generate alerts and reports
- Design and development of algorithms on an IoT platform with processing data tools for the creation and training of the algorithms
- Data visualization platform
- Manufacturing and laboratory confirmation of prototypes (conversion package, wireless node)
- Application and field testing in refrigeration plants.

3 Edge Machine Learning

One of the stages of the system is to predict leaks to optimize the operation and increase the system efficiency. The prediction problem is solved by using Machine Learning techniques. The final goal of the system is to reduce operating costs, as well as to avoid system failures that have a large cost to the business. As Machine Learning in the past few years has shown encouraging results and applications in the edge environment, we applied inference at the edge, which refers to running pre trained machine learning models directly on edge devices.

Inference on the edge allows sensitive data to remain on the device, reducing the risk of data breaches during data transmission. This enhances privacy and data security, making it particularly important for applications involving personal or sensitive information. Edge devices can perform inference locally, reducing the time taken to process data and make decisions. This low-latency capability is essential for applications that require real-time or near-real-time responses, such as autonomous vehicles and smart surveillance systems. Furthermore, can perform inference locally, leading to a more distributed and scalable approach. This enables applications to handle increasing workloads without overwhelming centralized servers. Finally, running inference on the edge reduces the need for expensive cloud computing resources, resulting in cost savings, especially in deployments with a large number of edge device.

In summary, inference on the edge solves various challenges associated with traditional cloud based inference, offering benefits such as reduced latency, enhanced privacy, scalability and cost efficiency.

Data

The system consists of 5 features. One of these features pertains to the power consumption of the cooling unit in Watts. The other 4 features relate to the output of the condenser and the input of the compressor. For the condenser output, data is collected for the temperature in Celsius degrees and the pressure in bar. Similarly, for the compressor input, measurements are taken for the temperature and the pressure. Since there is no output/target value in the stored data, unsupervised learning algorithms were used for predicting leaks, specifically clustering algorithms or outlier detection algorithms.

Algorithms

The algorithms used follow different approaches. Specifically, the following three algorithms were applied:

1. Clustering Algorithm: The clustering algorithm was used to separate data into different groups based on their characteristics, without using output labels. This allows the detection of distinct behaviors and patterns in the data.
2. Support Vector Machines (SVMs): SVM algorithm was applied to create a hyperplane that separates data categories with the largest possible distance between them. This hyperplane is then used to predict the category or behavior of new data.
3. Decision Trees: The decision tree algorithm was used to create a tree that classifies data based on their characteristics. Each node of the tree represents a choice for a feature, and each leaf corresponds to a category or behavior.

With these algorithms, the detection and categorization of different system behaviors based on the data collected for the cooling units can be achieved.

Results

Firstly, the DBSCAN clustering algorithm was applied, achieving an accuracy close to 88%. Similarly, good values were obtained for Precision and False Positive Rate. However, the Recall and False Negative Rate did not perform as well. The OCSVM algorithm, which incorporates techniques from Support Vector Machines (SVMs), was then applied, but its results were worse than DBSCAN's, as well as the subsequent Isolation Forest algorithm. A detailed analysis of the results is presented in Table 1. Observing the table, the Recall for all three algorithms is relatively low. In practice, this indicates that none of the three algorithms correctly recognize the extreme values, highlighting their poor effectiveness in detecting system leaks.

According to the literature, proposed solutions that could be implemented in future work include tuning the hyperparameters of the algorithms, increasing the dataset size with longer experiments, and augmenting the dataset artificially.

Algorithm	Accuracy (%)	Precision	Recall	F1	False Positive Rate	False Negative Rate
DBSCAN	87.5	0.99	0.50	0.67	0.001	0.49
OCSVM	80.37	0.77	0.31	0.44	0.030	0.69
Isolation Forest	83.50	0.97	0.35	0.51	0.003	0.65

Table 1. DBSCAN vs OCSVM algorithm.

4 Conclusion

A system has been developed using conventional sensors that have been modified accordingly to fit the measurement environment variables. This includes power autonomy, transmission medium (wired, wireless), communication protocols and electromechanical interfaces. This enabled one of the key system characteristics, the retrofitting.

The system developed has been trained using data from a refrigeration test set up with controlled inputs and outputs so as to get all the system performance signatures for the algorithms training. However, training has been proven not sufficient due to the limited dataset sizes required. This has been proven in both ML evaluation using different algorithms as well as real field testing on a standard large scale refrigeration system.

The main reason is the significant difference in system behavioral parameters due to the system scale as well as the limited dataset.

The system is currently further improved by collecting more data of the same large refrigeration system to continuously improve the training algorithms.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 21 – 22

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Implementation of a Beehive Health Monitoring System Based on Sound

Dimitrios Kampelopoulos[1], Giannis Sofiannidis[1], Vasileios Konstantakos[1], Spyridon Nikolaidis [1], Kostas Siozios[1]

[1] *Aristotle University of Thessaloniki, Thessaloniki, GR*
(dkampelo, isofiani, bkons, snikolaid, ksiop} @physics.auth.gr

Abstract

A beehive monitoring system was implemented for monitoring the internal and external conditions of the hive, as well as detecting beekeeping phenomena, like swarming, queen bee's absence and arrhenotoky, based on the acquired sound.

1 Introduction

Given how vital the honeybees are for the stability of the ecosystem by being responsible for the reproductive process of several plant species and for maintaining biodiversity, a lot of scientific effort is given into developing methods and systems to monitor and protect bee populations. In this context, these approaches focus on developing acquisition systems equipped with sensors for measuring temperature, humidity, carbon dioxide, weight, as well as acoustic sensors. Acoustic methods are non invasive and are proven to be successful in detecting and, even, predicting various beekeeping phenomena like swarming, and the queen bee's absence [1]. Some of the proposed approaches involve the extraction of the Mel's Frequency Cepstral Coefficients (MFCC) which are employed for multiple sound-related applications, including in beehive health monitoring [2].

In this work, a sensor system was implemented, equipped with environmental, weight, motion sensors, microphones and more, in order to monitor the condition of different beehive populations under different phases. By monitoring the sound of multiple beehives for a prolonged period, it was possible to detect phenomena like the queen bee's absence, arrhenotoky, and swarming. The proposed system is able to detect the phenomena through a statistical analysis of the MFCCs extracted from the acquired audio data.

2 Method – Sound Processing

First, a large audio dataset was created from the recordings of multiple hives over a three-year experimental process, during which the beekeepers interfered at multiple occasions in order to force the different beekeeping phenomena. The MFCCs were extracted for the different phases of the experiment and their statistical distributions were analyzed. As a result, it was possible to make a clear distinction between the different phases of the experiment based on the values of these features.

The decision-making process is based on creating an initial reference of the beehive's sound and then comparing each subsequent measurement to that reference. The reference is eventually determined

by the means and standard deviations of the MFCC. Each subsequent measurement's MFCC values are compared to the reference distributions and a relative probability is calculated. The closer a feature value is to the mean the greater the probability. As a result, when a phenomenon takes place, certain mel bands exhibit drastically lower probability values, thus alarming for a possible occurring phenomenon. Depending on which mel bands are differentiated, it is possible to distinguish between the different phenomena and generate an estimated probability.

3 Implementation

The system is based on the SAMD21 microcontroller of Arduino Nano 33 IoT which is responsible for acquiring environmental data, performing digital sound recording, weight acquisition, bee traffic monitoring, lid status control and motion detection and on Arduino MKR NB1500 for narrowband IoT connectivity. It is equipped with an environmental sensor, a PDM MEMS microphone, weight load cells, infrared reflective object sensors, a magnetic reed switch and an accelerometer. For the sound processing, the MFCC extraction was optimized in terms of memory given only 32KB available in SAMD21 and a microSD card is used for local storage. The results of the analysis are transmitted to a webpage dashboard using the narrowband IoT connection. The system is also equipped with an RTC clock for accurate timestamps, as well as a photovoltaic cell. The harvested energy from the cell is used to charge a Li-ion battery, powering the system.

4 Conclusion

The proposed system is a low-power solution for beehive monitoring. The current implementation can be installed on a beehive and autonomously monitor its conditions and protect it.

5 Acknowledgement

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T2EDK 01681).

6 References

1 Terenzi, A., Cecchi, S., & Spinsante, S. (2020). On the importance of the sound emitted by Honey Bee hives. Veterinary Sciences, 7(4), 168. https://doi.org/10.3390/vetsci7040168

2 Nolasco, I., Terenzi, A., Cecchi, S., Orcioni, S., Bear, H. L., & Benetos, E. (2019). Audio based identification of Beehive states. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp.2019.8682981

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 23 – 24

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

An Audio Fingerprinting Approach Based on the 2DFT for Byzantine Hymn Recognition

Dimitrios Kampelopoulos[1], Lazaros Moysis[1], Konstantinos Karasavvidis[1], Achilles D. Boursianis[1], Sotirios K. Goudos[1], Spyridon Nikolaidis[1]

[1] *Aristotle University of Thessaloniki, Thessaloniki, GR*
({dkampelo, lmousis, kokarasa, bachi, sgoudo, snikolaid} @physics.auth.gr

Abstract

An audio fingerprinting technique is proposed, in this work, for the challenging task of Byzantine hymn recognition. This approach involves the extraction of speed and scale invariant fingerprints, which is crucial to the characteristics of the application, as well as an evaluation process in order to reliably match a query to the correct hymn of the database.

1 Introduction

Audio fingerprinting is a technique commonly applied in voice recognition, sound event detection, musical information retrieval and cryptography [1]. In the field of music, there are established methods to recognize songs in a timely manner, but the task is mainly to match a distorted recording of a track to the official track in the database. However in Byzantine music the hymns are primarily performed live by different chanters and in varying tempo and scale. So for hymn recognition, it is important to not only eliminate the noise but also to be resistant to these kinds of variations.

In this work, the focus was given on the extraction of time and scale invariant fingerprints through a process involving the 2-dimensional Fourier Transform (2DFT), originally proposed by [2] for cover song identification, that was customized to the Byzantine music's characteristics. Also, a statistical polling technique was introduced as a final step and the whole process was optimized in order to achieve reasonable querying times. The method was tested on a hymn database containing 359 Byzantine hymn executions by multiple chanters.

2 Method

The first part of the process is fingerprint extraction. The Constant-Q Transform (CQT) was applied on the raw audio data in order to extract a spectrogram with a logarithmic frequency axis divided into the musical scales. This logarithmic transformation makes the tone changes linear and are eliminated in a later part of the process. The next step is performing adaptive thresholding on the CQT which is done by applying a median filter. The last step is to apply the 2DFT on the filtered spectrum in order to extract a sequence of Fourier coefficients. This sequence is the actual fingerprint that will be used for the

recognition. This process is performed on each track of the hymn database and the resulting fingerprints are precalculated and stored.

To compare two hymn executions the process involves the calculation of a distance metric. This distance is calculated by the similarity matrix of two fingerprint sequences. First, the mean of the similarity matrix is calculated and then a Gaussian filter is applied, in order to create distinct diagonals. The distance is calculated as the sum of the three main diagonals. As a next step, the track is resampled in varying sampling frequencies and the one with the minimum distance is selected. The correlation metric was chosen for the similarity matrix distance calculation which was found to produce better results than the Euclidean distance used in [2].

So, in order to match a query audio to a hymn in the database, the distance is calculated between the query and every hymn execution of the database. The executions exhibiting the minimum distance are the match candidates and usually the minimum distance corresponds to the same hymn. However, an additional step was introduced in which a probability is associated to every match candidate, analogous to the inverse of the distance. As a result, the choice is not always linked to the minimum distance, but the most probable hymn found in the match candidates.

3 Results

For the purpose of this work, a database of 359 executions were recorded on a set of nine different hymns by different chanters in varying tempo and speed. To evaluate the proposed method, each execution of the database was used as a query and was compared to the rest. The process was, also, optimized in terms of speed, so that a single query can be performed in a few seconds depending on its duration. A parametric execution of the algorithm followed which determined the best combination of parameters for this specific application. With this choice of parameters the algorithm exhibited 99.44% accuracy with the statistical polling method versus 99.16% with the original minimum distance one.

4 Conclusion

With the proposed method, it was possible to make a clear distinction between different executions of the same hymn performed by different chanters, in varying tempo and scale. The algorithm exhibits high accuracy and improved results by introducing a statistical polling step.

5 References

[1] Cano, P., Batlle, E., Kalker, T., Haitsma, J. (2005). A review of audio fingerprinting. Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology, 41(3), 271–284. https://doi.org/10.1007/s11265-005-4151-3

[2] Seetharaman, P., Rafii, Z. (2017). Cover song identification with 2d Fourier transform sequences. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp.2017.7952229

EMERGING TECH CONFERENCE 2023
Edge Intelligence

Papers

Session 1.2 | Novel applications

Session Chairs: **Nikolaos KONOFAOS**

IDEAN Neurosciences Education Platform
Aggelos Kostas, Anastasios Vittas, Eirini Georgia Dimitriou, Konstantinos Kalafatakis, Kalliopi Basiakou, Nikolaos Giannakeas, Alexandros Tzallas, Nikolaos Katertsidis and Markos G. Tsipouras

SmartGlove Cloud Platform for Parkinson's Patients Data Collection & Analysis
Eirini Georgia Dimitriou, Vasiliki Fiska, Konstantinos Kalafatakis, Nikolaos Giannakeas, Alexandros Tzallas, Nikolaos Katertsidis and Markos G. Tsipouras

Rapid virtual prototyping of battery storage systems
Sotirios Athanasiou and Georgios Koukos

S2W-Med Health Status Monitoring
N. Sionas, G. Pomportsis. P. Christodoulou, I. Tzamtzis

Fish Shape Alignment Based on Deformable Shape Tracking Suite of Tools
Nikos Petrelis, Georgios Keramidas, Christos Antonopoulos and Nikolaos Voros

HETiA
EXPANDING THE DIGITAL FRONTIERS

ARISTOTLE UNIVERSITY OF THESSALONIKI

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 25 – 27

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

IDEAN Neurosciences Education Platform

Kostas Aggelos[1], Vittas Anastasios[1], Dimitriou Eirini Georgia[1], Konstantinos Kalafatakis[1,2], Kalliopi Basiakou[1], Nikolaos Giannakeas[3], Alexandros Tzallas[3], Nikolaos Katertsidis[1], Markos G. Tsipouras[1,4]

[1] *Univeye IKE, Ioannnina, Greece*
[2]*Institute of Health Sciences Education, Barts and the London School of Medicine & Dentistry (Malta Campus), Queen Mary University of London, Victoria, Malta*
[3]*Dept. of Informatics and Telecommunications, University of Ioannina, Arta, Greece.*
[4]*Dept. of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece*
agelos.kwstas@gmail.com, tasosvittas@gmail.com, dim.georgian@gmail.com, k.kalafatakis@qmul.ac.uk, kelly.basiakou@gmail.com, giannakeas@uoi.gr, tzallas@uoi.gr, nkaterts@univeye.com, mtsipouras@uowm.gr

Abstract

The objective of this project is the development and integration of the IDEAN Platform, an online educational platform that provides personalized education in Integrative Neuroscience at the secondary and tertiary levels of education, both undergraduate and postgraduate. The platform encompasses a large volume of information related to the aforementioned knowledge domain, with English as the base language. Dissemination of the information is achieved through a variety of cutting-edge online audiovisual media and a range of educational approaches and tools. The aim of the project is to facilitate the learning of these fields using a modern and interactive approach, enabling users to acquire theoretical and practical knowledge in the available areas of study.

1 Introduction

The IDEAN Platform represents a modern and comprehensive e-learning system designed to facilitate the provision of educational content and resources to users. This application has been implemented using Moodle, a popular open source platform for managing the educational process. The configuration of the application has been tailored to adapt to the needs and requirements of the IDEAN Platform.

2 Platform Functionalities

User authentication is based on the rights assigned to each user within the IDEAN Platform. The administrator of the platform has the authority to add or remove users. All users are defined by the administrator, who assigns them appropriate privileges. User registration does not occur through self-registration instead, individuals must be verified by an administrator. Otherwise, they have only guest privileges, allowing them to have a preview of the available courses without access to their content. Types of users include Administrator, Professor, Student & Guest.

Course Organization. The IDEAN Platform offers courses in Neural Networks and Neurosciences.

The platform administrator oversees course creation and instructor assignment for new users. These courses aim to provide in-depth coverage of each chapter, offering ample subject-related material. Additionally, they incorporate interactive elements to engage students effectively. The course material is developed and provided by the assigned instructor. Each course offers video presentations with subtitles and voice-over for accessibility, quizzes based on each chapter, adapted levels based on quiz performance (Beginners, Intermediate, Expert), and live point tracking and grading for student engagement and progress. The ranking of students is visible to all based on their grades. In order to achieve the above, the platform utilizes a vast array of plugins such as: Custom Certificate, Open Forum, Adaptable Theme, Advanced Notifications, Level Up XP – Gamification and Boost Navigation Fumbling.

Forum & Chat. The Open Forum activity facilitates the exchange of ideas between students and instructors by posting comments as part of a "thread." Students access the forum by clicking on the icon and have the ability to create a new discussion topic. Instructors have three additional dots on the right side of the "Enroll" option, allowing them to pin, star, or lock discussions. When responding to a post, instructors can optionally send a Private Reply that is visible only to the specific student. Students cannot reply to this private reply. On the other hand, the Chat feature enables participants to engage in real-time conversations.

3 Platform Design and Architecture

The central platform architecture is entirely based on cloud computing services to ensure platform availability and scalability. This approach allows the platform to dynamically allocate additional resources as needed, ensuring smooth system operation and releasing them when no longer required. This approach offers cost savings, efficient resource management, and scalable flexibility through cloud services, reducing capital expenditure and enabling seamless adaptation to growing needs. Furthermore, no initial hardware purchase is required as the platform utilizes cloud provider services. Also, costs are tied to the resources utilized, enabling effective management of financial resources. Finally, the platform can automatically adapt and scale to meet growing needs.

The main dashboard of the platform (Fig.1) shares similarities with the preview available to regular users. The key differences lie in the accessibility of courses, where students can enter the visible ones, while visitors do not have this capability. Additionally, the platform's forums are visible, allowing students to exchange opinions, seek assistance, and communicate with instructors. Lastly, the live status of other users/students is displayed.

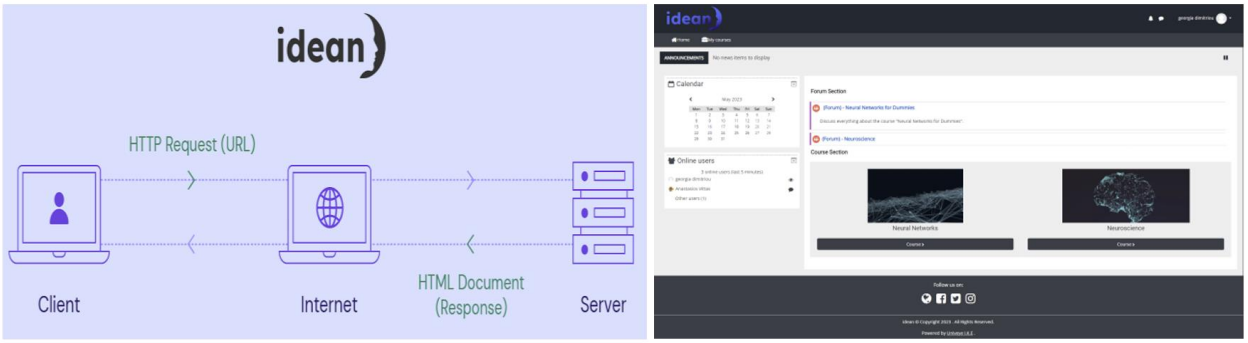


Figure 1. (a) Cloud Architecture, and (b) Application's Interface.

4 Discussion

The IDEAN Platform is an online educational platform that focuses on Integrative Neuroscience education at the secondary and tertiary levels. It offers personalized learning experiences using cutting-edge online audiovisual media and a variety of educational approaches and tools. The platform incorporates gamification to create a modern and interactive learning environment for Integrative Neuroscience education. By including gamification elements such as quizzes, challenges, and rewards, the IDEAN Platform aims to engage learners and enhance their overall learning experience. Its comprehensive range of resources, functionalities, and gamified components work together to facilitate learning and knowledge acquisition in the field of Integrative Neuroscience.

5 Acknowledgment

This work is part funded from the Operational Programme Competitiveness, Entrepreneurship and Innovation 2014 2020 (EPAnEK) (Project Code: ΓΓ1CL-0058895).

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 28 – 30

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

SmartGlove Cloud Platform for Parkinson's Patients Data Collection & Analysis

Dimitriou Eirini Georgia[1], Vasiliki Fiska[2], Konstantinos Kalafatakis[1,3], Nikolaos Giannakeas[4], Alexandros Tzallas[4], Nikolaos Katertsidis[1], Markos G. Tsipouras[1,2]

[1] *Univeye IKE, Ioannina, Greece*
[2]*Dept. of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece*
[3] *Institute of Health Sciences Education, Barts and the London School of Medicine & Dentistry (Malta Campus), Queen Mary University of London, Victoria, Malta*
[4] *Dept. of Informatics and Telecommunications, University of Ioannina, Arta, Greece.*
dim.georgian@gmail.com, dece00073@uowm.gr, k.kalafatakis@qmul.ac.uk, giannakeas@uoi.gr, tzallas@uoi.gr, nkaterts@univeye.com, mtsipouras@uowm.gr

Abstract

Parkinson's disease is a neurodegenerative disorder that affects millions of individuals worldwide. To aid in the diagnosis, monitoring, and treatment of Parkinson's patients a comprehensive platform has been developed, comprising a characterized database and a user interface (UI) for data access, visualization, and monitoring. The platform utilizes sensor signal recordings acquired using a smart glove for the automatic assessment of the patients' motor status.

1 Introduction

Parkinson's disease is a chronic neurodegenerative disorder characterized by motor symptoms such as tremors, rigidity, and bradykinesia. As the disease progresses, these symptoms worsen, significantly impacting the quality of life of affected individuals. Traditional methods for monitoring Parkinson's patients primarily rely on subjective clinical assessments, leading to potential inaccuracies and limited understanding of disease progression. However, advancements in technology, including the use of smartphones, have opened new avenues for objective and continuous monitoring of patients. Innovative devices, such as smart gloves equipped with sensors [1] [2], offer the ability to collect precise and quantitative data on motor function [3]. Cloud platforms provide scalable and secure environments for managing the vast amount of data generated by Parkinson's patients. By leveraging cloud infrastructure, researchers and clinicians can access, analyze, and store sensor signal recordings and processed results efficiently.

2 The SmartGlove Cloud Platform

The proposed platform serves as a secure and intuitive cloud-based solution for storing and analyzing patient data. By establishing communication with the mobPark application, the platform facilitates the retrieval of patient records, including data from the SmartGlove device, questionnaire responses,

and test results obtained through gaming applications. Additionally, the platform offers an interactive interface for users to access, visualize, and compare the collected data. It also acts as a Decision Support System, providing therapeutic suggestions, medical alerts, and additional medical data to enhance clinical decision-making and patient management.

To utilize the central management platform, both medical staff and patients are required to authenticate using assigned usernames and passwords. This authentication process ensures secure access to the platform's features, particularly the SmartGlove functions. Access to SmartGlove features is only possible with valid credentials. User authentication encompasses two distinct categories: 1) Patients access the platform's services through the SmartGlove application.

Their authentication is performed within the application using the provided API from the platform. 2)Medical Staff and System Administrators access the central platform directly. They authenticate themselves using their designated usernames and passwords, enabling them to leverage the platform's comprehensive functionalities.

The Central Platform is built in PHP, the popular language for dynamic web content. It utilizes the Yii2 Framework, offering a convenient starting point for development and supporting the Model-View-Controller (MVC) architecture. Yii2 integrates HTML5, Bootstrap CSS, and JavaScript, enhancing the platform's capabilities. It is cloud-based, offering availability and scalability (Fig. 1). It eliminates the need for hardware equipment and reduces capital expenditures. The cloud environment includes a public and private subnet, with communication through an internet gateway. The public subnet consists of a load balancer for workload balancing and a NAT gateway for secure traffic redirection to the private subnet. This cloud-based approach ensures flexibility, scalability, cost-effectiveness, and strong security measures. Within the private subnet, the web server facilitates user interaction with the system. Users can authenticate and interact through two methods: via the API for patients using the mobPark application and through the web application's corresponding screens. Additionally, the system includes a Processing Unit responsible for the intelligent functionality. It is divided into two main parts. The first part involves training models for automatic patient evaluation, while the second part focuses on implementing the computerized system.

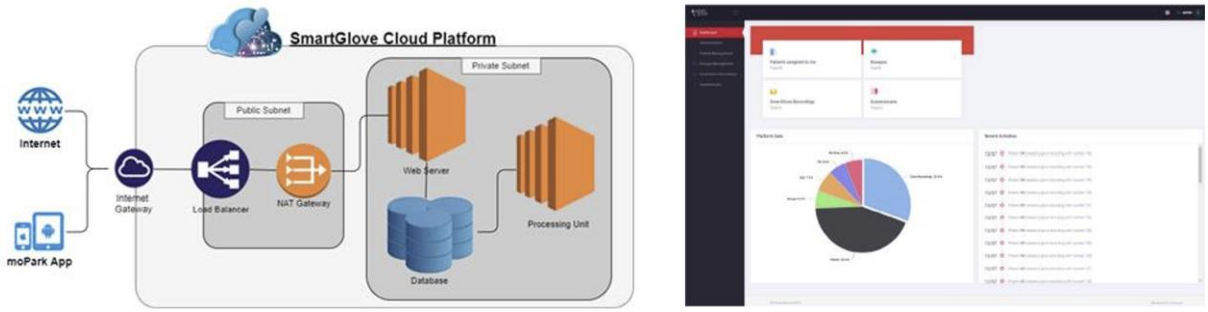


Figure 1. (a) SmartGlove Cloud Architecture, and (b) Main Dashboard.

3 Discussion

Overall, the proposed platform facilitates the creation of a standardized and accessible resource for sensor signal recordings from Parkinson's patients. This enables collaborative research efforts, facilitates data sharing, and promotes the development of predictive models and personalized

interventions. Furthermore, cloud-based storage ensures data durability and availability, minimizing the risk of data loss and providing a foundation for long-term studies and continuous monitoring. By combining advanced data processing techniques and an intuitive UI, the proposed cloud platform holds great potential for advancing our understanding of Parkinson's disease and improving the management of affected individuals.

4 Acknowledgment

This work is part funded by the project "SmartGlove for Assessment of the Motor Condition of Patients with Neurodegenerative Diseases (SmartGlove)", co-financed by the European Union and Greek national funds through the Operational Program for Research and Innovation Smart Specialization Strategy (RIS3) of Epirus (Project Code: HP1AB-0028207).

5 References

[1] V. Fiska, V. Gakilazou, N. Katertsidis, A. T. Tzallas, N. Giannakeas and M. G. Tsipouras, "Automatic Parkinson's tremor assessment with data analysis from daily activities," 2021 6th (SEEDA-CECNSM), Preveza, Greece, 2021, pp. 1-5, doi: 10.1109/SEEDA-CECNSM53056.2021.9566225.

[2] Mourtzis D, Angelopoulos J, Panopoulos N. Smart Manufacturing and Tactile Internet Based on 5G in Industry 4.0: Challenges, Applications and New Trends. Electronics. 2021; 10(24):3175.

[3] V. Fiska, N. Giannakeas, N. Katertsidis, A. T. Tzallas, K. Kalafatakis and M. G. Tsipouras, "Motor data analysis of Parkinson's disease patients," 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), Cincinnati, OH, USA, 2020, pp. 946-950, doi: 10.1109/BIBE50027.2020.00160.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 31 – 38

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Rapid virtual prototyping of battery storage systems

Sotirios Athanasiou[1] and Georgios Koukos[2]

[1] *Sunlight Group Energy Storage Systems, 14564 Athens, Greece*
[2] *Sunlight Group Energy Storage Systems, 67200 Neo Olvio, Greece*
s.athanasiou@sunlight.gr, g.koukos@sunlight.gr

Abstract

With this work we present the current approach into virtual prototyping of battery storage systems with MATLAB-Simscape. We show the robustness and effectiveness of the approach as well as the easiness to implement. Multiple components of the battery can be modeled this way and we show that even complicated behavior can be observed. A SOLIDWORKS design provides a more detailed approach to the battery design.

1 Introduction

Studies of battery storage systems in the industry can be roughly divided into 2 categories. Feasibility studies are done for a client, where the designers try to evaluate whether the selected battery cells can fit an application based on rough electrical and mechanical requirements. Virtual prototyping on the other hand can help in evaluating new technologies tested before they reach the actual product development teams. In this work we focus on virtual prototyping where we highlight the basic Battery Module and Battery Management System (BMS) components that are modeled for a specific application. Previously Sunlight Group has worked on Hybrid Storage Systems based on Lead Acid Batteries (LABs) [1,2] and more recently on smaller systems with Lithium-Ion Batteries (LIBs) [3]. Through these works we have identified the importance of simulating battery systems at prototype level and how this can affect sizing and configuration of the battery.

2 Battery storage system technology

2.1. Battery cell models

One of the key aspects in simulating battery storage systems is the battery cell models used. There are 4 major categories for battery cell models: Data driven, Empirical, Equivalent Circuit Models (ECM) and Electrochemical Models [4]. In this case we will be using the default ECM models provided by Simscape:

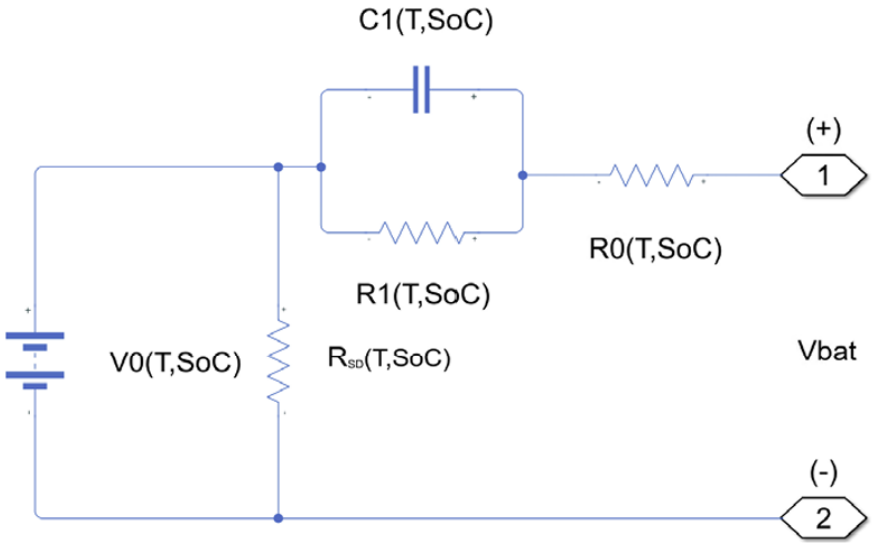


Figure 1: Battery cell ECM used.

Equivalent Circuit Models are models based on electrical elements that are used to emulate behavior of an electrochemical cell. In our case we are modeling the cell dynamic behavior with an RC circuit. Multiple cells are connected in series or in parallel for a battery pack which form the basis the battery. For this work we selected NMC chemistry lithium-ion cells with a capacity of 20Ah. The battery pack configuration selected was the 1P4S configuration.

2.2. Battery Management System

The battery management system is one of the key components of a Lithium-Ion battery. It is responsible for ensuring the safety of the battery, as well as calculating various parameters; for example, State of Charge (SOC) and State of Health (SOH). Regarding the safety of the battery storage system, the BMS is reacting towards opening the switching elements (effectively cutting-off power) in case a set of safety conditions is violated.

2.3. Battery charger / load

The energy storage system is connected to either a load or a charger at any given time. In some cases, it can be connected on both, but the battery can be either on Charging or Discharging state. Charging algorithms can be divided in different categories depending on the conditions used to charge the battery; Constant current constant voltage (CC-CV) charges the battery with a constant current, while afterwards the charger maintains a constant voltage. Constant power constant voltage (CP-CV) charges the battery with a constant power, while afterwards the charger maintains a constant voltage. In our case we will be using a CC-CV charger algorithm to charge the battery.

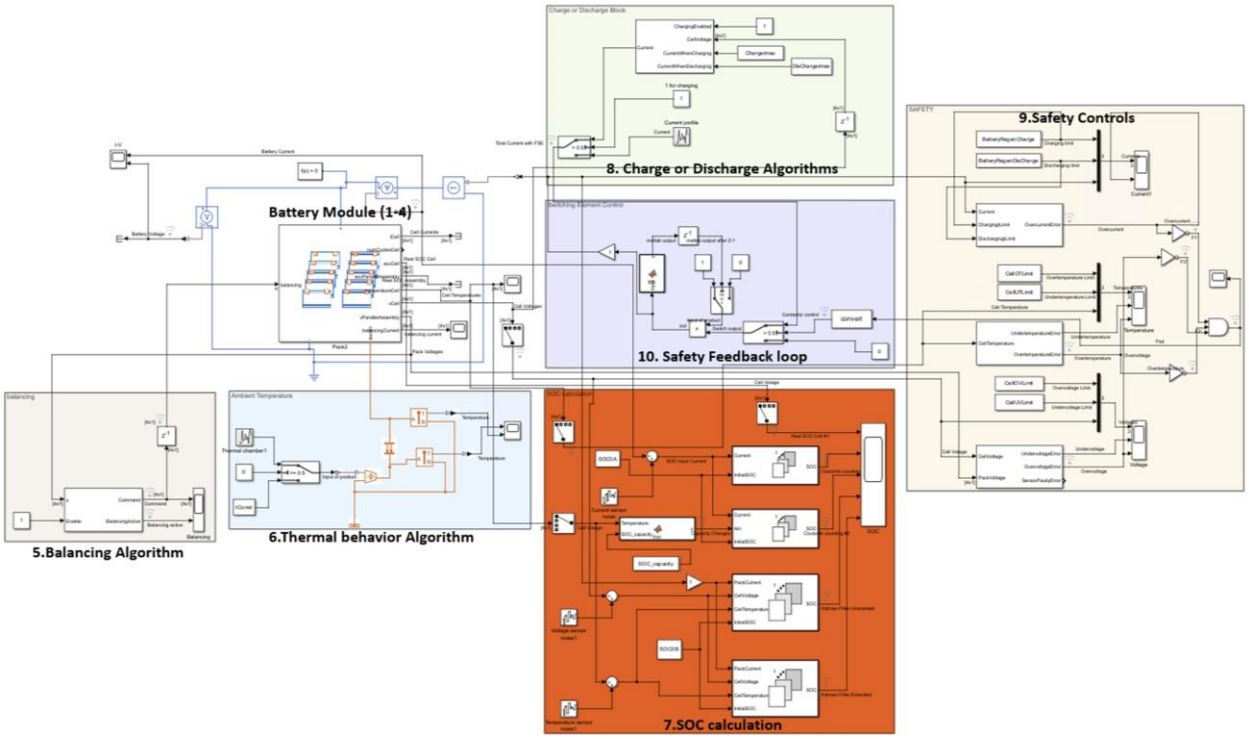


Figure 2: Battery testbench global view

3 Implementation

Battery parametrization is done through specific MATLAB-Simscape blocks and custom MATLAB functions. We parameterize cell dimensions and electrical connectivity of the battery pack. This virtual battery is then customized through initializing the components of the cells' ECM. Specific functionality was implemented to randomize the cell's status and force the energy balancing behavior. Figure 2 presents the experimental setup of the battery, with all its behavior modeled. All parameters were adjusted from parameters used in actual products to create a more accurate behavior. In figure 3 below we can see the methodology followed to create the battery model and its testbench. We will be briefly presenting the options of most of these blocks in the following paragraph and then we will be focusing on the safety control and safety feedback loop.

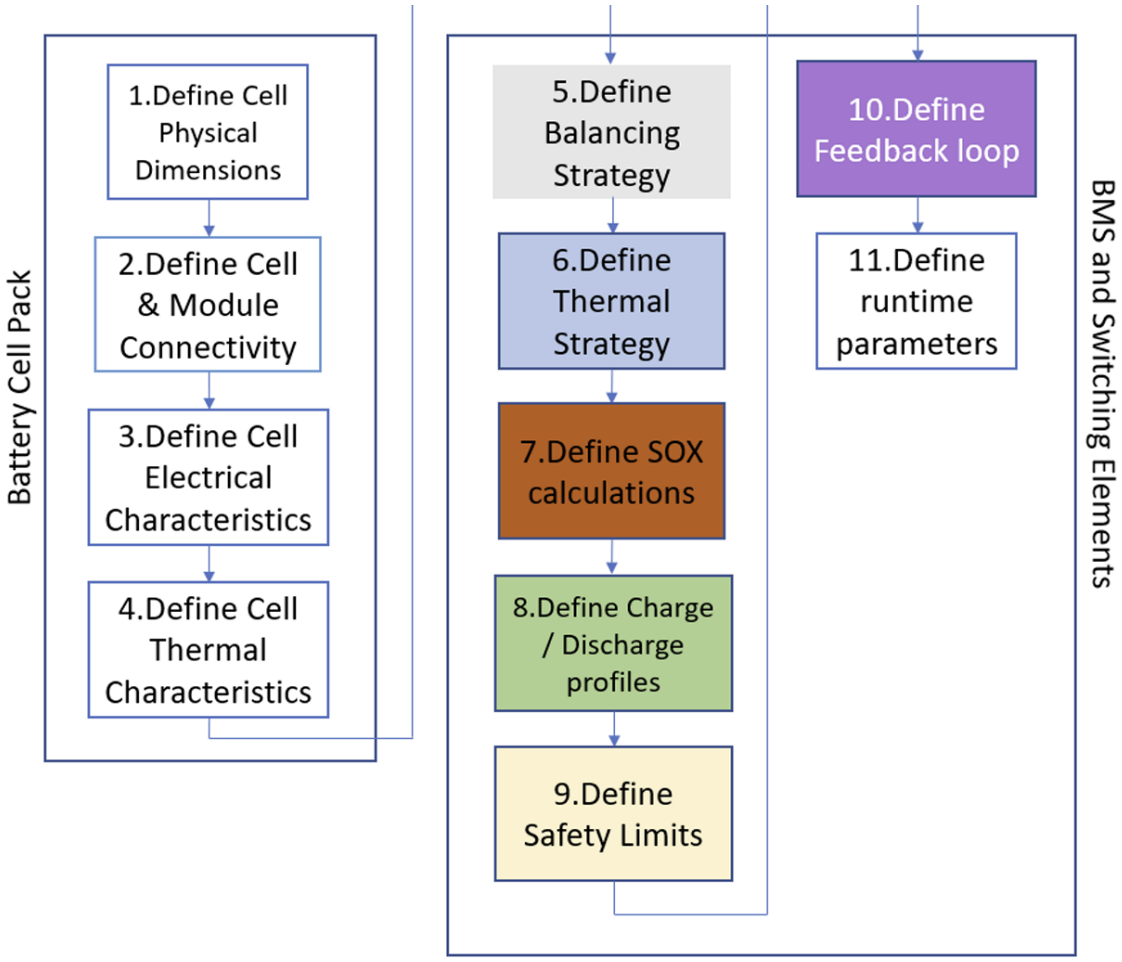


Figure 3: Methodology flowchart

3.1. Main setup

Generally, to define the battery pack you need to base your model on actual battery cell parameters. By using a cell datasheet, the mechanical (Figure 5) and some of the electrical characteristics can be defined (steps 1-4). For the missing parameters, cell characterization must be performed to obtain them. Alternatively, for parameters with a reduced impact on the results, simplified assumptions can be made. For the balancing strategy (step 5) we chose passive balancing with typical balancing threshold values. For the thermal strategy (step 6) we are assuming a constant environmental temperature that affects our cells through convective heat transfer. For the SOC subsystem (step 7) we compared the following algorithms:

1. Coulomb counting
2. Coulomb counting with variable capacity
3. Unscented Kalman filter
4. Extended Kalman filter

We can clearly see in Figure 4 the difference in calculated SOC values depending on the method used.

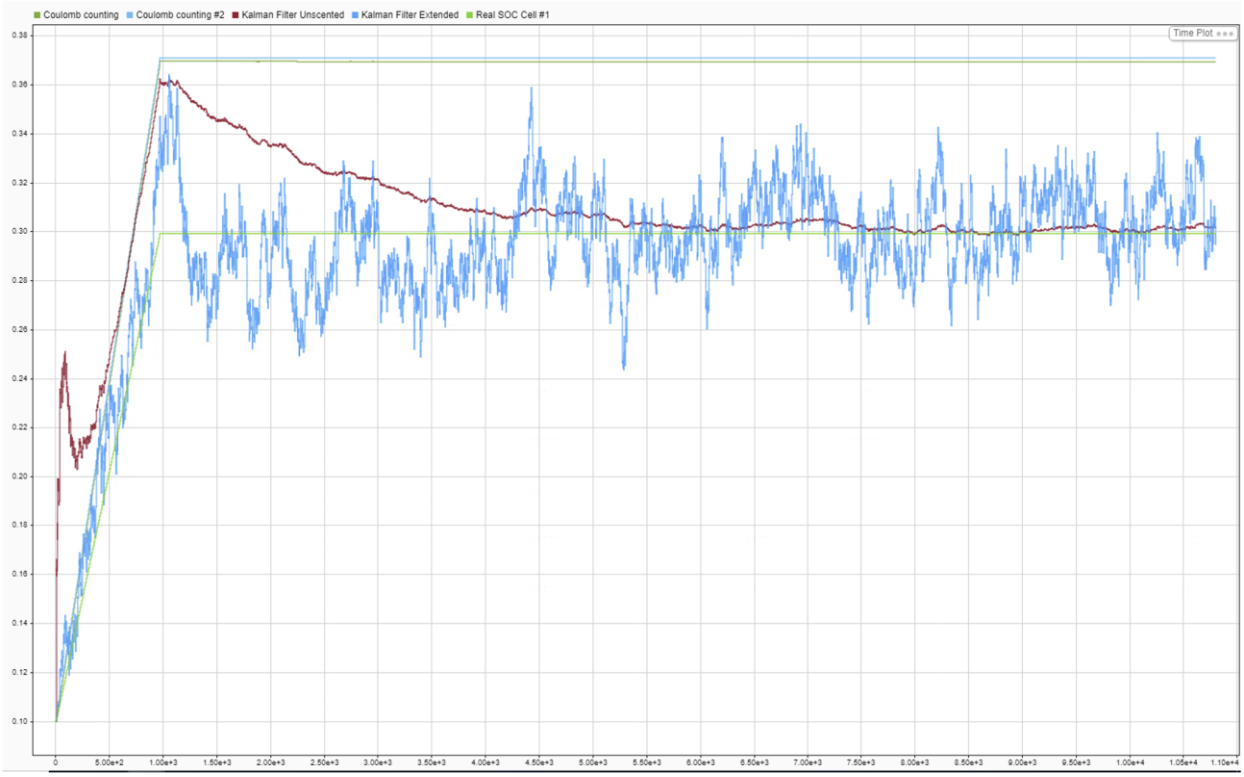


Figure 4: Comparison of different SOC calculation methodologies

Furthermore, in the case of charge and discharge algorithms (step 8) we applied a CC-CV charge algorithm with the maximum charge voltage extrapolated from the maximum voltage defined by the cell datasheet.

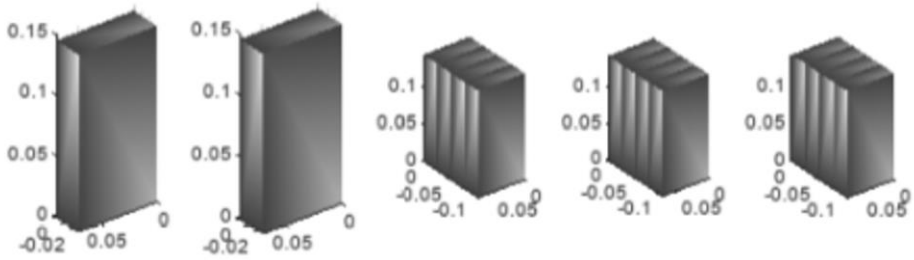


Figure 5: Battery pack physical modeling

3.2. Safety

One of the key aspects of a battery is its safety of Li-ion batteries [6]. To this end most of battery products need to comply with very strict regulations regarding functional safety [7]. In Figure 6 we demonstrate the way functional safety is implemented in Simscape. 3 different blocks simulate different BMS functions: Overcurrent protection, over/under voltage protection and over/under temperature protection are implemented. The system generates an error in case any of these limits are violated, which in turn opens a switching element. This action cutoffs completely the current flow

in or out of the battery, thus leading the battery to a safe state.

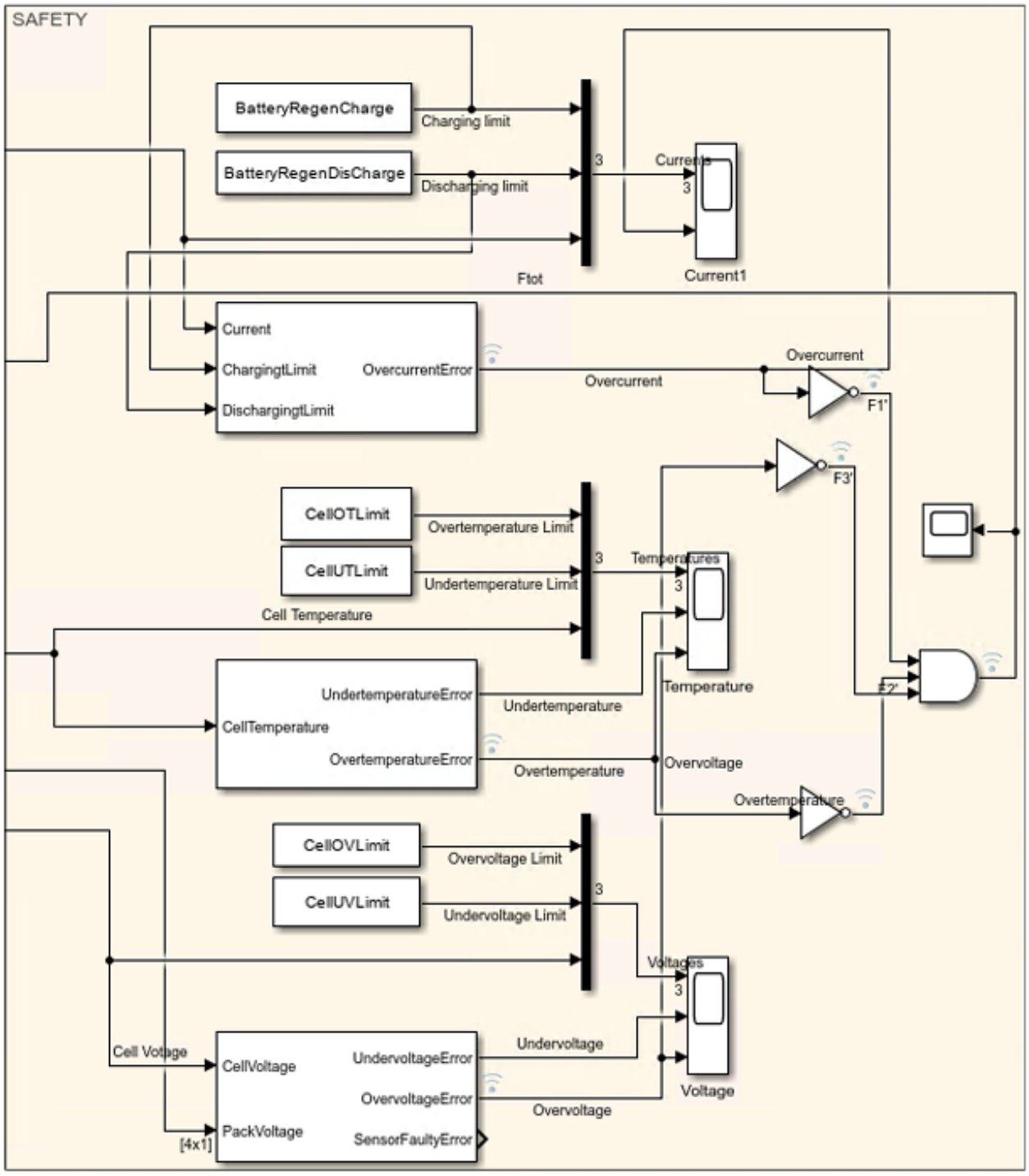


Figure 6: Safety Implementation of a Li-Ion battery

4 Results

We present below the results of temperature monitoring. As the temperature exceeds the overtemperature limit, we open the switching elements and temperature decreases to the initial value. The effect is seen on the current behavior where a drop is observed due to the open circuit.

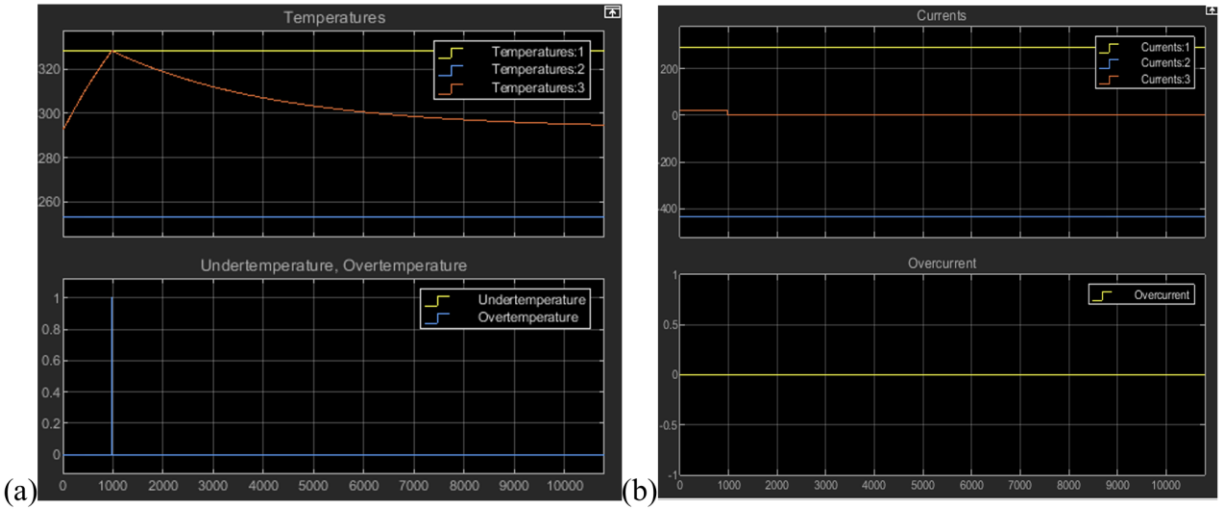


Figure 7: (a)Temperature monitoring and over/under temperature safety limits. (b) Current limit monitoring

This kind of behavior can be observed for all the battery subsystems, including voltage monitoring and charging. We use this methodology to size a battery for a specific application. The step that follows is creating in a CAD software the battery itself. Below we show the SOLIDWORKS drawing of the battery. In 3 different configurations. We see observe the electronic components on top of the Printed Circuit Board (PCB) as well as the cell configuration and the busbars.

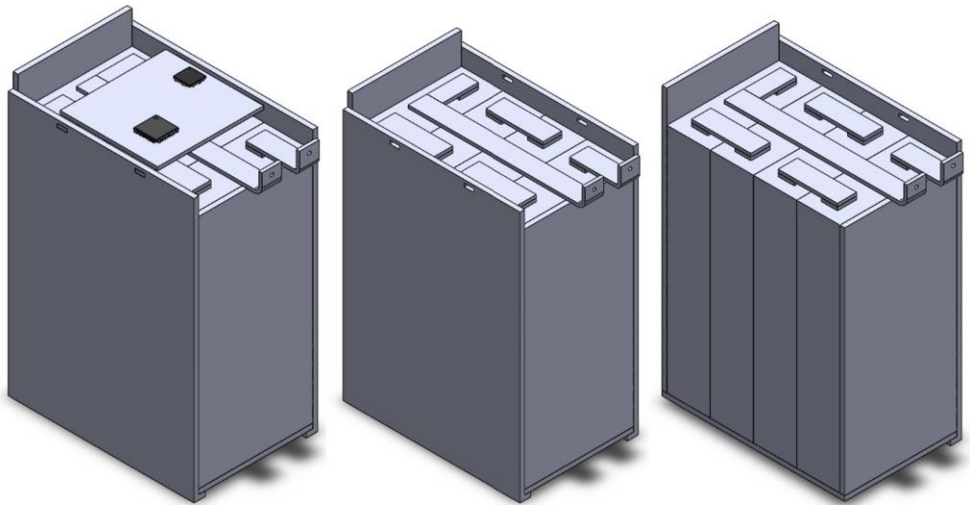


Figure 8: SOLIDWORKS design of the Lithium Ion Battery

5 Conclusions

With this work we presented a rapid prototyping approach. We use MATLAB-Simscape to insert the physical and electrical model of the cells in a specific configuration to form the battery module. We model the rest of the battery configuration by modeling the BMS functions as well as the battery charger behavior. By simulating the whole system, we can extract its behavior and use it to carefully size the battery components. We present the final step of this process by designing the key aspects of the battery with CAD software.

6 Acknowledgements

This project was partially funded by SHB (Smart Healthy Batteries) project from Eastern Macedonia & Thrace region as well as IPCEI EuBatIn project. We would like to thank MathWorks for their support during this work. We thank multiple Sunlight Group colleagues for providing various information for these simulations.

7 References

[1] Chrysovalantou Ziogou, Dimitris Ipsakis, Costas Elmasides, Fotis Stergiopoulos, Simira Papadopoulou, Panos Seferlis, Spyros Voutetakis, Automation infrastructure and operation control strategy in a stand-alone power system based on renewable energy sources. Journal of Power Sources Volume 196, Issue 22, 15 November 2011, Pages 9488-9499

[2] Damian Giaouris, Athanasios I. Papadopoulos, Chrysovalantou Ziogou, Dimitris Ipsakis, Spyros Voutetakis, Simira Papadopoulou, Panos Seferlis, Fotis Stergiopoulos, Costas Elmasides, Performance investigation of a hybrid renewable power generation and storage system using systemic power management models, Energy 61 (2013) 621e635

[3] Asimina Dimara, Christos Sougles, Sotirios Athanasiou, Konstantinos Grigoropoulos, Panagiota Sfakianou, Alexios Papaioannou, Stelios Krinidis, Dimitrios Triantafyllidis, Ioannis Tzitzios, Christos Nikolaos Anagnostopoulos, Aristoklis Karamanidis, Vaia Saltagianni, Dimosthenis Ioannidis, Dimitrios Tzovaras, Holistic plug-n-play autonomous solar system integration: a real-life small-scale demonstration—a practical approach, Electrical Engineering (2023)

[4] Hegazy Rezk, A. G. Olabi, Tabbi Wilberforce, Enas Taha Sayed, A Comprehensive Review and Application of Metaheuristics in Solving the Optimal Parameter Identification Problems, Sustainability 2023.

[5] Min Chen, Gabriel A. Rincon-Mora, Accurate Electrical Battery Model Capable of Predicting Runtime and I–V Performance

[6] Languang Lu, Xuebing Han, Jianqiu Li, Jianfeng Hua, Minggao Ouyang, A review on the key issues for lithium-ion battery management in electric vehicles

[7] UL, https://www.ul.com/services/functional-safety

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 39 – 42

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

S2W-Med Health Status Monitoring

N. Sionas, G. Pomportsis. P. Christodoulou, I. Tzamtzis*

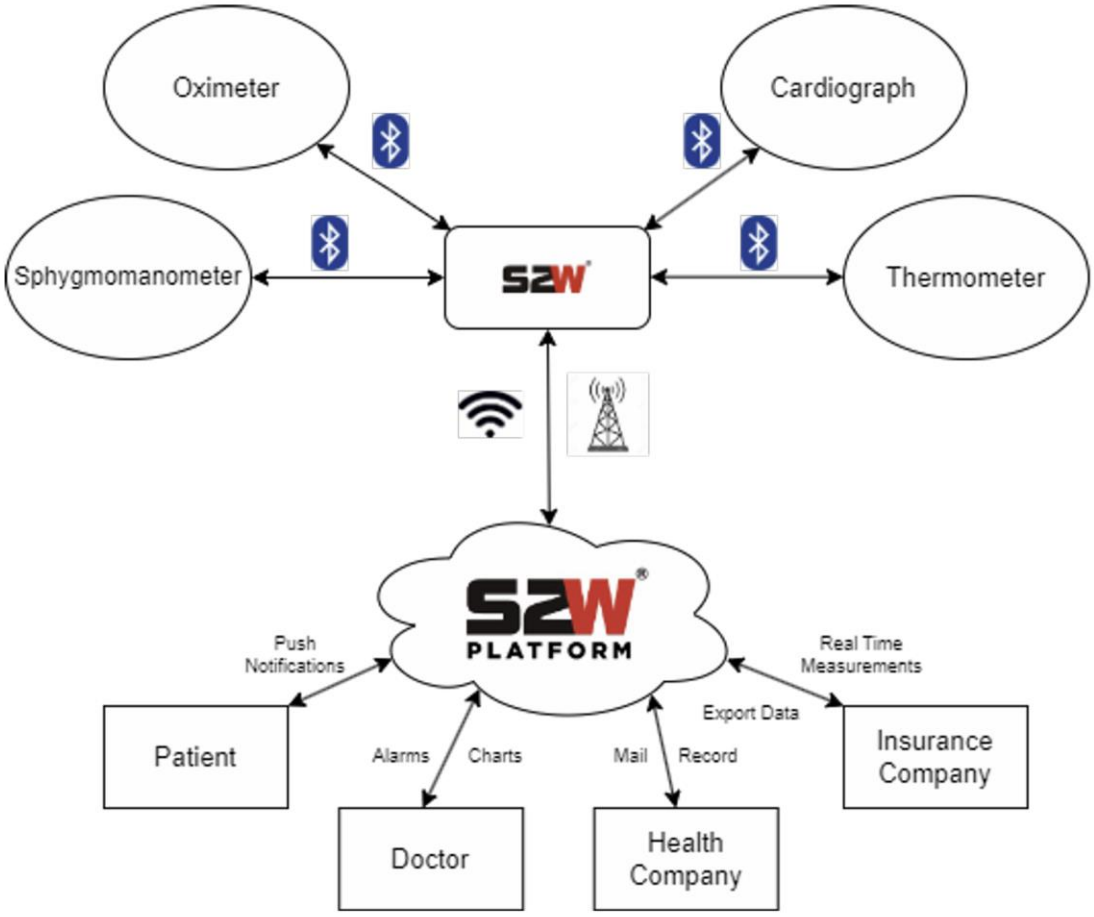
ELBIS PC, 14th km Thessalonikis – Moudanion ,Thermi – Thessaloniki, 57001, Greece
** Corresponding author. Tel.: +30 216 900 2750, E-mail address: itzamtzis@elbis.gr*

Abstract

It is common for patients who suffer from chronic health problems or are in a period of recovery to need constant monitoring of their body's vital signs by their personal doctor. Such a situation is time-consuming and costly for both the patient and the attending physician. S2W-Med that is presented in this paper is a system for monitoring patient's vital signs in a home environment. Its purpose is to gather the measurements from a group of medical devices so the attending physician has the results available in real time.

1 Introduction

The basic system devices are sphygmomanometer, oximeter, cardiograph and thermometer. The devices have the ability to send data via Bluetooth to a central device (S2W). S2W collects the data from all devices and sends them to the cloud via a GSM network. From the API of S2W-Med we monitor the measurements both through the Dashboard S2W application and the mobile application. In essence, Med-S2W upgrades existing tools and gives them IoT capabilities.



Basic Block Diagram

2 System description

2.1. Cloud infrastructure

The cloud infrastructure implements data management and storage methods and is also responsible for making the data accessible to users through a programming interface (API). The connection of the S2W device to the cloud infrastructure is achieved using a GSM network. The functionality of S2W allows to read values of sensors to which it is connected. To transmit and send data and settings to S2W devices, we use the MQTT protocol. The cloud infrastructure allows us to keep an overview of the device using shadows. Shadows are a representation of the last reported state of the device which includes the last settings we may have sent for the device to receive. Data is sent and received in JSON format. In particular, when sending the data from the device to the cloud, we use the SenML standard, which allows us to limit the size of the data transmitted. The general idea of SenML is that each value included in the JSON object sent is the result of subtracting the value from the first value of the object. The process is followed in reverse when the data reaches the cloud infrastructure.

2.2. Backend

We developed the backend in python's Django framework which is responsible for collecting and maintaining sensor measurements.

The backend (API) is a system consisting of smaller subsystems which are: MQTT broker, lambda functions, Time Series Database and HTTPS server. Lambda functions filter and store the data in the Time Series Database, where CRUD (create, read, change, and delete) operations can be performed on it. The HTTPS server application, which is also the system seen from the user application side, can perform operations from the database.

2.3. Frontend

It was developed using the ReactJS library. Also, to maintain a global state of the application, the Redux library was used in combination with the Redux-Sagas library. The purpose of the application is to monitor the values of the sensors as well as apply settings to them, this is achieved by calls to API-Django.

2.4. Application for mobile devices

A mobile application has been designed for iOS and Android.

Some of its features are:

- Real-time measurements.
- Setting measurement limits for each device
- Indication when exceeding measurement limits on the device
- Device status indication (connectivity, battery capacity, etc)

3 Development Tools

The tools used for the project development were:

3.1. *Software Tools*

- ESP-IDF (Programming interface for esp32)
- Django (Backend framework)
- Celery (Task queuing / event transfer from Lambda to Backend)
- Boto 3 (Amazon library for the interconnection of AWS cloud with Django)
- Influx DB (Time series database to store measurements and events)
- Docker (Virtual machine)
- ReactJS (JavaScript framework for frontend app)

3.2. *Development environments*

- Eclipse IDE (Programming environment for writing esp code in C language)
- VS-CODE, (Editor for backend and frontend development)

3.3. *Infrastructure*

- AWS-IOT (Amazon MQTT server for data transfer from device cloud)
- AWS-LAMBDAS (Serverless functions for data transfer from AWS-IOT to Backend)
- AWS-RDS (Cloud database to store settings and users)
- AWS-EC2 (Cloud host)

4 Conclusion

The S2W-Med system was created to fill a gap in patient's surveillance.

On one hand, the patient with a small cost feels the security of being constantly monitored by his doctor and even from the comfort of his home, without requiring frequent visits, thus saving time and money.

On the other hand, the doctor has a more complete picture of the patient's condition. Without burdening his time with continuous visits, he receives the indications in his office, automatically creates his patient's record and with notifications and alarms has immediate information about his patient's condition so that he can intervene when necessary.

Also, apart from the four basic devices (sphygmomanometer, thermometer, oximeter, cardiograph), new medical devices are being tested and integrated in the S2W-Med system in order to make the surveillance of a patient's condition more complete and also become a monitoring tool for more medical specialties.

5 Conclusions

With this work we presented a rapid prototyping approach. We use MATLAB-Simscape to insert the physical and electrical model of the cells in a specific configuration to form the battery module. We model the rest of the battery configuration by modeling the BMS functions as well as the battery charger behavior. By simulating the whole system, we can extract its behavior and use it to carefully size the battery components. We present the final step of this process by designing the key aspects of the battery with CAD software.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 43 – 49

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Fish Shape Alignment Based on Deformable Shape Tracking Suite of Tools

Nikos Petrellis[1], Georgios Keramidas[2], Christos Antonopoulos[1] and Nikolaos Voros[1]

[1] *Electrical and Computer Engineering, University of Peloponnese, Patras, Greece*
[2] *School of Informatics, Aristotle University of Thessaloniki, Greece*
npetrellis@uop.gr, gkeramidas@csd.auth.gr, ch.antonop@uop.gr, voros@uop.gr

Abstract

Fish morphological feature extraction based on shape alignment is used to estimate the dimensions, detect malformations, locate body parts like the eyes or the gills, classify fish orientation and species, etc. The Ensemble of Regression Trees machine learning approach is employed and specifically, the Deformable Shape Tracking package is adapted for fish shape alignment. Fish farms can benefit from the proposed approach for fish health assessment, fish growth and feeding needs estimation and harvest time selection. It can also be used for fish monitoring in open seas. Eighteen (18) landmarks are used to define the shape of a fish with an accuracy of approximately 95%. The training and testing were conducted using a custom dataset with low quality underwater images displaying seabream fish. The novelty of this approach is that the customized DEST package is implemented on the reconfigurable hardware of an embedded platform to support hardware acceleration and real time operation.[2]

1 Introduction

The morphological features that have to be measured daily in an aquaculture include the fish size, its mass, eye diameter, eye and gill color. The fish size and mass can determine when the harvest should be performed as well as the feeding needs. The eye and gill features as well as malformations in the shape of a fish can provide indications about its health, the welfare, etc. Measuring fish dimensions, weight, etc, was until recently performed manually, in an invasive way since fish had to be taken out of the water. The shape can also be used to track fish and interpret its behavior, classify fish species, observe fish populations and these procedures are also important for fish monitoring outside fish cultures (e.g., in rivers or the open sea).

A review of computer vision applications for aquacultures can be found in (Zion, Computers and Electronics in Agriculture, 2012). Applications for fish and egg counting, size measurement, estimation of the fish mass and gender, identification of species and for fish behavior monitoring, are presented in

[2] This work has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No 872614 -SMART4ALL: Self-sustained Cross-Border Customized Cyberphysical System Experiments for Capacity Building among European Stakeholder

(Zion, Computers and Electronics in Agriculture, 2012). Estimation of fish freshness from photographs and videos captured in a lab environment with controlled light exposure is reviewed in (Franceschelli et al, MDPI Sensors, 2021). The various sensors (biosensors, electric nose and tongue, colorimetric sensor array, spectroscopy, etc) that can be used for this purpose are presented in this paper. Fish classification in 12 classes based on Fast Regional-Convolutional Neural Networks (Fast R-CNNs) is described in (Li et al, Oceans, 2015). In (Lekunberri et al, Ecological Infromatics, 2022) various tuna species are counted and classified from image frames that display fish on a conveyor belt with 70% accuracy. Mask Regional-Convolutional Neural Network (Mask R-CNN) (He et al, 2023) and ResNet50V2 neural network architectures are employed to measure tuna fish sizes ranging from 23cm to 62cm. Low resolution images are also used in (Sun et al, CISP-BMEI, 2016) for fish recognition with 78% precision. Sonar imaging is used in (Martignac, Fish and Fisheries, 2015) for fish morphology and swimming behavior estimation. A size estimation error ranging from 2%-8% was measured for fish with size between 40cm and 90cm. The fish size and shape are measured using 18 landmarks in (Alsmadi et al.,Journal of Computer Science, 2010) using 3-layer artificial neural network.

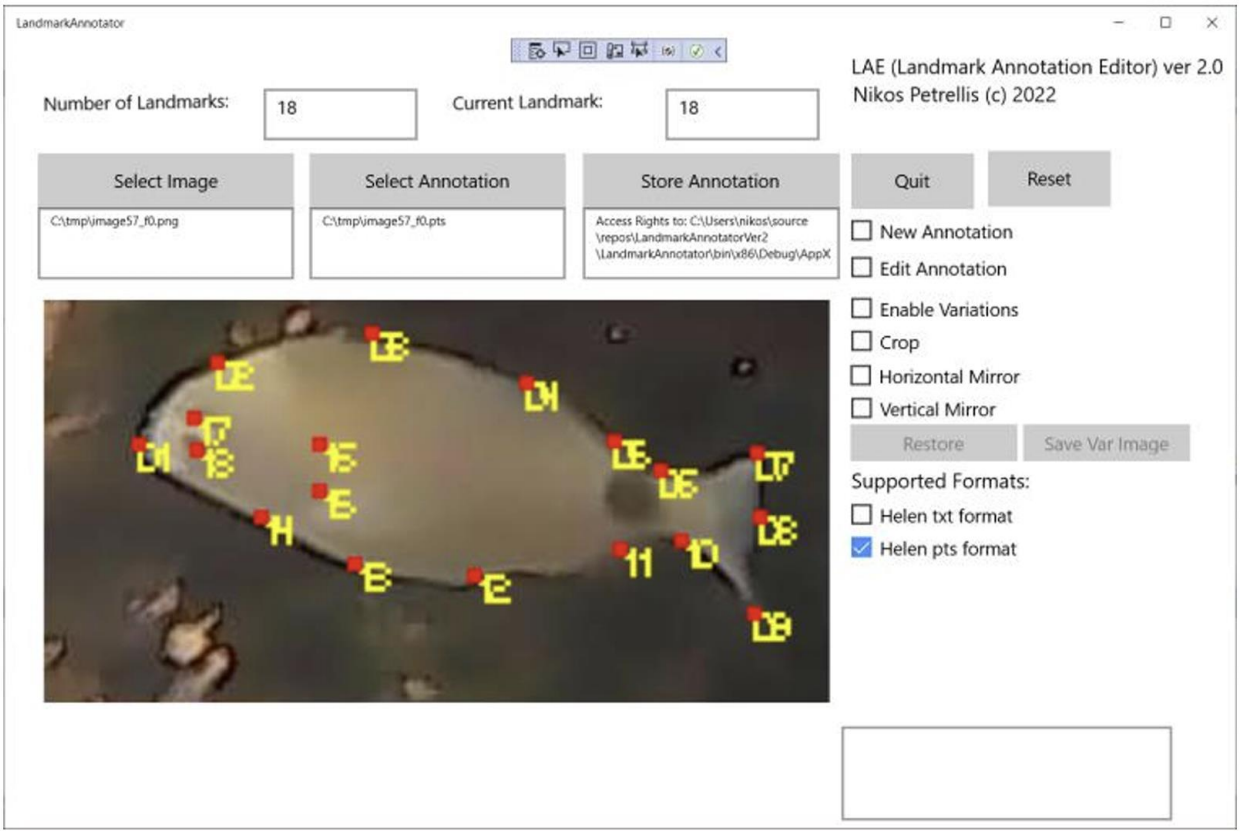


Figure 1: Fish landmark annotation in the LAE editor

In the approach described here, it is assumed that the fish is already detected and the bounding box of the fish is available as a separate patch from the frame it was derived. Fish detection methods like the one described in (Fish Detection, 2023) can be used for this purpose. Moreover, it is assumed that all the fish that are subject to shape alignment have the same orientation. The fish detection model of

(Fish Detection, 2023) can be trained to recognize e.g., only fish in a horizontal position facing at a specific direction. A lightweight shape orientation classification based on Principal Component Analysis (PCA) can be used as described in (Lendave, 2023) to classify the orientation of the fish shape. Then, the corresponding shape alignment model trained for this orientation can be employed for shape alignment.

The shape alignment applied on the image patch derived from the bounding box of the detected fish, is based on the machine learning (ML) approach called Ensemble of Regression Trees (ERT), presented by Kazemi and Sullivan in (Kazemi and Sullivan, CVPR, 2014) and is exploited in the Deformable Shape Tracking (DEST) suite of applications (DEST, 2023). The DEST package has been exploited for driver drowsiness applications in our previous work (Petrellis et al,VLSI-SoC, 2021). The source code of the DEST library was ported to Ubuntu host computer, Microsoft Visual Studio 2019 as well as Xilinx embedded Field Programmable Gate Array (FPGA) platform to support hardware acceleration. In another previous work (Petrellis, MDPI Appl. Sci., 2021) fish morphological feature measurement was performed for different fish species, using a different approach.

The contribution of this work can be summarized as follows: a) ERT has been used to align 18 landmarks on the fish shape, b) different ERT models can be trained for different fish orientations for more precise alignment, c) hardware acceleration techniques can be directly applied in shape alignment in order to support real time video processing, d) a new Landmark Alignment Editor (LAE) tool has been developed, e) a new dataset has been created and will be made public.

This paper is organized as follows. The ERT background and the use of the DEST package for training and testing fish shape alignment are discussed in Section 2. The experimental results and discussion are presented in Section 3. The conclusions of this work can be found in Section 4.

2 ERT Background and DEST Adaptation

The Ensemble of Regression Trees (ERT) as described in (Kazemi and Sullivan, 2014) is a shape alignment ML method originally developed for facial shape alignment and applications such as driver drowsiness detection, face expression recognition. In this paper, ERT is adapted for fish shape alignment. The ERT method is applied in a fish bounding box. The fish detection method described in (Fish Detection, 2023) is used to get the coordinates of the bounding box of a fish before applying the ERT method. Fish shape is not symmetrical and therefore the fish should have a specific orientation. The fish is allowed to have a tilt of about ±20 degrees. The employed fish detection method can be trained to recognize only fish in a specific direction, otherwise orientation classification methods should be used to verify that a detected fish has an acceptable orientation. In the rest of this paper, we focus on the fish shape alignment stage performed by the ERT method.

In ERT ML method, T_{cs} cascade stages are visited in order to gradually correct a mean shape stored in the trained ERT model. The correction is based on the comparison of the gray level of pairs of reference pixels belonging to a sparse representation of the input image. The default sparse representation of the input image consists of R_p=600 pixels. In face alignment applications, L=68 landmarks are used. In the fish alignment performed in this paper we use L=18 landmarks defined at the positions shown in Fig. 1.

In each cascade stage, T_{rg} binary regression trees with 2^{td}-1 nodes, are visited. In each node the difference in the gray level of a specific pair of reference pixels is compared to a threshold T_h and either

the right or the left child node of the regression tree is followed according to whether the difference is higher or not than T_h. For each node of the regression tree, the threshold T_h, the indices of the reference pixels that have to be compared and the next nodes that have to be visited are all stored in the trained ERT model. When a leaf is reached in the regression tree, a correction factor is found that is added to the current position of the landmarks. If the shape S∈ R^{2L}, is defined as the set of the xi pair of

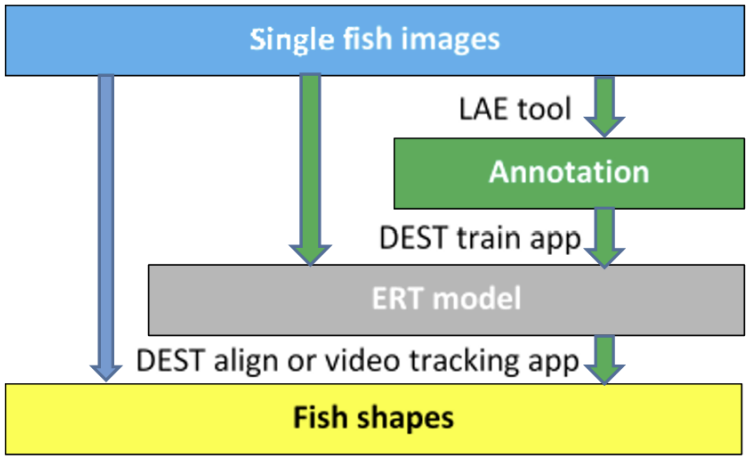


Figure 2: Training and testing fish shape alignment

coordinates: $S = \{x_0, x_1, \ldots x_L\}$, the ERT algorithm reads the mean shape *S* from the trained model. The, current shape estimate in the regressor t is $\hat{S}^{(t)}$ (t=1,.., T_{cs}) and the transition to the next cascade stage, requires the estimation of a correction factor r_t. This correction factor rt is added to $\hat{S}^{(t)}$ in order to generate the updated shape in the next regressor $\hat{S}^{(t+1)}$. As already mentioned, the r_t value is determined at the leaves of the regression trees that are visited. The regression factors r_t are determined during training using a gradient tree boosting algorithm with a sum of square error loss. During training, the values of rt are determined by the triplet: $\left(I_{\pi_i}, \hat{S}^{(t)}, \Delta\,\hat{S}_i^{(t)}\right)$ where I_{π_i} , $0 \le \pi_i < N$ is the image π_i in the set of N training images and $\hat{S}_i^{(t)}$ is the mean shape derived from all training image with $i \neq \pi_i$. If S_{π_i} is the shape in the π_i image, the residual $\Delta\,\hat{S}_i^{(t+1)}$ in the regressor r_{t+1} is estimated as: $\Delta\,\hat{S}_i^{(t+1)} = S_{\pi_i} - \hat{S}_i^{(t+1)}$. More details can be found in (Kazemi and Sullivan, 2014).

The procedure followed to train and test the fish shape alignment performed from the adapted DEST tools is shown in Fig. 2. During training the Landmark Annotator Editor (LAE) that has been developed, is used to annotate the L=18 landmarks in the images generated by the bounding boxes that display a single fish (see Fig. 1). In the LAE editor, the number of landmarks can be defined and a new annotation can start or an existing one can be modified easily by moving landmarks with a single mouse click. The annotations can be stored in a format compatible with the DEST tools. The LAE editor also offers dataset augmentation services (image mirroring, cropping). The resulting (image, annotation) pairs can be split in training and test images.

The DEST suite of tools (Deformable Shape Tracking, 2023) is developed in C++ exploits the Eigen math

library and contains applications for ERT model training (dest_train), shape alignment in single photographs (dest_align), shape alignment in video frames (dest_video_tracking), etc. This software package has been ported to Ubuntu Intel i5/i7 environment as described in (Petrellis et al, 2021) for a driver drowsiness monitoring application. From Ubuntu it was easy to port applications like dest_video_tracking to the Xilinx Vitis environment in order to accelerate in hardware the frame processing rate on an embedded platform (Xilinx ZynqMP Ultrascale+, ZCU102). In the context of porting DEST to Ubuntu, software was restructured and time consuming Eigen calls were replaced by optimized high speed C code achieving a software acceleration in the order of 240 times. Hardware acceleration on the other hand, leads to a further reduction of frame processing latency by 60%. In the context of this work, the DEST suite of tools has also been ported to Microsoft Visual Studio 2019 for operations in Windows environment. In fish shape alignment applications with real time constraints, the accelerated platforms in Ubuntu and ZynqMP Ultrascale+ can be exploited for high speed frame processing. With hardware acceleration and only 18 landmarks as is the case of the fish shape alignment described in this paper, the frame processing latency is estimated between 10 and 15 ms

The application dest_train has been used to train the ERT model for fish shape alignment based on the dataset created by the LAE editor. The trained ERT model can be used in real time when an input image frame is analyzed. The fish shape is determined by the L=18 landmarks that are defined as follows (see Fig. 1): landmark No. 1 is the fish mouth, landmarks 2-5 are the upper part of the contour, landmarks 6-10 are defined in each salient of the caudal fin, landmarks 11-14 make the bottom part of the contour, landmarks 14-16 denote the position of the gill and finally, landmarks 17-18 denote the location of the fish eye.

The position of the eye and the gills can be used to analyze the color of these body parts for disease diagnosis. The fish shape and its malformations can also reveal information about the fish health, variety, welfare, growing conditions, etc. The distance between landmarks 1 and 8 can be used to estimate the relative length of the fish while landmarks 3 and 13 can be used to estimate the height.

3 Experimental Results-Discussion

The ERT model was trained using 300 fish photographs. The orientation of the fish was horizontal, facing left. The contrast of the fish images was low as shown in Fig. 1. The images were selected on purpose with low quality in order to test the developed shape alignment method under worst case conditions. Concerning the resolution of the dataset images, the width ranges between 60 and 400 pixels, while the height between 20 and 300 pixels. The test set consisted of 100 fish images. Seven ERT models have been trained as shown in Table 1.

Model	T_{cs}	T_{rg}	Frame Process Latency (MS VS2019)	Normalized error	Fish Length estimation err.
M1 (default)	10	500	344 ms	4.81%	6.09%
M2	8	500	281 ms	5.16%	6.69%
M3	10	400	310 ms	4.81%	6.11%

M4	8	400	264 ms	4.91%	6.22%
M5	6	400	156 ms	6.21%	6.40%
M6	8	300	211 ms	5.06%	6.42%
M7	6	300	200 ms	6.41%	7.50%

Table 1: Comparison in the speed and error of the trained models

In the default model M1, T_{cs}=10 cascade stages have been employed with T_{rg}=500 regression trees in each stage. The cascade stages have been reduced to 8 in M2 while the regression trees have been reduced to 400 in M3. Both cascade stages and regression trees have been reduced to 8 and 400, respectively, in M4. The cascade stages have been further reduced to 6 in M5 along with 400 trees. In M6, 8 cascade stages have been used with only 300 trees in each cascade stage. Finally, in M7, 6 cascade stages with 300 trees are used. The frame processing latency is the one listed in the 4th column of Table 1 and is measured in a Microsoft Visual Studio 2019 environment. The frame processing latency is almost proportional to T_{cs}·T_{rg}. In an Ubuntu Intel Core i5-9500 CPU @3.00GHz, 6 core processor with 16GB RAM environment, the frame processing latency for M1 is less than 0.4us while this latency for M7 is less than 0.2us. On an embedded ZCU102 platform the corresponding latencies for M1 and M7 are less than 15ms and 8ms, respectively. The normalized relative error is estimated as the mean relative Euclidean distance between the estimated landmark position and the one annotated with LAE as ground truth. In the estimation of the normalized error the width and height distances are divided by the image width and height, respectively. As can be seen from the 5th column of Table 1, the normalized error in M3 is the same with M1 although the M3 latency is 10% shorter than M1. The mean error in M6 (5%) is quite close to M1 while its latency is 39% smaller. The estimation error in the fish length from landmarks No. 1 and 8, is listed in the last column of Table 1, for each ERT model. Compared with referenced approaches, (Lekunberri, et al., 2022) estimate tuna size with a standard deviation between 0.328 and 0.396, while (Martignac et al., 2015) estimate fish length with an error between 2% and 9%. Generally, we can state that our approach can achieve a high accuracy with a very low latency.

4 Conclusion

The DEST implementation of ERTs was adapted for fish shape alignment and morphological feature extraction. The position of 18 landmarks was estimated with an accuracy of about 95%. Future work will focus on the incorporation of the developed method for morphological feature estimation in a framework where automatic fish detection and tracking will also be supported.

5 References

[1] Alsmadi, M. K., Omar, K. B., Noah, S. A., Almarashdeh, I. (2010), Fish Recognition Based on Robust Features Extraction from Size and Shape Measurements Using Neural Network. Retrieved from https://doi.org/10.3844/jcssp.2010.1088.1094.

[2] Deformable Shape Tracking (DEST) (2023, June). Retrieved from https://github.com/cheind/dest.

[3] Fish detection (2023, June). Retrieved from https://github.com/kwea123/fish_detection.

[4] Franceschelli, L., Berardinelli, A., Dabbou, S., Ragni, L., and Tartagni, M. (2021). Sensing Technology for Fish Freshness and Safety: A Review. MDPI Sensors 21, (p. 1373). Retrieved from https://doi.org/10.3390/s21041373

[5] He, K., Gkioxari, G., Dollár, P., Girshick, R. (2023). Mask R-CNN. Retrieved from https://arxiv.org/abs/1703.06870.

[6] Kazemi, V., and Sullivan, J. (2014, June). One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 867-1874). Retrieved from 10.1109/CVPR.2014.241.

[7] Lekunberri, X., Ruiz, J., Quincoces, I., Dornaika, F., Arganda-Carreras, I., Fernandes, A.A. (2022). Identification and measurement of tropical tuna species in purse seiner catches using computer vision and deep learning. Ecological https://doi.org/10.1016/j.ecoinf.2021.101495.

[8] Lendave, V. (2023, June). Detecting Orientation of Objects in Image using PCA and OpenCV. Retrieved from https://analyticsindiamag.com/detecting-orientation-of-objects-in-image-using-pca and-opencv/

[9] Li, X., Shang, M., Qin, H., and Chen, L (2015). Fast accurate fish detection and recognition of underwater images with Fast 1R-CNN. Oceans (pp. 1-5). Retrieved from 0.23919/OCEANS.2015.7404464

[10] Martignac, F., Daroux, A., Bagliniere, J.L., Ombredane, D., and Guillard, J. (2015). The use of acoustic cameras in shallow waters: new hydroacoustic tools for monitoring migratory fish population. A review of DIDSON technology. Fish and Fisheries 16 (pp. 486–510). Retrieved from https://doi.org/10.1111/faf.12071

[11] Petrellis, N., Christakos, P., Zogas, S., Mousouliotis, P., Keramidas, G., Voros, N., Antonopoulos, C. (2021). Challenges Towards Hardware Acceleration of the Deformable Shape Tracking Application. In Proceedings of the 2021 IFIP/IEEE 29th International Conference on Very Large Scale Integration (VLSI-SoC). Retrieved from 10.1109/VLSI-SoC53125.2021.9606999.

[12] Petrellis, N. (2021). Measurement of Fish Morphological Features through Image Processing and Deep Learning Techniques. MDPI Appl. Sci. 11 (p. 4416). Retrieved from https://doi.org/10.3390/app11104416.

[13] Petrellis, N. (2021). Measurement of Fish Morphological Features through Image Processing and Deep Learning Techniques. MDPI Appl. Sci. 11 (p. 4416). Retrieved from https://doi.org/10.3390/app11104416.

[14] Sun, X., Shi, J., Dong, J., Wang, X. (2016). Fish recognition from low-resolution underwater images. In Proceedings of 9th IEEE International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (pp. 471–476). Retrieved from https://doi.org/10.1109/CISP-BMEI.2016.7852757

[15] Zion, B. (2012). The use of computer vision technologies in aquaculture – A review. Computers and Electronics in Agriculture 88 (pp. 125-13). Retrieved from https://doi.org/10.1016/j.compag.2012.07.010.

EMERGING TECH CONFERENCE 2023
Edge Intelligence

Papers

Session 1.3 | Hardware design

Session Chairs: Gregory DOUMENIS

Design and Implementation of Bone Wax Application Device
Michail Tsilikas and Dimitrios Papakostas

Towards a modular IoT device design and prototyping for the Sports domain
Evripidis Chondromatidis, Emmanouil Tsardoulias, Konstantinos Panayiotou and Andreas Symeonidis

High-Performance Design of SRT IP Blocks
Aristotelis Tsekouras, Giorgos Stagakis, Anastasis Avgoustidis, Grigoris Kokkonis, Konstantinos Gkekas, Vasilis Pavlidis, Thomas Noulis and Giorgos Keramidas

Fractional-N Phase Locked Loop for Wi-Fi 6/6E With 135 mW Power Consumption and Spur Reduction Techniques
Savvas Sgourenas, Christos Andriakopoulos, Stefanos Pokamisas, Charis Basetas, Chrysa Vassou, Vasilis Tsamis, Kostas Retsinas, Vasilis Kolios, Nektarios Sgourenas and Giorgos Kasapoglou

Enhanced Safety Architecture with Fault Preventive Mechanisms for Automotive Li-ion Battery Management Systems
Apostolos Delizonas, Christos Mademlis, Evangelos Tsioumas, Dimitrios Papagiannis, Nikolaos Jabbour, Christos Sansaridis and Tilemachos Matiakis

Energy Efficient ML Accelerator using a Decoupled Vector Engine and a Systolic Array attached to a Low-Cost RISC-V Scalar Core
Giorgos Dimitrakopoulos, Vasileios Titopoulos, Christodoulos Peltekis, Dionysios Filippas and Kosmas Alexandridis

HETiA
EXPANDING THE DIGITAL FRONTIERS

ARISTOTLE UNIVERSITY OF THESSALONIKI

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 50 – 56

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Design and Implementation of Bone Wax Application Device

Michail Tsilikas[1*] and Dimitrios Papakostas [1+]

[1] *Department of Information and Electronic Engineering International Hellenic University*
Thessaloniki Greece
michalis.tsilikas@gmail.com, dpapakos@ihu.gr

Abstract

The basis of this research paper is a Diploma Thesis and presents the process of designing and implementing an innovative device to combat bone haemorrhage. The device is based on a patented idea that proposes heating the hemostatic wax, which is already used in cases where bone haemorrhage needs to be addressed, and its application through a special device. The device was fabricated following a thorough examination of the existing procedures of bone haemorrhage hemostasis. After defining the key factors that needed to be considered so that this device could be innovative yet effective in various conditions, without compromising the patient's safety, the device was fabricated following a thorough examination of the existing procedures of bone haemorrhage hemostasis. The results demonstrate the device's efficacy in the experimentation phase while pointing out the room for improvement. Thus, the fabricated device could impact the approach to bone hemostasis treatment and could be applied to various fields, from orthopedic surgeries and neurosurgeries to combat and emergency medical aid.

1 Introduction

Technology brings forth novel advancements that push the boundaries of innovation and progress. Medical science employs these innovations to provide the optimum medical care, resulting in the least possible side effects. The evolution in both technology and medical science fields has been rapid over the past decades. The research delves into the theoretical background of haemorrhage treatment with the utmost goal of establishing the basis for implementing a device that acts as an assistive medical tool, improving the procedures and the quality of treatment.

Comprehending the bone's internal structure is crucial, as it facilitates a better grasp of the reasons behind the need for hemostasis. Being porous and traversed by veins, while being the source of production of the red blood cells, bones when ruptured, result in blood loss and bleeding. The implications of that blood loss can be increased healing time but can also be more severe when not treated accordingly. During surgeries, blood loss can also obscure the view of the inflicted areas and

[*] Master's Student at the Department of Information and Electronic Engineering, International Hellenic University

[+] Professor at the Department of Information and Electronic Engineering, International Hellenic University

increase the difficulty of the procedure.

The aim of the Diploma Thesis[1] was the research, design and development of a device for applying bone wax, based on the general idea that the initial patented idea [2] described. The main functions of the device should be the heating and the extracting of the bone wax to deal with the haemorrhage of bones.

2 Theoretical Background and Literature Review

Currently, there are several ways of addressing bone haemorrhage, using a unique mixture of beeswax and paraffin, forming the hemostatic bone wax[3], [4]. The medical personnel apply this mixture, called bone wax, by "heating" and forming the hemostatic bone wax, with their hands, to be malleable and ready to be applied to the wounded area. Therefore, it is time-consuming and results in an increased difficulty of application. Furthermore, the hemostatic bone wax prevents the complete restoration of bone tissue to the areas where it has been applied, introducing undesirable consequences.

The objectives of implementing a bone wax application device were:

- Easier application procedure
- Sterile device and hemostatic bone wax
- Portable and easy-to-use operatio
- Capabilities of increased accuracy of treatment with less material applied
- Ability to access difficult wounded areas
- Reusability
- Emergency response on site
- Employable under extreme conditions

Some characteristics, temperature of the bone wax, materials of the device, and portability, were also defined through the research to make the device easy to use and safe for the patients. Osteonecrosis [5] is the necrosis of bone tissue and in that case, a limiting factor that needed to be considered was the heat applied. Osteonecrosis occurs at 50oC and it is a temperature that should not be applied to the patient's body by the device. The device's sterilisation was considered, and specific materials were selected. The design includes removable parts, which can be sterilised before use and discarded.

3 Methodology

To achieve the aforementioned objectives and fabricate the final device Figure 2-C, the implementation of the device went through certain stages. First, the theoretical background had to be studied and considered as the primary source of providing the necessary information on how the device should operate. The theoretical background consisted of research in medical science and the applications and studies of hemostatic bone wax[4]. After that, the circuitry was experimentally designed only for controlled heating. The feedback was positive and after that, some experimental circuits were designed and adjusted accordingly, to come up with the final design that implements all

the necessary features. Then the casing and the extruding mechanism were designed in 3D CAD software. The next step was to 3D print all the necessary components and test fit while checking the functionality. Following that, the device was fabricated from all the individual components and then the code was composed to enable the device to work according to the needs defined by the research that had been done. To ensure optimal performance of the device, testing and experimentation were carried out. This involved measuring power consumption, amperage, and operation times. Based on the results, adjustments will be made in subsequent steps to enhance the device's functionality.

3.1. Circuitry Design

The requirements for ease of use, portability, heating and extrusion of the bone wax, provided the core for determining the design of the circuits. The need for portability resulted in choosing batteries as the power source, while the heating procedure set forth the need for batteries capable of providing high current. The battery of choice was a 26650 Li-ion battery, providing up to 30 Ampere at a nominal voltage of V_{Nom} = 3.7V. Hence, it was crucial to have a charging and Battery Management System (BMS). The ICs of choice to have optimal performance, while reducing size, were the TP4056 as the charging IC and the DW01 as the BMS.

The microcontroller responsible for performing the different tasks, calculations and temperature control, was the STM32g030f6P6 microcontroller and was chosen because of its capabilities (Pulse Width Modulation pins, Timers, etc.), the number of I/O (input/output pins) that it had and the input voltage requirements.

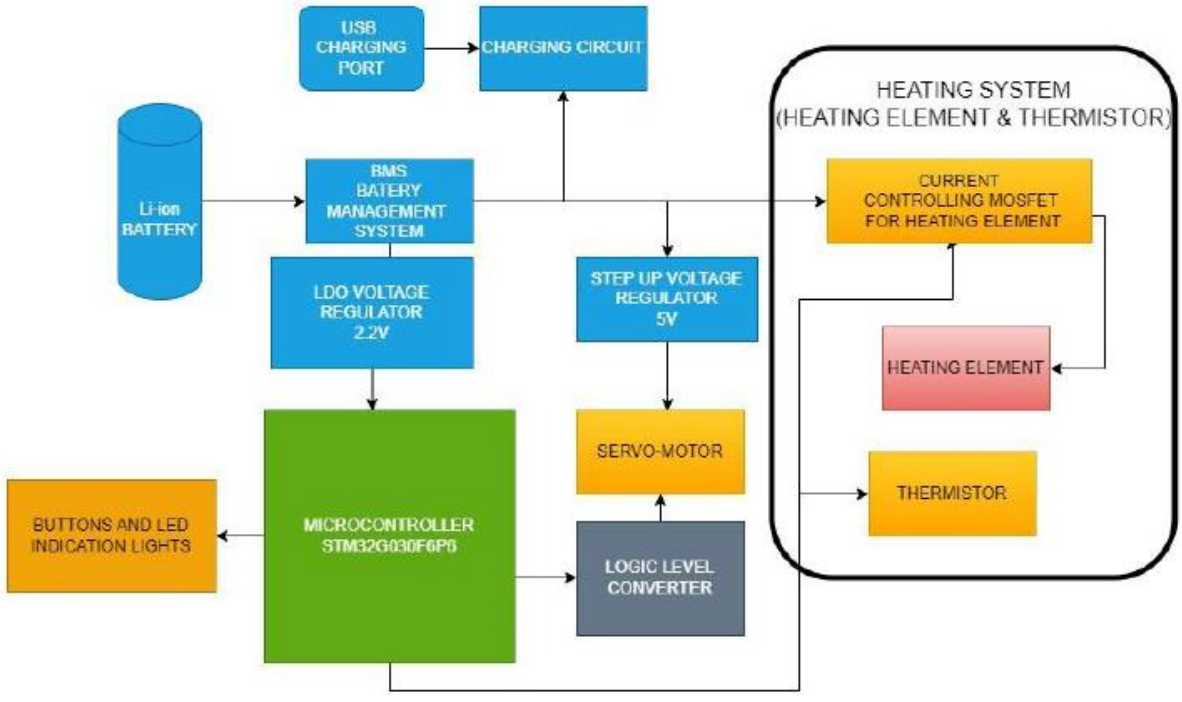


Figure 1: Block Diagram of the Circuit

A stable power supply to the microcontroller had to be considered for best performance and accurate temperature control, for that reason a circuit of low voltage dropout voltage regulator was added

providing 3.2V to the microcontroller. A step-up voltage regulator, with an output of 5V, was also added to provide the voltage to the servo-motor, which is responsible for the extruding procedure, which needed 4-6V. Having a low-power microcontroller meant that the logic level was shifted towards the 3V and a need for a logic level shifter emerged to communicate with the servo-motor that had a 5V logic level.

The high current requirement of the heating element was taken into account. Using a MOSFET the current can be regulated through the microcontroller. That way it was possible to use PID and PWM to control the heating procedure more accurately. The low voltage provided by the battery was crucial for selecting the MOSFET and the $V_{GSthreshold}$ should be close to the minimum voltage of the battery, which is 2.4V, so that the full potential of the battery can be used. The CSD16340Q3 was chosen and provides up to 20A of current with a $V_{GSthreshold} \approx 0.85V$, $I_{ds} = 250\mu A$.

3.2. Printed Circuit Board (PCB) Design and Fabrication

Following the completion of the theoretical circuit design, the subsequent stage was the fabrication of the circuit in its physical form. The PCB was crafted by utilising online PCB design software, a 3D model produced is shown in Figure 2-A, followed by printing and soldering of various components. Initially, the first versions of the PCB were printed for testing purposes. To check only the heating procedure, and various components and eventually to check for any problems. The final version was sent to an external manufacturer to be printed and after that, the PCB was hand-soldered with the use of a handmade solder plate.

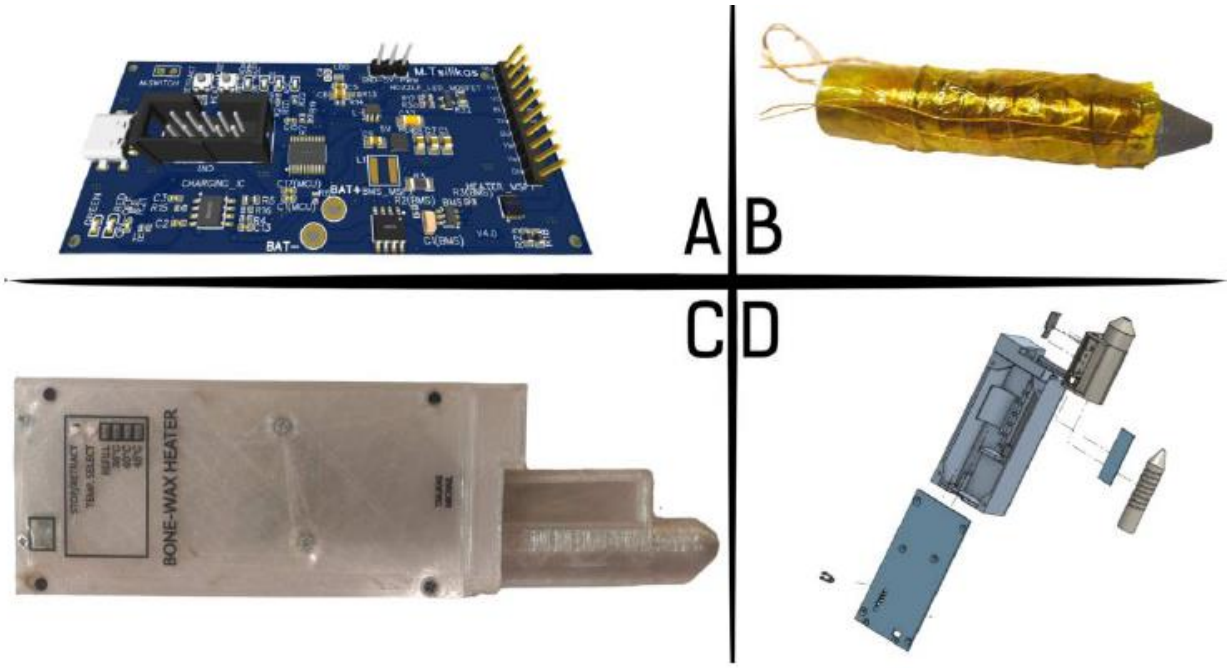


Figure 2: **A.** 3D Preview of PCB **B.** Heating Chamber **C.** The Final Device **D.** 3D Preview of Device`s Parts

Heating displacement through the copper pads was taken into consideration and dual-layer PCB was chosen for the prototype to decrease the size of the board and to make it more accessible for testing.

The components used were Surface Mount Technology (SMT) to decrease the size and produce a compact and efficient result.

3.3. 3D Casing Design and Mechanical Extrusion

To create a device that meets specific requirements, 3D CAD software was utilised to design a custom casing and mechanical extrusion mechanism. The device needed to house all the electronic components PCB, buttons, LEDs, battery and the extrusion mechanism with the servo-motor. These parts had to be snugly fit and secured in place to make the most out of the available space. Incorporating a detachable nozzle that holds the bone wax, heater element, thermistor, and extrusion button was a crucial aspect of the device's design. This feature enhances ease of use and promotes sterility. The chosen detachable mechanism uses a linear slot where parts slide and lock together through friction. Meanwhile, electrical connection happens through female and male connectors found on the two individual pieces that fit and slide together.

3.4. 3D Printing Parts

The materials selected are crucial to the construction and performance of the device. Considering heat and sterilisation, the device needed to be manufactured with strong, yet lightweight materials that can be sterilised easily. The best choice was aluminium for its strength, smooth and easily sterilised surface, as well as its self-sterilizing characteristics[6] and ability to conduct heat efficiently. However, during the prototyping phase and due to cost consideration, the material chosen was PET-G and only the container of the bone wax was manufactured with aluminium. All the parts shown in Figure 2-D were printed in a 3D printer except from the wax container that was sent to an external manufacturer to be 3D printed, but in aluminium.

3.5. Device Assembly

With all the essential components at hand, the assembly of the device is initiated. The SMT components were soldered to the final PCB and inserted into designated positions. PCB was fitted with soldered wires connecting the servo-motor, buttons, LEDs, battery and the female connector.

As shown in Figure 2-B the bone wax container was covered in Kapton tape to make it electrically non-conductive. After that the Nichrome wire, acting as the heating element, was wrapped around the container in prefabricated slots. The thermistor also had to be mounted by special non-electrically conductive heat-transferring glue. Along with one button and two inspection LEDs, all the components were soldered to the male connector and fitted to the detachable nozzle part. The final device as shown in Figure 2-C has dimensions of 21cm length, 6.1cm width and 4.2cm height.

3.6. Code Composition

The code was composed in the Cube IDE software for the microcontroller. In order not to halt the code execution, interrupt functions were used to detect the press of buttons and the proper function of the PID controller alongside the PWM generation. The code checks if the temperature selection button is pressed. If pressed, the button initiates the heating procedure (PID controller, PWM generation, temperature calculation). When the temperature reaches the desired temperature the corresponding LED lights up continuously declaring that the device is ready for extruding the bone wax. Using an interrupt, the device detects when the extrude or stop button is pressed and acts accordingly. To

change the detachable nozzle, the user needs to press the refill button, and then the extruding piston returns to the refill position and awaits the user's action.

Lastly, the code incorporates some safety features. The device checks for hazardous temperatures above 49°C and if it detects a temperature greater than 49°C stops all actions. In the design a magnetic sensor was placed on the main body of the device, detecting if the nozzle is present, and if it is not the device deactivates until the nozzle is placed in its proper position.

4 Results

Experimenting and assessing the device was done in a controlled environment, of 25°C and 38% humidity, to receive results that provide a clear understanding of the performance.

Time to heat in each of the three different predefined temperatures:

- 35°C → 20 seconds
- 40°C → 29 seconds
- 45°C → 50 seconds

To test the device in extreme conditions it was inserted in a freezer (-17°C) for 3 hours, so the inner core of the bone wax could reach the desired -17°C of the freezer, and then the heating procedure was initiated. At the same time, an oscilloscope and a temperature sensor monitored the temperature and the voltage levels. With the selected temperature being 45°C, the time the device needed to reach the desired temperature was 2 minutes and 35 seconds.

Additionally, an experiment of continuously heating at each of the heating temperatures was conducted to test the consumption of the device and its performance. As presented in Figure 3 the device starts drawing a current of around 3.5A until it reaches the selected temperature. After that, it draws small amounts of current, around 200mA to sustain the selected temperature.

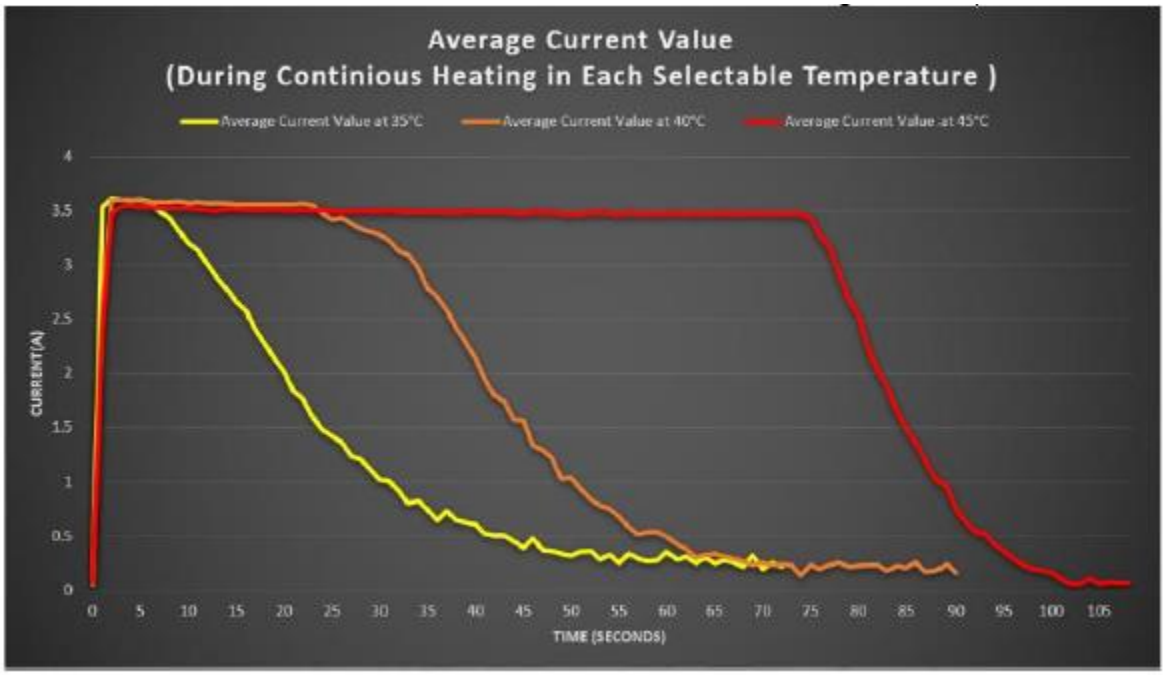


Figure 3: Continuous Heating, Average Current

Considering all of the above, it is concluded that the device is capable of performing the tasks that it was designed to do while being portable and able to operate when the temperature varies from -17°C to 40°C according to the tests.

There is room for improvement and the procedure of optimizing the device commenced. The final device will be using aluminium and will optimize the performance of the microcontroller to limit the current consumption while decreasing the size of the device.

5 Conclusion

The objective of designing and implementing a device to apply bone wax has been achieved. The device possesses the necessary features, which were determined through research on the medical use of bone wax for hemostasis. The device implemented is just a prototype, improvements and further research and clinical trials are needed to reach its final form.

The PCB will be improved by the use of 3 layers so that the copper traces will cover less distance, thus improving performance and freeing space for the components to be oriented more compactly.

The material of choice for the outer shell will be aluminium, decreasing the thickness of the outer walls and consequently, the size of the device, while having more strength and better sterilization properties.

The battery has a performance that is more than sufficient. However, by decreasing its capacity, the size of the final device would be reduced, which is crucial for creating a portable medical device.

As an ongoing process, research, trials, and experimentation provide ample feedback, allowing for continual improvement.

6 Bibliography

[1] M. Tsilikas, “Design and implementation of bone wax application device”, Diploma Thesis, Department of Information and Electronic Engineering, International Hellenic University, Sindos, Thessaloniki, 2023.

[2] C. Cholevas and S. Grigoriadis, “ΣΥΣΤΗΜΑ ΤΗΞΗΣ ΚΑΙ ΕΦΑΡΜΟΓΗΣ ΑΙΜΟΣΤΑΤΙΚΟΥ ΚΕΡΙΟΥ”, Oct. 2015. [As seen Online on 5/06/2023]. Available: http://www.obi.gr/wp-content/uploads/2022/06/EDBI_A_2015_09.pdf

[3] H. Zhou, J. Ge, Y. Bai, C. Liang, and L. Yang, “Translation of bone wax and its substitutes: History, clinical status and future directions”, *J. Orthop. Translat.*, vol. 17, pp. 64–72, Apr. 2019, doi: 10.1016/j.jot.2019.03.005.

[4] J. M. Das, “Bone Wax in Neurosurgery: A Review”, *World Neurosurg.*, vol. 116, pp. 72–76, Oct. 2018, doi: 10.1016/j.wneu.2018.04.222.

[5] K. Kniha, N. Heussen, E. Weber, S. C. Möhlhenrich, F. Hölzle, and A. Modabber, “Temperature Threshold Values of Bone Necrosis for Thermo-Explantation of Dental Implants—A Systematic Review on Preclinical In Vivo Research”, *Materials*, vol. 13, p. 3461, Oct. 2020, doi: 10.3390/ma13163461.

[6] S. V Gudkov, D. E. Burmistrov, V. V Smirnova, A. A. Semenova, and A. B. Lisitsyn, “A Mini Review of Antibacterial Properties of Al2O3 Nanoparticles”, *Nanomaterials*, vol. 12, p. 2635, Oct. 2022, doi: 10.3390/nano12152635.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 57 – 63

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Towards a modular IoT device design and prototyping for the Sports domain

Evripidis Chondromatidis, Emmanouil Tsardoulias, Konstantinos Panayiotou, and Andreas Symeonidis

School of Electrical and Computer Engineering, AUTH, Thessaloniki, Greece
evrichon@auth.gr, etsardou@ece.auth.gr, klpanagi@ece.auth.gr, symeonid@ece.auth.gr

Abstract

As the ever-growing demand of Internet of Things (IoT) devices continues, so does the necessity for rapid and systematic ways of designing their hardware and software. The above-mentioned issue becomes aggravated when the final requirements of a system cannot be determined a priori, especially when the system is expected to scale up in the future.

This work proposes a modular architecture. We discuss our approach in context of a device targeted in the Fitness vertical. Subsequently, we proceed by instilling these principles into a working prototype solution in the direction of evaluating its practical usage.

1 Introduction

The unprecedented collection and exchange of data that our Information Age offers, is largely driven by the emergence of compact and cost-effective devices, the so-called Edge & Internet of Things Devices. The total number of such devices is expected to triple from 9.7 billion in 2020 to more than 29 billion IoT devices in 2030 [1]. This undoubtedly shows that the industry already has and will continue to heavily invest in the direction of IoT devices, reaching USD 3,352.96 billions by 2030 [2].

Following the existing trend, many Sports Organizations employ IoT in order to tackle three fundamental objectives [3]: a) Player Development, b) Player Safety and, c) Fan Engagement. IoT devices can provide an abundance of data which when combined with advanced analytics can provide to Coaches suitable metrics in order to assess the performance and efficiency of their athletes in a much more accurate and unbiased manner. Moreover athletes are able to recover faster from injuries as well as improve their overall health conditions, as Sports Physicians and Team Doctors can continuously monitor their biometrics.

Despite the undisputed advantages that IoT technologies bring to the Sports domain, currently their large scale impact is hindered by a number of reasons. Most of the commercial IoT sport devices are quite expensive with regard to their hardware. Moreover, this issue is further enlarged by the fact that these solutions are closed-source. Licensing imposes restrictions on the number of users who can use the software on such devices. As a result, sports organizations are unable to adapt the products to their specific needs due to the special permissions required to do so. Furthermore existing IoT SportS devices are dedicated to execute specific tasks. This means that small modifications in the training protocol may demand new functionality, rendering existing devices obsolete. These reasons

allow only a limited percentage of the Sports Community to embrace these new technologies and magnify the skill gap in the Sports industry.

In the current work, the methodology of design, implementation and deployment of a modular IoT device is presented, with the aim of tackling all the aforementioned issues, dedicated to the sports domain. Our solution encapsulates not only the hardware design and assembly process, but also the software implementation using the powerful open source Riot-OS [7].

2 State of the Art

Currently there are many Smart Products targeting the Sports Industry, that are commercially available or developed for research applications. In general, these Products consist of smart devices that are capable of both transmitting stimulus via actuators to the athletes as well as receiving and recording their reactions using sensors. Such devices can either operate solely or they can be part of a larger interconnected System. Depending on their specific hardware, they can be incorporated in different exercise scenarios where they measure important physical and cognitive performance metrics [6]. Such scenarios could range from measuring the raw physical abilities of an athlete, such as strength and speed to performing physical abilities while following a series of evolving logical rules aiming to enhance the mind-body responsiveness [4]. The accumulated data are then transmitted either throughout a cable or wirelessly to a platform where they can be stored, processed and displayed. Fitlight[‡] & BlazePot[§] are two examples of systems that comprise wirelessly interconnected devices. The above mentioned devices contain a light emitting source that can be adjusted to different colors, a ToF sensor that can measure distance and a touch sensor. Remedes [6] is a system that comprises similar smart devices in a star configuration and can be utilized both in sports and medical fields, extending the reflection measurements process in the direction of detecting neurodegenerative diseases. Another paradigm of such Solutions is Chronojump [5], which is rather a collection of different Products. Most of them comprise a single device that is connected via a USB cable (serial communication) and a sensor unit, while recently a wireless solution for race analysis was added. They include a jumping pad so as to measure jump height, piezoelectric sensors in order to measure force and linear encoders which are attached to an athlete via a string so as to analyze their acceleration performance.

The aforementioned Sport products, undoubtedly constitute solid and well-tested solutions, however all of them pose the issue of specialization in regard to their hardware, which inherently limits their adaptability and overall effectiveness. Different teams have different requirements and one can argue that the same rule applies to team's members as well. Technology in the context of Sports promises to provide a more personalized experience. It is the technology that should adapt to the player's need and not vice versa. Therefore while the monolithic design approach may offer additional robustness to the end product, it contradicts its original purpose. In this work, a modular architecture is presented where the hardware of such devices is separated into functionally independent blocks.

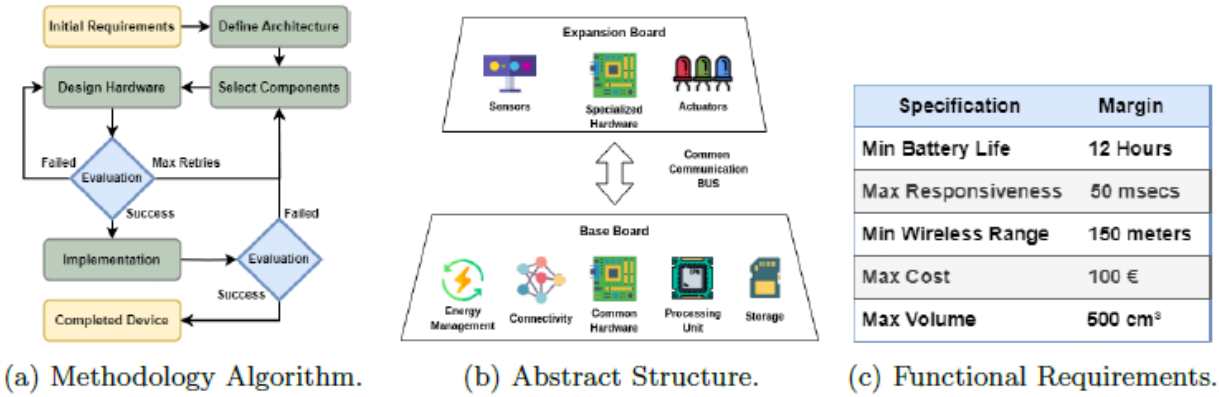
[‡] https://www.fitlighttraining.com/
[§] https://blazepod.eu/?tw_source=google&tw_adid=631785973612&tw_campaign=15546904716

3 Design & Implementation

3.1. Methodology

The methodology we adopted for the implementation of the proposed device follows the process described in figure 1a. In this process, the most crucial steps are the correct definition of the initial requirements and the two evaluation processes. The criteria of the first evaluation relied on the component's datasheets, schematic simulations using LTspice[**], as well as testing and experimentation on a breadboard shown in figure 2a. In addition, the second evaluation was conducted in each of the constructed prototype devices using the Analog Discovery 2[††]. In this work, we achieved the specifications in just two full iterations.

The initial requirements derived from the collaborative knowledge of professional coaches, who utilize cutting edge technologies in the field of Sports as well as the experimental results gathered from an on-going project, Remedes[‡‡]. Remedes is a field-tested project of the ISSEL[§§] laboratory that relies upon similar devices and therefore it will be used as reference in order to assess the performance of this work. Along with user safety, some of the most important functional requirements are presented in the Table 1c.



(a) Methodology Algorithm. (b) Abstract Structure.

Specification	Margin
Min Battery Life	12 Hours
Max Responsiveness	50 msecs
Min Wireless Range	150 meters
Max Cost	100 €
Max Volume	500 cm^3

(c) Functional Requirements.

Figure 1: Requirements & Methodology of the proposed device.

3.2. Architecture

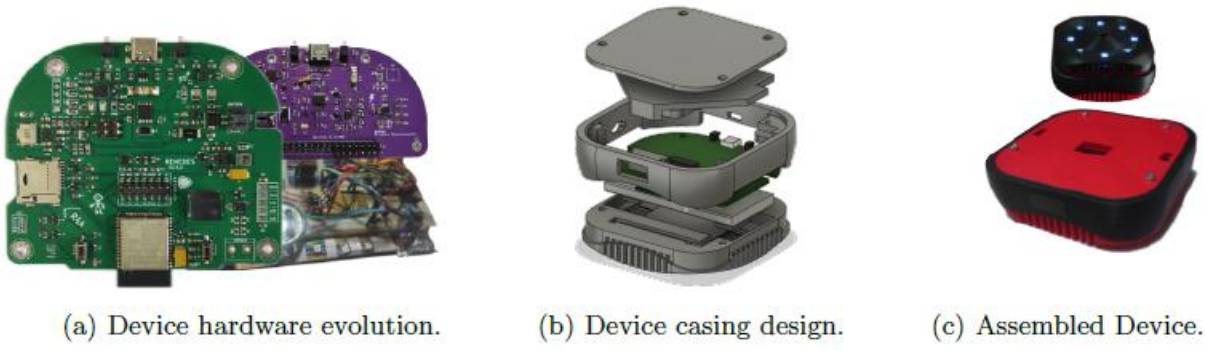
The adapted architecture structurally presents similarities with the paradigm of an Arduino ecosystem, where there are general purpose boards that can connect to one or multiple shield boards. In the same fashion, the proposed methodology assembles by the combination of components that we will refer to as Bases and Expansions. Bases are general purpose components that contain all the common functionality among devices such as power management, connectivity, storage & processing power. On the contrary, Expansions have all the specialized hardware to implement a specific Sports exercise. These components communicate through a common interface

[**] https://www.analog.com/en/design-center/design-tools-and-calculators/ltspice-simulator.html
[††] https://digilent.com/reference/test-and-measurement/analog-discovery-2/start
[‡‡] https://lab.issel.ee.auth.gr/remedes/
[§§] https://issel.ee.auth.gr/en/13-2/

bus. In contrast, however, to the Arduino ecosystem our approach stands out in the following major aspects: 1) Specialization: The electronic parts of our device have been carefully selected in order to meet the requirements of Sports domain applications. 2) Robustness: The proposed device utilizes industry approved electronic components and is programmed using the powerful Riot RTOS, where Arduino is mainly used for education & rapid development purposes. 3) Software flexibility: Every Expansion can be virtually connected to any Base that possesses the peripheral support on the hardware level without any code modification. By utilizing Riot's abstraction layers, each software driver is written once and can be compiled to any of the supported architectures. 4) Ease of Use: Bases can automatically identify the connected Expansions through the common interface bus and load the appropriate drivers, resulting in a seamless transition between them.

3.3. Hardware Design

During the hardware design phase, the selection of the appropriate microcontroller unit is of paramount importance. In the current work, the Base board was developed using the ESP32- Wroom-32D7 MCU,



(a) Device hardware evolution. (b) Device casing design. (c) Assembled Device.

Figure 2: Construction steps of the proposed device.

Wroom-32D*** MCU, developed by Espressif†††, as it satisfies most of the derived requirements from Section 3.1. Sports devices rely on sensors in order to accurately measure quantities such as distance or touch. ESP32 offers an abundance of peripherals (SPI, UART, I2C, ADC) that enables it to connect to hundreds of different sensors. Another great aspect of ESP32 is the integrated wireless connectivity via WiFi & Bluetooth, allowing it to publish the collected data to the cloud or into another smart device, where it can be processed and visualized in order to help coaches evaluate the performance of their athletes. In addition, ESP32 has a long range wireless communication protocol, ESP-NOW‡‡‡, which is useful for deploying outdoor track exercises over long distances. All in all, ESP32's high processing power, versatility and power management capabilities renders it an ideal choice.

Furthermore, the Base board includes an SD Card storage system, allowing the device to save the aggregated data while being in an environment without internet access. In order to facilitate the programming and debugging process a USB-to-UART bridge IC has been added which enables serial

*** https://www.espressif.com/sites/default/files/documentation/esp32-wroom-32d_esp32-wroom-32u_datasheet_en.pdf

††† https://www.espressif.com/

‡‡‡ https://www.espressif.com/en/solutions/low-power-solutions/esp-now

communication between the ESP32 and a PC. The Base also includes status leds and a passive buzzer so it can notify the user both visually and acoustically. Lastly, in the final design we attach a fully functional sound subsystem consisting of a dedicated class D PCM amplifier and two 8 ohm speakers, allowing the device to give vocal instructions to the athletes.

After identifying all the functional hardware components of the Base board and estimating the average power consumption of the Expansion boards, the Power Management System (PMS) was developed. The PMS is responsible for providing the appropriate voltage and sufficient current to all electronic components of both the Base and the Expansion Boards. This subsystem implements: 1) A continuous power delivery function, which automatically switches to the currently available power source, reassembling a single virtual power source that operates continuously. Currently there are two power sources in our device, a 2600mah LiPo Battery and the external power supply via the USB Type C port. 2) Battery charging capabilities: When the USB port is connected, the battery disconnects from the load and enters charging mode. 3) Battery safety mechanisms. 4) Power consumption gnostic functions. 5) Two switching voltage regulators in order to stabilize the voltage at 3.3v/4.3w & 5v/10w. 6)Power saving mode. During this state the ESP32 enters deep sleep mode, the 3.3v regulator enters pulse skipping mode and the rest of the components are disconnected from the battery via MOSFET switches. All the above can be viewed in the figures 2a, 3 & 4.

Figure 3: Reflexes & Jump Pad Expansions.

Figure 4: Balance & Screen Expansions.

Figure 5: Codin programming tool.

3.4. Implementation

The PCB's that were designed using the open source tool KICAD[§§§] were sent for manufacturing. Then, the PCB assembly process took place by hand using the reflow soldering method. In figures 2b & 2c , the 3D model as well as the actual proposed device are presented. The 3D case designs were developed in Fusion360[****]. The device was programmed using the IoT friendly Riot-OS [7]. Riot is a lightweight and energy efficient RTOS, containing multiple abstraction layers which hides the internal architecture of the host device. Therefore, the written firmware can be compiled into many different architectures. Finally, the device was successfully connected to to ISSEL's remote IoT inspection platform Codin[††††], as shown in figure 5. Codin is a useful tool for creating customized GUI's, that can control and monitor IoT devices. The connection was possible through an intermediate Raspberry Pi node running Commlib [8].

[§§§] https://www.kicad.org/
[****] https://www.autodesk.com/products/fusion-360/overview
[††††] https://codin.issel.ee.auth.gr/

3.5. Empirical Evaluation

In the direction of evaluating the practical usage of the proposed device regarding the modularity, four Expansion boards were developed, presented in figures 3 & 4. The first Expansion is similar to many of the devices presented in chapter 2 and consists of RGB leds and a TOF distance sensor, making it ideal to measure the reflexes of an athlete after a given optical stimulus. The second Expansion connects to a custom jumping pad enabling our device to measure the height of the athlete's jump. Athletes can receive more detailed instructions via the screen we installed on the third Expansion. Lastly the fourth Expansion has a IMU sensor and can be used to assess the athlete's balance capabilities by measuring the inclination of the XY plane.

4 Conclusions & Future work

In this work, we present and discuss our methodology and the implementation strategy of a modular, open-hardware/software IoT device targeting the Sports domain. Following the methodology described in chapter 3.1 we managed to create a working prototype Base with four different Expansion boards and meet the initial requirements by achieving an average battery life of 52 hours in normal Operation & 257 hours in Deep Sleep Mode, at a total cost of 74€ per device in small scale production. Due to COVID-19, compromises regarding the hardware component selection were made. In addition, direct testing in real conditions was difficult, therefore the verification process was conducted by comparing the proposed device with the devices of the on-going project Remedes, in which it showed superior functionality regarding the battery life & overall speed. The adopted architecture showed the following advantages: 1) Scalability: New athletic exercises of measurement configurations can be imported by just adding a new Expansion into the system. This is especially useful when a Sports team is expanding and therefore it cannot define its demands beforehand. 2) Reusability: Old Expansions can be combined with new Bases and via verse , creating multiple combinations. 3) Cost reduction: While each device separately may be more expensive, the total cost for a Sport team is greatly reduced. 4) Rapid Development: Each new Base and Expansion board can be developed independently and software drivers can be reused. While meeting the initial requirements, more Base boards should be designed in order to further investigate the capabilities of this architecture. An interesting extension in the current solution would be to further expand its modularity, by separating the connectivity form the Base and deploying it in a separate component. All the hardware schematics, software code and demos are available on github, in the form of open-source projects[‡‡‡‡].

5 References

[1] Lionel Sujay Vailshery "Number of IoT connected devices worldwide 2019-2021, with forecasts to 2030", statista, 2022, [Online]. Available: https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

[2] "Internet of Things (IoT) Market Size, Share COVID-19 Impact Analysis, By Component (Platform, Solution & Services), By End-use Industry (BFSI, Retail, Government, Healthcare, Manufacturing, Agriculture, Sustainable Energy, Transportation, IT & Telecom, and Others), and Regional Forecast, 2023-2030", FORTUNE BUSINESS INSIGHTS, 2023, [Online]. Available: https://www.fortunebusinessinsights.com/industry-reports/internet-of-things-iot-market-

[‡‡‡‡] https://github.com/robotics-4-all/remedes-riot

100307

[3] ”Internet of Things (IoT) inSports: Bringing IoT to Sports Analytics, Player Safety, and Fan Engagement”, Delloitte Digital, 2018, [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consumer-business/us-cb-internet-of-things-sports.pdf

[4] Ahmad Hassan Kobeissi, Riccardo Berta,Francesco Bellotti, Hussein, Chible Alessandro De Gloria ”Design and implementation of an IoT system for enhancing proprioception training”, In Proceeding of the 2017 29th International Conference on Microelectronics (ICM)

[5] Pueo B. , Jimeney-Olmedo J. M. , Lipi´nska P. , Bu´sko K. , Penichet-Tomas A. ”Concurrent validity and reliability of proprietary and open-source jump mat systems for the assessment of vertical jumps in sport sciences”, In Proceeding of the October 2018 Acta of bioengineering and biomechanics / Wroclaw University of Technology

[6] Poptsi, Eleni, Despina Moraitou, Emmanouil Tsardoulias, Andreas L. Symeonidisd, and Magda Tsolaki. ”Is the discrimination of subjective cognitive decline from cognitively healthy adulthood and mild cognitive impairment possible. A pilot study utilizing the R4Alz battery.” Journal of Alzheimer’s Disease 77, no. 2 (2020): 715-732.

[7] Emmanuel Baccelli, Oliver Hahm, Mesut G¨unes, Matthias W¨ahlisch, Thomas C. Schmidt ”RIOT OS: Towards an OS for the Internet of Things”, In Proceedings of the 2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)

[8] Panayiotou, Konstantinos, Emmanouil Tsardoulias, and Andreas L. Symeonidis. ”Commlib: An easy-to-use communication library for Cyber-Physical Systems.” SoftwareX 19 (2022): 101180

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 64 – 70

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

High-Performance Design of SRT IP Blocks

Aristotelis Tsekouras[1], Giorgos Stagakis[1], Anastasis Avgoustidis[2], Grigoris Kokkonis[1], Konstantinos Gkekas[2], Vasilis F. Pavlidis[1], Thomas Noulis[2] and Georgios Keramidas[3,4]

[1] *Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece*
[2] *Department of Physics, Aristotle University of Thessaloniki, Thessaloniki, Greece*
[3] *Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece*
[4] *Think Silicon, S.A. An Applied Materials Company, Patras, Greece*

{aristotet gstagakis}@ece.auth.gr, anavgous@physics.auth.gr, gkokkonis@ece.auth.gr, kogkekas@physics.auth.gr, vpavlid@ece.auth.gr, tnoul@physics.auth.gr, gkeramidas@csd.auth.gr

Abstract

The design and verification of three floating-point components (division, reciprocal, inverse square root), based on the Sweeney, Robertson and Tocher (SRT) algorithm and implemented in SystemVerilog, that operate with up to three pipeline stages are presented. The design flow for these blocks utilizes the Genus tool from Cadence for synthesis. To ensure proper functionality, Chipware components provided by Cadence serve as a golden reference and a means of validating the correctness of the blocks. A SystemVerilog testbench is developed, instantiating the designed SRT components as well as the respective Chipware components, enabling comprehensive functional testing across various rounding modes and precisions. While the components are specifically designed to support half, single, and bfloat16 precisions, they are also adaptable to custom precisions. The functional correctness of the components is evaluated through coverage analysis using the Integrated Metrics Center (IMC) of the Xcelium tool, achieving industry-level toggle, block, and expression coverage for all parameter combinations. Furthermore, equivalency checking is performed using the Conformal tool during different synthesis stages. Power consumption results are obtained using Joules.

1 Introduction

The motivation behind this work stems from the increasing demand for these components in scientific and data-intensive applications, where accurate and efficient computations are crucial. This work aims to enhance computational capabilities, improve numerical accuracy, and enable complex mathematical operations with low power in relevant application domains.

The floating-point components reciprocal (1/a), division (a/b) and inverse square root ($1/\sqrt{aa}$) have been developed in SystemVerilog, synthesized, and verified using Cadence tools. The block diagrams at register transfer level (RTL) for the three components are depicted in ***Figure 1***. The mantissa width, the exponent width, the rounding modes, and the number of pipeline stages are all parameterized. Following the RTL design of the components, a testbench was created to facilitate verification and

coverage analysis. Subsequently, the designs undergo formal verification, synthesis, and logic equivalency checks. Finally, the power of the components is measured through Joules, and the components can be used as soft IPs.

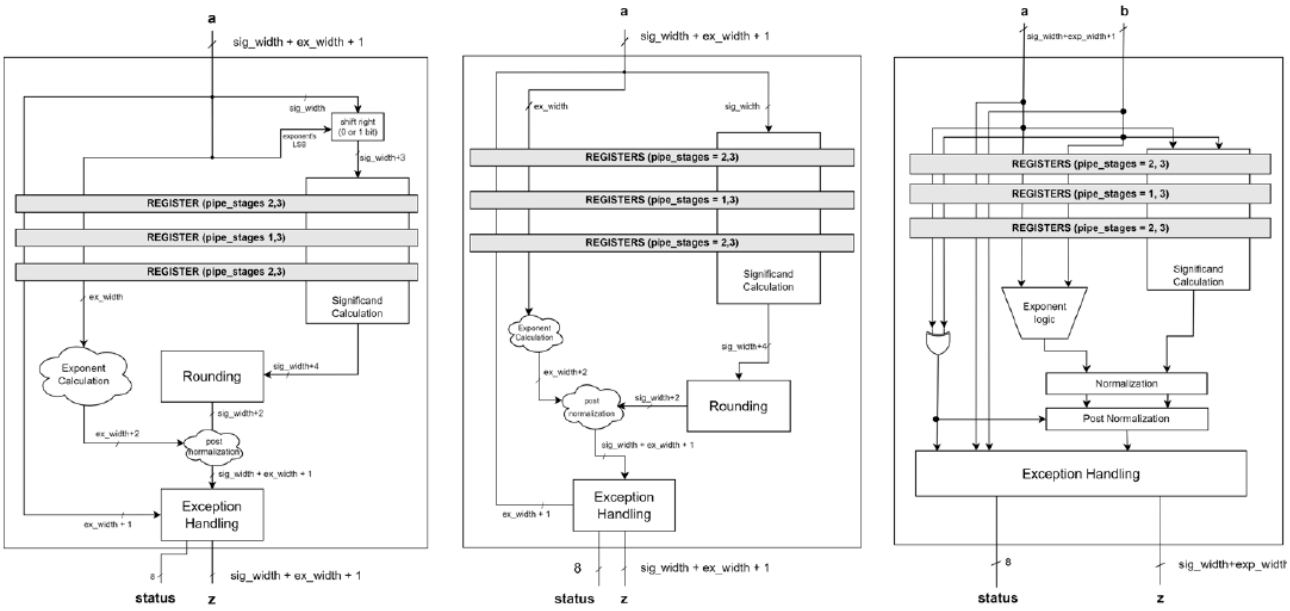


***Figure 1**: Block diagram of the inverse sqrt, reciprocal, and division components respectively*

Despite the undisputed advantages that IoT technologies bring to the Sports domain, currently their large

2 Design Description

The three components have many similar characteristics since all of these components use the SRT algorithm to determine the significant part of the floating-point number. Each component is divided into four parts, which are responsible for the exponent calculation, the significand calculation, the rounding and post-normalization, and the exception handling. The input number of a floating-point root component is a normalized number represented as

$$a = (-1)^s \cdot 2^{e-bias} \cdot (1.f) \quad (1)$$

where *s* is the sign, *e* is the biased exponent and *f* is the fraction of number *a*.

2.1. Exponent Calculation

The exponent of the result (*z*) can be determined through the following equations depending on the specific operation of the component.

Division: $e_x = (e_a - bias) - (e_b - bias) + bias = e_a - e_b + bias$ (2)

Reciprocal: $e_z = -(e_a - bias) + bias = 2 \cdot (bias - e_a)$ (3)

Inverse square root: $e_z = -\frac{e_a - bias}{2} + bias = \frac{3}{2} bias - \frac{e_a}{2} - e_a[0]$ (4)

where e_a (and e_b) are the unbiased exponents of inputs *a* (and *b*). Note that in the case of the inverse square root component, if the biased exponent of the input is odd, the significand must be shifted 1-bit left such that the division by 2 is exact.

2.2. Significand Calculation

The SRT algorithm is a digit-recurrence method producing an approximation of the result. More precisely a radix-4 (r = 4), carry-save version of the algorithm is implemented using redundant representation. This redundancy implies that each quotient digit is chosen from five possibilities: {−2, −1, 0, +1, +2}, and is encoded using One-Hot encoding. Consequently, these possibilities are encoded as: {1000, 0100, 0000, 0001, 0010} respectively.

In order to avoid the normalization step after calculating the significand of z, the inputs are shifted to the right. In this way, the inverse square-root component $\alpha \in \left[\frac{1}{4}, \frac{1}{2}\right)$ results to $z = \frac{1}{\sqrt{a}} \in (1,2]$ and the reciprocal component $\alpha \in \left[\frac{1}{2}, 1\right)$ results to $z = \frac{1}{a} \in (1,2]$. On the contrary, for the division component in order to use the same selection function with the reciprocal component $\alpha \in \left[\frac{1}{8}, \frac{1}{4}\right)$ and $b \in \left[\frac{1}{2}, 1\right)$ results to $z = \frac{a}{b} \in \left[\frac{1}{8}, \frac{1}{2}\right)$ thus the result must be shifted 2 or 3 bits to the left.

For computing the result, a residual is defined for each operation as:

- Inverse square-root $\left({}^{1}/_{\sqrt{d}}\right)$: $w[j] = 2^{-1} \cdot 4^{j} \cdot (1 - dQ[j]^2)$ (5)
- Division $\left({}^{x}/_{d}\right)$: $w[j] = 4^{j} \cdot \left(\frac{x}{8} - \left(\frac{d}{2}\right) Q[j]\right)$ (6)
- Reciprocal $\left({}^{1}/_{d}\right)$: $w[j] = 4^{j} \cdot (1 - dQ[j])$ (7)

where $Q[j] = Q[0] + \sum_{i=1}^{j} q_i \, 4^{-i}$ is the partial result up to iteration *j*.

From the above residual definitions, the recurrences are:

- Inverse square-root $\left({}^{1}/_{\sqrt{d}}\right)$: $w[j+1] = 4w[j] + q_{i+1}Q[j]d \; - 2^{-1} \, q_{j+1}^2 4^{-(j+1)d}$ (8)
- Division $\left({}^{x}/_{d}\right)$: $w[j+1] =\cdot \, 4w[j] \; - q_{j+1^d}$ (9)
- Reciprocal $\left({}^{1}/_{d}\right)$: $w[j+1] =\cdot \, 4w[j] \; - q_{j+1^d}$ (10)

To simplify the implementation of the inverse square-root recurrence steps, two variables are defined:

$$D[j] = dQ[j] \text{ and } C[j] = 2^{-1} \cdot 4 - [j + 1] \tag{11}$$

Using these variables, we can obtain the recurrences:

$$w[j+1] = 4w[j] - q_{j+1}D[j] - q_{j+1}^2 C[j] \;\; D[j+1] = D[j] + 2q_{j+1}C[j], \qquad C[j+1] = 4^{-1}C[j] \tag{12}$$

Since q_{j+1} is an encoded digit, MUXs can be used for the multiplications of q_{j+1}d, q_{j+1}D[j] *etc*. The q_{j+1}digits are selected in a way to bound the absolute error of $Q[j]$ ([1], [2]). The selection function uses an estimation of the residual *w* and an estimation of *d* (*D* for the inverse square root component).

D_L	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
m_{-1}	12	13	14	14	15	15	16	17	17	18	19	19	20	21	21	22
m_0	3	4	4	4	4	4	4	5	5	7	7	6	5	6	8	8
m_1	-5	-5	-5	-6	-6	-6	-7	-7	-7	-8	-8	-8	-9	-9	-9	-10
m_2	-13	-14	-14	-15	-16	-17	-18	-18	-19	-19	-20	-21	-23	-23	-24	-25

*D_L is scaled by 32 and m_k is scaled by 16

Table 1: *Selection function for division and inverse square root components*

D_L	8	9	10	11	12	13	14	15
m_{-1}	12	14	15	16	18	20	20	24
m_0	4	4	4	4	6	6	8	8
m_1	-4	-6	-6	-6	-8	-8	-8	-8
m_2	-13	-15	-16	-18	-20	-20	-22	-24

*D_L is scaled by 8 and m_k is scaled by 16

Table 2*: Selection function for reciprocal component*

The selection is performed using ***Table 1*** and ***Table 2*** as follows: for $\tilde{d} = D_l(i)$ ($\widetilde{D} = D_l(i)$ in case of inverse sqrt),

$$q_{j+1} = -2 \; if \; 4\tilde{w} < m_{-1} \tag{13}$$

$$q_{j+1} = k(-1 \le k \le 1) \; if \; m_k \le 4\tilde{w} \le m_{k+1} \tag{14}$$

$$q_{j+1} = 2 \; if \; 4\tilde{w} \ge m_2 \tag{15}$$

Where $\tilde{d}$: 4 bits after the MSB of (the MSB of *d* is not an input as it is always 1), $\widetilde{D}$: 5 MSBs of D and $4\tilde{w}$: 7 MSBs of *w*.

Since the digits *q* are encoded, there is a need to convert these digits on the fly from redundant signed-digit representation to conventional (Q). This conversion is performed in each step of the recurrence and the result depends on the encoded digit *q* as well as the decoded quotient Q of the previous step (***Table 3***). For a detailed explanation see [3, p. 256].

q_{j+1}	Q[j+1]	QM[j+1]
+2	{Q[j], 2'b10}	{Q[j], 2'b 01}
+1	{Q[j], 2'b 01}	{Q[j], 2'b 00}
0	{Q[j], 2'b 00}	{QM[j], 2'b 11}
-1	{QM[j], 2'b 11}	{QM[j], 2'b 10}
-2	{QM[j], 2'b 10}	{QM[j], 2'b 01}

Table 3*: On the fly conversion*

The SRT algorithm produces two bits of the quotient in each step meaning that in order to produce the full *sig*_width width significand plus the guard, round, and sticky bit the required number of steps are $\left\lceil \frac{sigwith + 3}{2} \right\rceil$. Finally, a full addition is needed to calculate the final residual *w* and determine the correct quotient depending on the sign of *w* (if w≥ 0: Q *else* QM).

2.3. Rounding

Since the result is already normalized, there is only a need for rounding and, if needed, re-normalization, in case the rounding leads to an overflow. The rounding can be implemented by assigning to the *round* parameter 1 out of 6 rounding modes as reported in ***Table 4***. For the rounding, three extra bits (guard, round, and sticky) after the mantissa are used to determine whether the result should be incremented. The guard bit determines whether the unrounded result is above or below the middle of the two nearest representable values, while the round bit acts as a tiebreaker, and the sticky bit as an indication of what is contained in the lesser significant bits that are not kept.

Value	Description
"IEEE_near"	Round to the nearest representable value, if both are equally near, output the result with an even significand
"IEEE_zero"	IEEE round towards zero
"IEEE_pinf"	IEEE round to +Infinity
"IEEE_ninf"	IEEE round to -Infinity
"near_up"	Round to the nearest representable value, if both are equally near, output the result closer to +Infinity
"away_zero"	Round away from zero

Table 4: *Supported rounding modes*

2.4. Exception Handling

Special case values, listed in ***Table 5*** for input *a* (and *b* in case of division), are handled separately by flushing the result *z*. For example, $\frac{1}{0} = +\infty, \frac{1}{\sqrt{-0}} = \infty, \frac{+\infty}{-0} = -\infty$Denormals are flushed to sign preserved
± Zero or ± MinNorm, depending on the rounding mode, where Denormals are defined to be nearer to Zero than MinNorm.

Name	Sign	Exponent	Significand	Value
±Zero	±	0	0	±0
±Inf	±	11…11	0	±Inf
Denormal	S	0	≠0	$(-1)^s \cdot 2^{1-bias} \cdot 0.sig$
Normal	S	0<exp<11…11	sig	$(-1)^s \cdot 2^{exp-bias} \cdot 1.sig$
±MinNorm	±	1	0	$(-1)^s \cdot 2^{1-bias}$

Table 5: *Special case values*

3 Verification

To verify the correctness of the designs, three comprehensive testbenches compare the behavior of the division, reciprocal, and inverse square root components with the output from the respective Chipware/Designware components. The testbenches cover all possible parameter combinations and utilized covergroups to capture inputs and outputs. Cross-coverage is also employed for precise analysis. Special consideration is given to denormals and NaNs with illegal bins designated for these cases at the outputs, as well as for the unreachable status flags.

Regarding the covergroups, directed bins are defined differently for each of the components. For

reciprocal and inverse square root, we use 64 bins for inputs and outputs, while for division we use two covergroups, where the first group is used to employ cross coverage exclusively on normal numbers with 32 bins, and the second group considers the normal numbers as one bin and uses 8 bins for the corner cases. We use twelve corner cases during the testing of all three components, which are the following: One specific positive/negative signaling NaN, one specific positive/negative quiet NaN, one specific positive/negative normal, one specific positive/negative denormal, positive/negative Infinity, and positive/negative Zero.

The coverage analysis for division involved comparing the outputs of 10 million inputs and the corner cases as previously mentioned. The inverse square root and reciprocal blocks are covered by 5 million inputs as they have only one input. The coverage results for all the three implemented circuits are reported in ***Table 6***.

Coverage Type	Division	Reciprocal	Invs Sqrt
Block Coverage	100%	100%	100%
Expression Coverage	100%	100%	100%
Toggle Coverage	88%	94%	92%
Covergroup Coverage	100%	100%	100%

Table 6: *Percentage of each coverage type for the three designed circuits*

4 Results

The synthesis results are produced using the 45 nm fast library gsclib045 provided by Cadence. The default Genus synthesis flow [4] is used in addition to an incremental optimization at the end of the synthesis stage. The results are summarized and compared with the respective Chipware in ***Table 7***. All the results have been produced after Synthesis. Area and timing results are produced by Genus, while power results are produced by Innovus.

Note that our designs have a disadvantage in terms of area when compared to their respective Chipware counterpart. However, the power consumption of our SRT blocks are comparable or lower to Chipware, enabling their use for low power applications.

Component	Area (μm^2)	TNS (ps)	Power (mW)
Division @ 6 ns	9956.17	0	2.238
Division CW @ 6 ns	6634.87	0	1.667
Reciprocal @ 6 ns	5555.24	0	1.435
Reciprocal CW @ 6 ns	5746.01	1	1.47
Inverse Sqrt @ 10 ns	23831.12	0	3.792
Inverse Sqrt CW @ 10 ns	14823.24	0	2.243

Table 7: *Power, performance, and area (PPA) results for the SRT components for zero pipeline stages compared to the Chipware components*

5 Conclusion

Three floating-point SRT components, demonstrating the accuracy, functionality, and power efficiency are designed and verified. These components contribute to the advancement of

computational capabilities and numerical accuracy in a range of applications. The presented blocks exhibit comparable performance with the commercial IP blocks offered by Cadence while offering higher versatility through a customizable range of pipeline stages to suitably adapt to the requirements of the target application.

6 References

[1] E. Antelo, T. Lang, P. Montuschi and A. Nannarelli, "Low latency digit-recurrence reciprocal and square-root reciprocal algorithm and architecture," 17th IEEE Symposium on Computer Arithmetic (ARITH'05), Cape Cod, MA, USA, 2005, pp. 147-154, doi: 10.1109/ARITH.2005.29.

[2] T. Lang and E. Antelo, "Radix-4 reciprocal square-root and its combination with division and square root," in IEEE Transactions on Computers, vol. 52, no. 9, pp. 1100-1114, Sept. 2003, doi: 10.1109/TC.2003.1228508.

[3] M. D. Ercegovac and T. Lang, *Digital Arithmetic*, 1st ed., Morgan Kaufmann, 2003, pp. 247-330.

[4] *All Products | Cadence - Cadence Design Systems*. Available at: www.cadence.com/en_US/home/tools/tools-a-z.html (Accessed: 01-10-2023)

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 71 – 77

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Fractional-N Phase Locked Loop for Wi-Fi 6/6E With 135 mW Power Consumption and Spur Reduction Techniques

Savvas Sgourenas[1], Christos Andriakopoulos[1], Stefanos Pokamisas[1], Charis Basetas[1], Chrysa Vassou[1], Vasilis Tsamis[1], Kostas Retsinas[1], Vasilis Kolios[1], Nektarios Sgourenas[1], and Giorgos Kasapoglou[1]

[1] *MEICSi*
ssgourenas@meicsi.com, candriakopoulos@meicsi.com, spokamisas@meicsi.com, cbasetas@meicsi.com, cvassou@meicsi.com, vtsamis@meicsi.com, kretsinas@meicsi.com, vkolios@meicsi.com, nsgourenas@meicsi.com, gkasapoglou@meicsi.com

Abstract

This paper presents a fractional-N phase locked loop (PLL) for Wi-Fi 6/6E applications in 28 nm FD-SOI CMOS technology. The PLL consists of an LC voltage-controlled oscillator (VCO), a dual edge phase-frequency detector (PFD), a programmable charge pump with linearization scheme, a prescaler, a fully integrated low pass filter and a 3rd order Sigma-Delta modulator to control a programmable divider. Notable features include fast automatic frequency locking, and optimal loop bandwidth (LBW) and VCO amplitude calibration. The PLL draws current from a 1.8 V supply, having a power consumption of 135 mW. The reference clock is 60 MHz, RMS jitter is 161.9 fs (1 kHz – 100 MHz) and the reference spur level is -55 dBc.

1 Introduction

Phase-Locked Loops (PLLs) (Razavi, 2020) play a crucial role in modern wireless communications. The PLL is one of the key building blocks of RF front-end transceivers, generating the local oscillator (LO) signal. The LO signal specifications, i.e. phase noise and frequency, are determined by the wireless standard. The PLL presented in this paper targets Wi-Fi 6/6E standard specifications. Key performance metrics of a PLL include strong suppression of reference and harmonic spurs, low RMS jitter and minimal power consumption. This work presents a fractional PLL with output frequency range 6.5 – 9.7 GHz.

2 System Architecture

The proposed fractional-N PLL generates the LO clocks for the wireless transceiver. Figure 1 shows the block diagram of the Fractional-N frequency synthesizer architecture. Charge-pump based topology is selected to achieve high-end phase noise performance. Multiple oscillator cores ensure VCO noise optimization over frequency in the expense of additional area. To suppress reference spur and Sigma-Delta modulator noise, a 4th order, programmable low pass filter has been implemented. The optimum loop bandwidth for the specific design is around 500 kHz. Digital logic controls the PLL parameters so that it operates as close as possible to the optimum bandwidth and achieves fast lock

time. The lock time including calibration is less than 50 μs. Fast lock time is required to enable fast channel hopping in certain wireless protocols such as Bluetooth Low Energy. Special techniques to compensate for the dynamic phase error introduced by the Sigma-Delta modulator, which elevates fractional spur power levels, have been deployed.

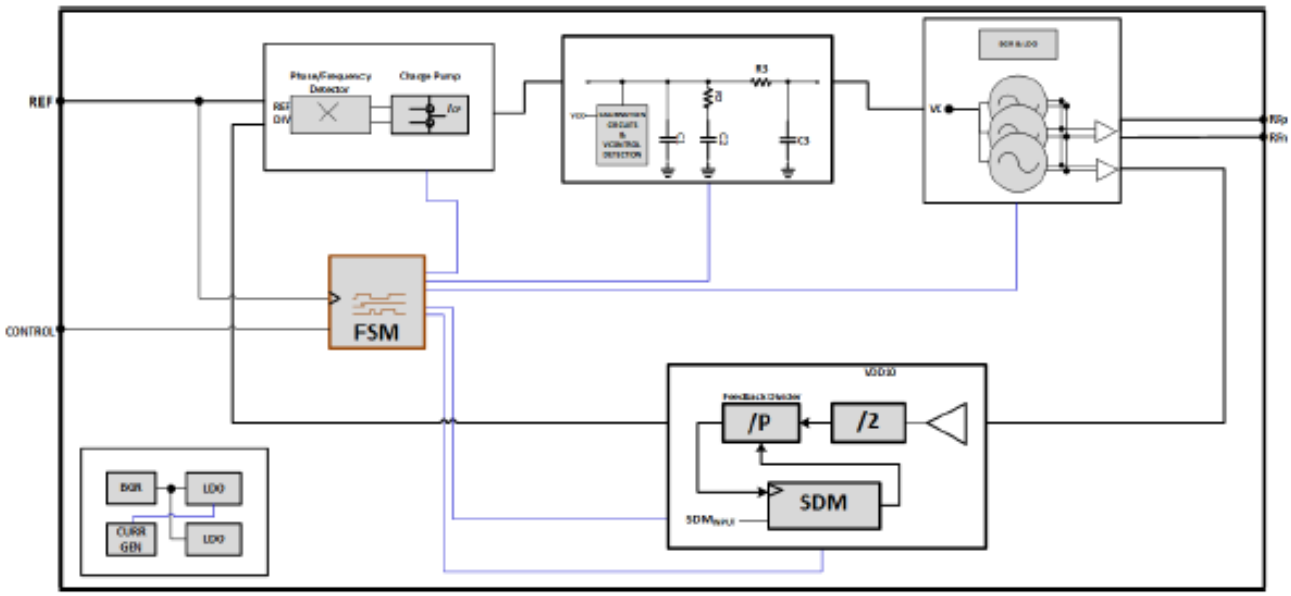


Figure 1: PLL block diagram

2.1. PFD & Charge Pump

The phase frequency detector (PFD) and the charge pump (CP) are used to convert the phase difference between the reference and the feedback signals to a proportional current signal. An important PFD operation is to "recognize" which pulse is leading or lagging, causing the charge pump to either sink or source current. The PFD can detect the phase difference of both positive and negative edges of the reference and feedback signals, effectively doubling the reference signal frequency. Using a clock doubler instead would increase the phase noise by 6 dB, since the phase noise of the multiplied signal increases with the multiplication factor.

When designing the PFD and CP, the main targets are to minimize spurious frequencies, phase noise and power consumption. Common factors that can increase the spurious tones are the dead zone of the PFD, and CP current mismatches and off-state leakage. Isolating the CP current mirrors from the local bias using sample & hold scheme results in reduced output noise. Special attention has been paid to fractional spur suppression. Dedicated circuits are used to force the CP to operate in a linear region and to compensate for the instantaneous phase error introduced by the Sigma-Delta modulator. As shown in Figure 2 (b), to reduce current mismatch, the bias current I_{bias} is generated by a bandgap reference and the operational amplifier, used as a buffer, forces CP_{out} and the other branch to have the same voltage. The top left operational amplifier regulates the PMOS current mirror gate voltage, while CP_{out} is fluctuating.

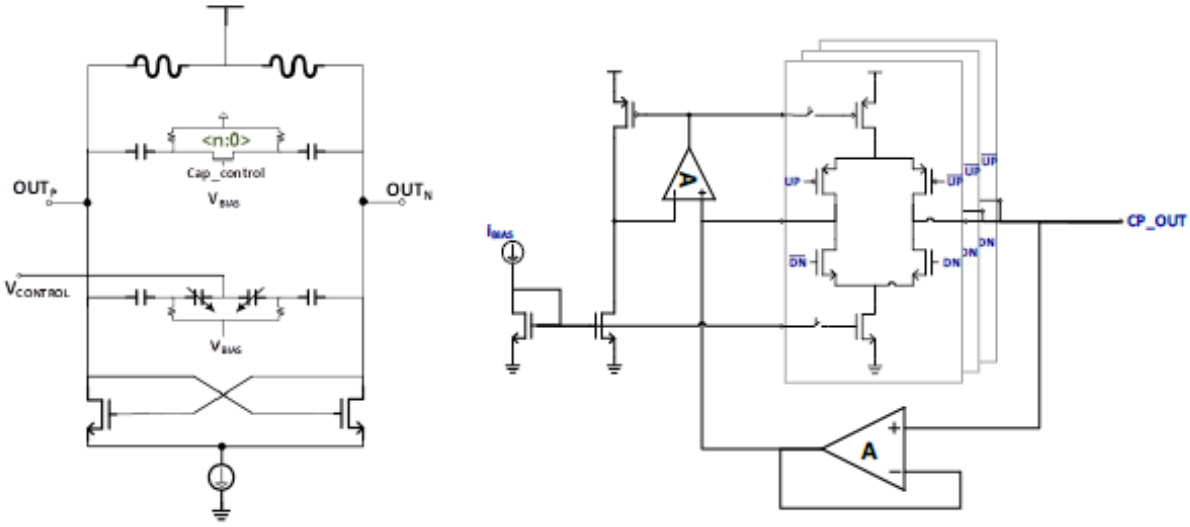


Figure 2: (a) VCO architecture, **(b)** Charge Pump

2.2. VCO

Figure 2 (a) shows the VCO architecture, which has a frequency range from 6.5 GHz to 9.7 GHz. To cover this frequency range, the VCO consists of three cores. Each core utilizes a custom 8-shaped, high Q inductor, which is the best shape to minimize the effect of PA pulling. The architecture employs digitally controlled binary and thermometer-coded capacitor banks controlled by digital logic to allow tuning of the VCO to the desired frequencies over process, voltage, and temperature (PVT) variations. The capacitor banks are designed to achieve more than 60% frequency overlap between the bands over PVT variations and mismatch. Also, a network of varactors is employed to realize fine frequency tuning with a gain that ranges between 35 MHz/V and 120 MHz/V over PVT variations and frequency bands. High level layout EM extracted views have been used in the simulations, resulting in excellent correlation of the simulation results with the lab measurements.

2.3. Sigma-Delta Modulator

The digital Sigma-Delta modulator (SDM) is used to modify the division ratio of the PLL feedback divider to achieve fractional division. A 3rd order modulator has been implemented so that there is minimal impact on the PLL phase noise due to the SDM quantization noise. The SDM design uses a single stage multiple feedback topology to avoid the frequency spurs that are present in the output spectrum of typical MASH 1-1-1 SDM. Furthermore, the output range of the SDM used in the presented PLL can be decreased to reduce PLL near-in frequency spurs due to charge pump non-linearity.

2.4. Digital calibration logic

Digital calibration logic has been implemented to boost PLL performance and automate PLL configuration to allow for faster lock time. The digital calibration finite state machine (FSM) performs four kinds of calibration: VCO frequency tuning, VCO amplitude, PLL loop bandwidth, and VCO core selection.

VCO frequency tuning calibration is used to find the VCO capacitor bank configuration that results in an open loop VCO output frequency as close as possible to the target one for a given VCO varactor control voltage. To achieve the highest possible margin due to temperature variations, the VCO varactor control voltage during VCO frequency tuning calibration should be equal to the middle value of the varactor control voltage range.

VCO amplitude calibration is used to target a specific VCO amplitude that minimizes VCO phase noise. Due to PVT variations the VCO current that results in a specific VCO output amplitude is not constant. The calibration logic utilizes the VCO amplitude detector to set the VCO output amplitude to the target value. Since VCO output amplitude is affected by the VCO frequency, VCO amplitude calibration is performed in tandem with VCO frequency tuning calibration.

PLL loop bandwidth calibration is used to set the PLL loop bandwidth to a certain target value that minimizes phase noise when the PLL is locked. It is possible to control the PLL loop bandwidth by modifying the closed loop gain, which depends on the VCO frequency gain (MHz/V), the feedback signal gain, the PFD/CP gain, and the low-pass filter gain. For a given PLL output frequency only the low-pass filter and the PFD/CP gains can be modified since all other gains cannot be directly controlled. However, to accurately set the PLL loop bandwidth all the gain values must be known. The feedback signal gain is dictated by the feedback path division ratio and therefore it is known. The PFD/CP and low-pass filter gains are controlled by PLL loop bandwidth calibration, so their values are also known. VCO frequency gain depends on PVT variations and therefore it must be measured by the digital calibration FSM. Once the VCO frequency gain is known, all the different gain values are combined to set the CP current and the PLL low-pass filter resistance to the values that result in the target PLL loop bandwidth.

Due to PVT variations some PLL output frequencies, near the boundaries of the VCO core frequency ranges, can be generated by different VCO cores. Therefore, we cannot always assign the same VCO core for a given PLL output frequency. VCO core selection calibration is used to automatically select the VCO core for a specific PLL output frequency.

3 Modeling and Measurements

PLL low pass filter design, and optimum CP current and VCO gain selections were based on a PLL system model developed in Matlab. The system model utilizes all the PLL sub-block noise transfer functions and noise specifications to calculate the PLL output phase noise contribution of each PLL sub-block as shown in Figure 3.

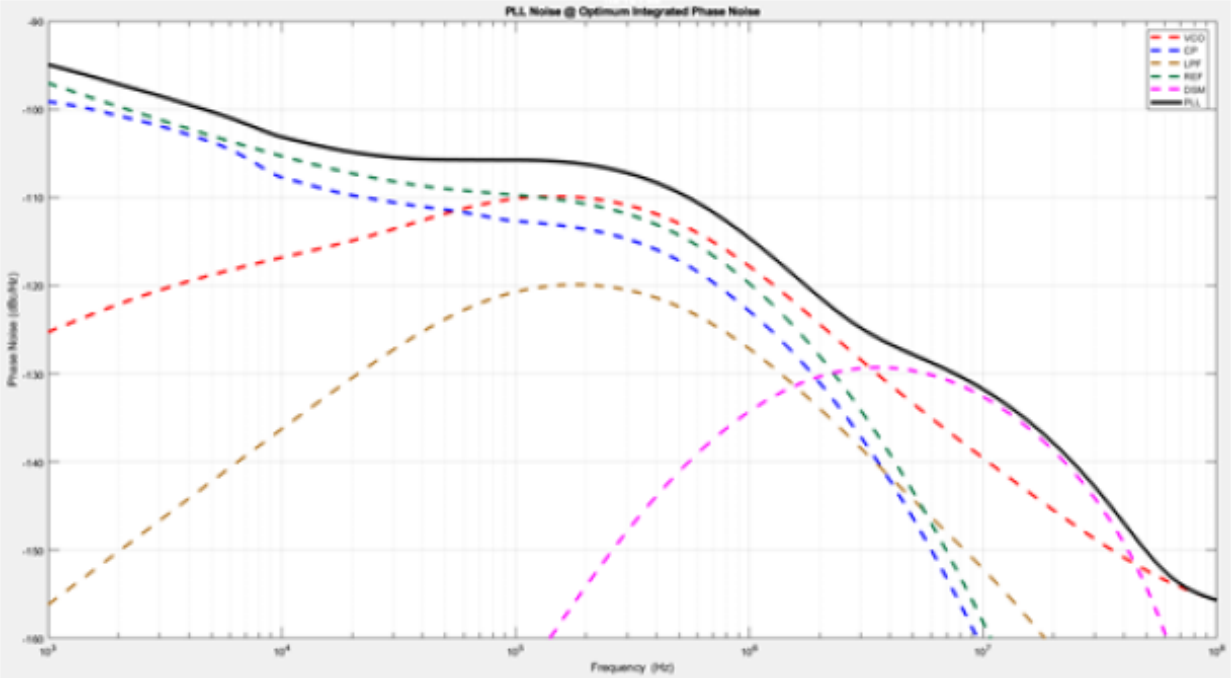


Figure 3: PLL phase noise contributions calculated using the system model

The PLL measurements were conducted through the transmitter chain output, meaning that the Tx LO signal is measured rather than the actual PLL output. The phase noise measurements at 5.18 GHz and at 7.12 GHz are depicted in Figure 44 (a) and (c) respectively. Please note that there are no visible in-band spurs (Figure 4 (b)). The reference spurs are 55 dB lower than the carrier, which is extremely low given the low reference signal frequency. Finally, in Figure 4 (d) the PLL layout is shown.

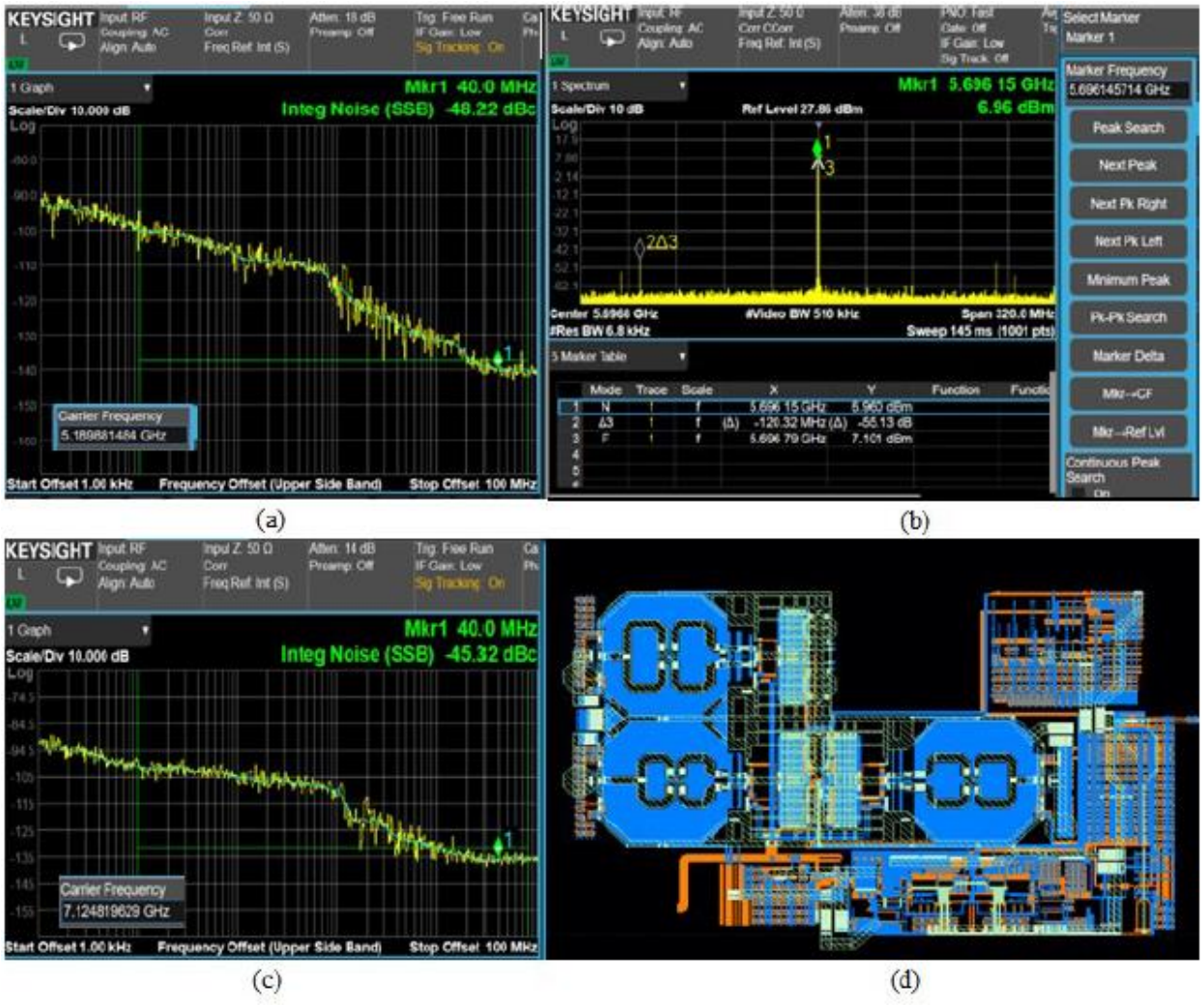


Figure 4: (a) Phase noise measured at @5.18 GHz, **(b)** Reference spur level @5.06 GHz **(c)** Phase noise measured @7.12 GHz, **(d)** PLL layout

4 Conclusions

In this paper, a PLL with output frequency range 6.5 – 9.7 GHz and many different spur reduction techniques has been presented. Furthermore, digital logic has been implemented to enable loop bandwidth calibration, fast automatic frequency locking, automatic VCO core selection and VCO amplitude calibration.

The comparison of our PLL with other state-of-the-art PLLs for Wi-Fi applications is shown in Table 1. To facilitate comparison, phase noise performance of all PLLs has been normalized to 1 GHz using the formula $PN_{norm} = PN - 20log\left(^{F_{out}}/_{1\ GHz}\right)$. Since the PLLs we compare to use different reference signal frequencies, we cannot directly compare phase noise performance. A well-known figure of merit (FoM) that takes into account RMS jitter, power consumption and the reference signal frequency (Bae, 2022) is:

$$FOM_{JR} = 10\ log\left[\left(\frac{\sigma_{rms}}{1s}\right)^2 \quad \left(\frac{P_{pLL}}{1\ mW}\right) \quad \left(\frac{f_{ref}}{1\ MHz}\right)^{1.5}\right]$$

where σ_{rms} is the PLL RMS jitter, P_{pLL} is the PLL power consumption and f_{ref} is the PLL reference signal frequency. As it can be seen in Table 1 the FoM of the presented PLL achieves the same or better performance than other works.

	(Didem Turker, 2018)	(Mayank Raj, 2017)	(Yue-Fang Kuo, 2022)	**This work @7.12 GHz**
Technology	16 nm FinFET	16 nm FinFET	TSMC 180 nm	**28 nm FD-SOI**
Number of LC VCO(s)	2	2	1	**3**
Reference freq. [MHz]	500	450	198.6 – 216.3	**60**
Frequency range [GHz]	7.4 – 14.0	9.0 – 18.0	6.42 – 6.92	**6.5 – 9.7**
Phase noise @100 kHz	-135.9	-129.2	N.A.	**-122**
Phase noise @1 MHz	-139.1	-132.4	-132	**-137**
RMS jitter [fs]	53.6	164	N.A.	**161.9**
Reference spur [dBc]	-75.5	N.A.	N.A.	**-55**
Power [mW]	45	29.2	3.1	**135**
Area [mm^2]	0.35	0.39	N.A.	**1.66**
FoM_{JR}	-148.4	-141.3	N.A.	**-147.9**

Table 1: Performance summary and comparison

The proposed PLL exhibits phase noise performance that can support the highest data rate (MCS-11) for Wi-Fi 6/6E applications, while using a low frequency reference signal. Furthermore, the low reference spur power level minimizes interference from neighboring channels. Finally, due to the digital calibration logic employed in this PLL the lock time is less than 50 μs, allowing its usage in applications that require fast frequency hopping, such as Bluetooth Low Energy.

5 References

[1] Bae, W. (2022). Benchmark Figure of Merit Extensions for Low Jitter Phase Locked Loops Inspired by New PLL Architectures. *IEEE Access, vol. 10*, 80680-80694.

[2] Didem Turker, A. B. (2018). A 7.4-to-14GHz PLL with 54fsrms Jitter in 16nm FinFET. *ISSCC* , 3.

[3] Mayank Raj, A. B. (2017). A 164fsrms 9-to-18GHz Sampling Phase Detector based PLL with In-Band Noise. *Symposium on VLSI Circuits*, C182-C183.

[4] Razavi, B. (2020). *Design of CMOS Phase-Locked Loops: From Circuit Level to Architecture Level.* Cambridge University Press.

[5] Yue-Fang Kuo, S. H. (2022). Low-Power Optimization Design of CMOS Phase-Locked Loop for WiFi-6E Applications. *IEEE International Conference on Consumer Electronics - Taiwan*, 9-10.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 78 – 84

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Enhanced Safety Architecture with Fault Preventive Mechanisms for Automotive Li-ion Battery Management Systems

Apostolos Delizonas[1], Christos Mademlis[2], Evangelos Tsioumas[2], Di-mitrios Papagiannis[2], Nikolaos Jabbour[2], Christos Sansaridis[1], Tilemachos Matiakis[1]

[1] *KENOTOM P.C., Kalamaria-Thessaloniki, Greece*
a.delizonas@kenotom.com
[2] *School of Electrical and Computer Engineering Aristotle University of Thessaloniki, Thessaloniki, Greece*
mademlis@auth.gr

Abstract

Compliance of electric vehicle (EV) Battery Management Systems (BMS) with functional safety standard ISO 26262 is considered mandatory due to several risks associated with Lithium-Ion (Li-ion) batteries (BT). However, the high-criticality safety goals formed in accordance with the standard can significantly increase the implementation and verification complexity of the EV-BMS. Therefore, the aim of this paper is to propose an improved safety architecture which can provide effective mechanisms to prevent or mitigate both systematic and random faults and avoid single instances of failure. This architecture is applicable to a wide range of EV BT topologies and therefore, it can simplify and accelerate the development lifecycle of modern EVs.[16]

1 Introduction

Despite their multiple benefits, Li-ion BTs are vulnerable to venting, fire, and explosion in case of electrical, thermal, and mechanical abuses [1]. Thus, the main objectives of a BMS are to ensure safety, protection of the BTs' lifespan and satisfactory performance of the BTs. Therefore, the EV-BMSs must comply with the automotive safety standards since potential failures may have a great impact on the vehicle's safety.

ISO 26262 [2] safety standard imposes a set of processes and provides guidelines which can ensure the functional safety of an automotive electronic control unit (ECU). According to the standard's orders, hazardous events, that may be caused by system malfunctions, should be analyzed at vehicle level. The result of this process is the risk assessment of these events and the derivation of the

[16] The project KMP6-0126719 was implemented under the framework of the Action «Investment Plans of Innovation» of the Operational Pro-gram «Central Macedonia 2014 2020» that is co-funded by the European Regional Development Fund and Greece

respective safety goals.

In the ISO 26262, risk is assessed with automotive safety integrity level (ASIL) indicators, which combine severity, exposure, and controllability of a hazardous event. There are four ASIL indicators (A, B, C, D) rated from least to strictest one, respectively. Safety goals with higher ASILs (C and D) require stricter safety mechanisms and verification methods and consequently increase the complexity of the system.

Although safety goals for an automotive BMS have been determined in [3], the combination of criticality and complexity for the functionality of the state-of-charge estimation, significantly increases the implementation and validation effort. In [4], derived safety requirements are independent from the state-of-charge estimation of the BTs; however, the high ASILs (D, C) and due to the fact that the potential random faults have been ignored, several complexity and safety issues may arise. Mechanisms for prevention of random faults have been examined in [5]. However, the absence of redundancy may lead to adverse single instances of failure.

Although safety goals for an automotive BMS have been determined in [3], the combination of criticality and complexity for the functionality of the state-of-charge estimation, significantly increases the implementation and validation effort. In [4], derived safety requirements are independent from the state-of-charge estimation of the BTs; however, the high ASILs (D, C) and due to the fact that the potential random faults have been ignored, several complexity and safety issues may arise. Mechanisms for prevention of random faults have been examined in [5]. However, the absence of redundancy may lead to adverse single instances of failure.

The aim of this paper is to introduce a low-complexity automotive safety architecture compliant with the ISO 26262. Specifically, the high-ASIL safety goals are decomposed to lower-ASIL function-alities which can be implemented by redundant and adequately independent hardware and software elements. Thus, any fault of these independent elements is prevented from causing a single point failure and also, the overall development complexity is decreased. Moreover, a combination of Fault Tree Analysis (FTA) and Failure Mode and Effects Analysis (FMEA) is performed to detect the possible random faults that can violate the determined safety goals and prevent or mitigate these faults. Various EV-BT systems are considered to validate that the proposed architecture can be adopted by almost all automotive BMSs.

2 Derivation of Safety Goals for the Automotive BMS

Modern EV BTs contain multiple Li-ion cells which are connected in series and parallel and ar-ranged in modules. Considering that the dominant operating voltages are 400Vdc and 800Vdc, EV BTs should contain several modules and thus, a high number of BT cells. Therefore, the BMS adopts a modular architecture with two subsystems, as presented in Figure 1. The Supervisor Units are respon-sible for the voltage, temperature, and current monitoring of the cells of each module. The Master Unit undertakes the task of

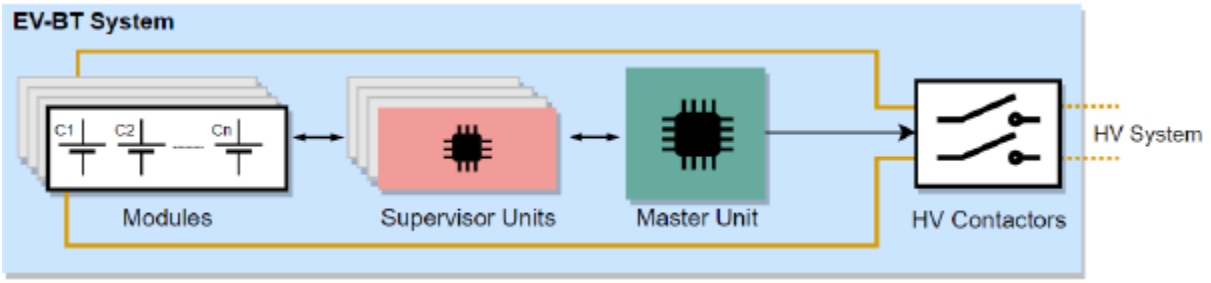


Figure 1 Overview of the modular BMS for Li-ion BTs

undertakes the task of the BT state estimation, the communication with other vehicle ECUs, as well as the connection/disconnection of the BT system from the rest HV system.

Severity	Exposure	Controllability		
		C1 (Lower)	C2	C3(Higher)
S1 (Lower)	E1 (Lower)	QM	QM	QM
	E2	QM	QM	QM
	E3	QM	QM	A
	E4 (Higher)	QM	A	B
S2	E1	QM	QM	QM
	E2	QM	QM	A
	E3	QM	A	B
	E4	A	B	C
S3 (Higher)	E1	QM	QM	A
	E2	QM	A	B
	E3	A	B	C
	E4	A	C	D

Table 1 ASIL ratings

Li-ion battery cells are susceptible to thermal runaways when abused, either electrically or ther-mally. Since Li-ion cells have significantly high energy density, the rapid self-heating may lead to ex-tended fire and violent explosions [6]. Regarding electrical abuses, Li-ion cell overcharge is the main cause for thermal runaways. In this situation, it operates above its voltage limits which leads to chemical reactions that can trigger a thermal runaway. On the contrary, deep overdischarges and excessive cur-rents may form dendrites leading to internal short circuits within the cell and eventually, thermal runa-way. Moreover, operation above 55°C - 60°C may bring the battery cell to a situation where excessive heat cannot be properly dissipated and a thermal runaway event may be initiated. On the other hand, operation of Li-ion cells at low temperatures may also form dendrites that in the worst case may pene-trate the cell's separator and lead to internal short circuits. Low temperature operation is mostly dan-gerous when charging a Li-ion cell below 0°C since charge-transfer resistance of discharged cells is much higher to that of charged ones [7].

Table 1 shows the various combinations of Severity, Exposure and Controllability classes that form the four ASIL ratings. The quality management (QM) rating represents hazards of low risk that does not need

to comply with ISO 26262 processes.

A thermal runaway event in a single Li-ion cell inside a module may easily propagate to neighbor cells due to the firm assembly of an EV-BT system [6]. Under these conditions, multiple sources of potential fire and explosion are present within the BT system. In addition, a thermal runaway occurring during the end of a charging process may have critical consequences since the Li-ion cells have all their energy available [6]. Thus, these events must be assigned to the strictest severity class S3, according to Table B.1 of ISO26262-3. Moreover, internal reactions of the cells cannot be controlled or prevented by the driver and thus, the strictest C3 class is considered.

The probability of exposure for vehicle charging depends on modern EV ranges. Considering a typical range of 340km [8] and average monthly travel of 1000km, an EV is charged approximately three times per month. Therefore, E3 exposure class is assigned as per Table B.3 of ISO 26262. On the other hand, vehicle charges at low temperatures (below 0°C) are less probable considering yearly aver-age temperatures and hence, E2 exposure class is assigned. Moreover, as Li-ion cells have a slow self-discharging rate of 5% per month, periods of no charging should last more than two months to cause dangerous deep overdischarges, as defined in [1]. Thus, E2 class is considered according to Table B.3.

Tables 2 and 4 present the Hazard Analysis and Risk Assessment for various cause events of Li-ion BT fire or explosion, as well as the respective safety goals of a BMS.

Cause Event	(S)	(E)	(C)	ASIL
Cell overcharging at vehicle charging	S3	E3	C3	ASIL C
Cell overheating at vehicle charging	S3	E3	C3	ASIL C
Cell internal short circuit due to vehicle charging after long period of no charging	S3	E2	C3	ASIL B
Cell internal short circuit due to vehicle charging at temperatures below 0°C	S3	E2	C3	ASIL B
BT conducting excessive currents due to faulty DC charger or extended acceleration at low traction surfaces	S3	E1	C3	ASIL A

Table 2 HARA for the BMS

Safety Goal	Description	ASIL
1	Prevention of cell overcharging	ASIL C
2	Prevention of cell overheating	ASIL C
3	Prevention of cell overdischarging or charging after an overdischarge	ASIL B
4	Prevention of cell charging at temperatures below 0°C	ASIL B
5	Prevention of BT overcurrents	ASIL A

Table 3 Safety Goals of the BMS

3 Proposed Safety Architecture

To achieve the safety goals, the BMS must ensure the transition to a safe state within one second, whenever a systematic fault appears. Regarding the safety goals 1 and 3, the BMS should detect operation of cells outside the overvoltage (OV) and undervoltage (UV) limits. Since these limits occur for different Li-ion chemistries, they should be configurable for our BMS to cover various EV-BT systems. For safety goals 2 and 4, operation of cells above overtemperature (OT) limit and below undertemperature (UT) limit of 0°C must be detected. Configurability should also be ensured for overcurrent (OC) limits since they differ for charging and discharging process and depend on Li-ion chemistry.

Since the analyzed hazards are related to charging and discharging of the BT, the BMS should achieve a safe state by disconnecting the BT from the rest system and interrupting any charge transfer.

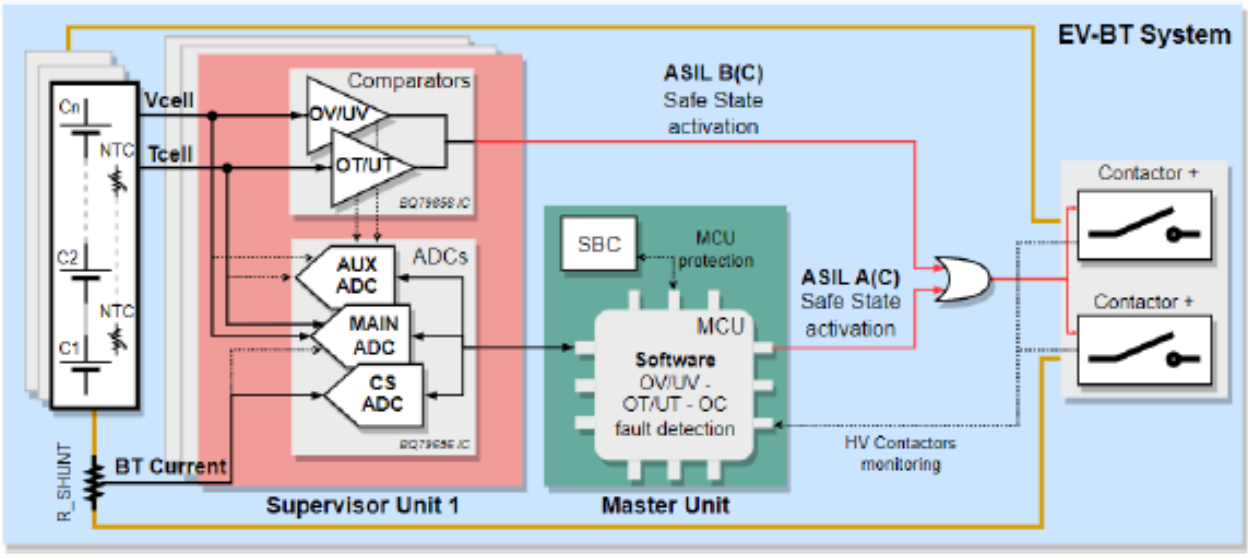


Figure 2 Layout of the proposed safety architecture

For the proposed architecture, the BQ79656-Q1 is utilized [9] as the main cell monitoring integrated circuit (IC) of the Supervisor Units and also, a microcontroller (MCU) protected by a system basis chip (SBC), for the Master Unit, is used. The proposed architecture is illustrated in Figure 2. The IC uses the main ADC for the measurements of cell voltages and temperatures and the current sense (CS) ADC for current measurement. A third auxiliary (AUX) ADC is used for the additional safety mechanisms. The measurements are transmitted via UART communication protocol. Moreover, the IC provides independent hardware comparators with configurable limits, directly connected to cell voltages and thermistor outputs with the ability to trigger a hardware reaction.

Therefore, safety goals 1-4 are decomposed to two redundant functionalities with lower ASILs. The first one [ASIL B(C)] takes over the hardware comparison of cell voltages and temperatures and also, immediate disconnection of the BT from the rest of the system, in case of faults. The second one [ASIL A(C)], is responsible for voltage and temperature measurements from the ADCs and the detection of potential faults as well as the disconnection of the BT from the software of the MCU. Safety goal 5 is

allocated only to the second functionality.

4 Safety mechanisms for prevention of random hardware faults

The proposed architecture can prevent any hazardous event due to systematic faults, as analyzed above. Since random hardware failures may also lead to a violation of a safety goal, an FTA is performed as presented in Figure 3 which is referred to safety goals 1-4.

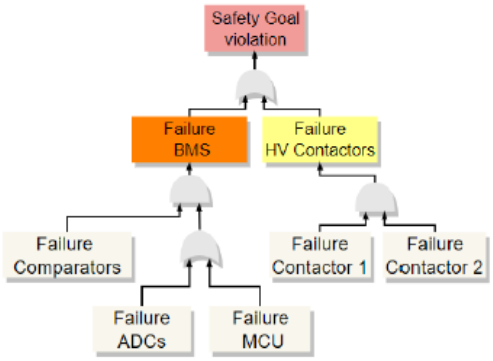


Figure 3 FTA for safety goals 1-4

From the FTA, it is revealed that none of the analyzed failures can lead directly to a violation of safety goals 1-4. Therefore, it is verified that the decomposition of these goals to redundant functional-ities has led to avoidance of single instances of failure. However, if combined, these failures may raise threats for overall system safety. Thus, the proposed architecture implements proper safety mechanisms to detect dual-point faults that may lead to these failures. The most relevant mechanisms based on FMEA are presented in Table 4. Moreover, as implied by ISO 26262, safety mechanisms for dual-point faults can be performed during start-up of the system and set to safe state instead of running continu-ously during its operation. Therefore, the implementation complexity of these mechanisms is signifi-cantly decreased.

Failure Mode	Random Fault	Safety Mechanism
Failure Comparators	Fault at value of OV/UV – OT/UT limit	Limits are measured from the AUX ADC and diagnosed by MCU
	Random Comparator Fault	MCU injects fault at comparators to verify their proper functionality
Failure ADC	Fault at Main ADC	Voltage and temperature are also measured by AUX ADC and deviations are diagnosed by MCU
	Fault at Current sense ADC	Current is measured also by Main ADC and deviations are diagnosed by MCU
	Communication fault	Both MCU and IC detect faults with communication timeouts and CRC calculations at UART messages
Failure MCU	Loss of power and other random faults	External watchdog implemented with SBC monitors the proper functionality of the MCU
Failure HV contactors	Contactors + and - stuck at closed	MCU monitors the actual state of Contactors and informs VCU in case both contactors are stuck

Table 4 Safety mechanisms for random hardware faults

5 Conclusions

The proposed BMS safety architecture provides improved safety and low complexity by covering both systematic and random faults, while critical single instances of failure can be avoided. Moreover, it can be adopted for a wide range of EV BT systems due to its modularity. Therefore, it effectively contributes to the overall safety of the EV and the acceleration of its development lifecycle.

6 References

[1] C. Yuqing, et al, "A review of lithium-ion battery safety concerns: The issues, strategies, and testing standards", *Journal of Energy Chemistry*, vol. 59, pp 83-99, Aug. 2021

[2] ISO 26262:2018 – Road Vehicles – Functional Safety – All Parts, ISO: Geneva, Switzerland, 2018

[3] W. Taylor, G. Krithivasan, J.J. Nelson, "System Safety and ISO 26262 Compliance for Automotive Lithium-Ion Batteries" in Proc. *IEEE Symposium on Product Compliance Engineering*, 2012

[4] B. Li, et al, "Research on Functional Safety of Battery Management System (BMS) for Electric Vehicles", in Proc. *Int. Conf. on Intelligent Computing, Automation and Applications (ICAA)*, 2021

[5] D. Marcos, et al, "A Safety Concept for an Automotive Lithium-based Battery Management System", in Proc. *Electric Vehi-cles Int. Conf. & Show (EV2019)*, Oct 2019

[6] F. Xuning, et al, "Thermal runaway mechanisms of lithium-ion battery for electric vehicles: A review", *Energy Storage Materials*, vol. 10, pp 247-267, Jan. 2018

[7] Shuai Ma, et al, "Temperature effect and thermal impact in lithium-ion batteries: A review", *Progress in Natural Science: Materials International*, vol. 28, pp. 653-666, Dec.2018

[8] "Range of full electric vehicles" ev-database.org, Available [Online]: https://ev-database.org/cheatsheet/range-electric-car, 2022

[9] Texas Instruments, "Automotive 16-S precision battery monitors, balancer, current sensor with ASIL-D compliance", BQ79656-Q1 datasheet, May 2021 [Rev. Jun. 2022

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 85 – 88

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Energy Efficient ML Accelerator using a Decoupled Vector Engine and a Systolic Array attached to a Low-Cost RISC-V Scalar Core

Giorgos Dimitrakopoulos, Vasileios Titopoulos, Christodoulos Peltekis, Dionysios Filippas and Kosmas Alexandridis

Integrated Circuits Lab, Electrical & Computer Engineering, Democritus University of Thrace, Xanthi, Greece

Abstract

The widespread proliferation and adoption of ML models has triggered the need to accelerate them directly in hardware. In this way, we satisfy the need for high performance needed by many ML applications with real time responses, and at the same time allow for energy efficient implementations as necessitated by edge or mobile applications. In this short report, we highlight the proposed architecture for ML acceleration that combines (a) a vector processor that supports structured sparsity, (b) a systolic array architecture with a fine-tuned pipeline organization for integer or reduced precision floating point arithmetic and (c) a low-cost RISC-V scalar core that allows dual issuing of compressed 16-bit instructions with minimal hardware overhead. Consequently, the proposed design can substantially improve instruction throughput and reduce execution times. The programmable nature of the proposed architecture allows handling efficiently both current as well as future extensions of ML applications, while its overall organization allows it to approach the energy efficiency of application-specific designs.

1 Introduction

Deep learning has had a significant impact on many rapidly emerging applications, such as computer vision [1], [2], natural language processing [3], and robotics [4]. From the outset, the widespread proliferation of various deep learning models necessitated their direct hardware acceleration, with the ultimate goal being to improve both performance and energy efficiency.

Matrix multiplications are at the heart of deep learning algorithms and their computation in hardware maps naturally to vector processors or even more parallel hardware structures such as Systolic Arrays (SA) [5]. Additional operations such as various forms of activation functions are also needed to complete computation for both inference and training.

ML applications involve dense or sparse data that follow various arithmetic formats. Sparse data may be the result of pruning during training for reducing the memory footprint of weights or even the result of activation functions such as ReLU that tend to increase the number of zero output elements. The

achieved sparsity can either be unstructured [6], or structured [7]. In unstructured sparsity, there is no constraint on the locations of the zeros and multiple metadata are also required to identify the original position of each non-zero element. On the contrary, in structured sparsity, there is an upper limit on the number of non-zero elements that may be present within a block of consecutive elements. The latter approach simplified and indexing and thus the corresponding hardware.

To increase energy efficiency, inference is typically executed using integer arithmetic, after appropriate data quantization [8]. However, recent studies have shown that FP arithmetic cannot be avoided, if one wishes to preserve the inference quality [9]. To enjoy both benefits, i.e., the low hardware cost of integer arithmetic and the accuracy/dynamic range of FP arithmetic, several reduced-precision FP formats of 16-bit or even 8-bit formats have been proposed [10], [11].

Thus, any hardware platform for ML acceleration should be able to handle the inherent diversity in data formats and structure of ML models, it should offer high throughput and should be programmable to cover current and future extensions of ML algorithms.

2 Proposed Architecture

To address these challenges in this work we combine three forms of computation paradigms, each one targeted for different parts of the workload. An organization of the proposed architecture is depicted in Fig. 1.

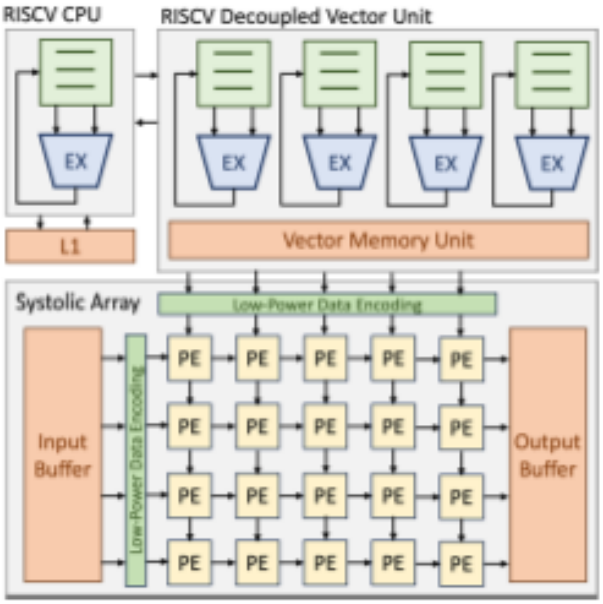


Figure 1 *Decoupled RISC-V CPU organization including a scalar core, a vector unit and a Systolic Array*

A low cost RISCV CPU [12] orchestrates the whole operation and can execute serial code or any unstructured parallelism. The latter is handled by dual issuing of compressed instructions that are sufficient to encode various ML kernels in a compact manner. The scalar code targets also high clock gating efficiency by design. For instance, the core utilizes a partitioned register file that capitalizes on

the skewed use of registers arising by the heavy utilization of the compressed instructions to improve energy efficiency through clock gating.

A decoupled vector core following RISC-V ISA extension can handle vectorized parallelism arising in many edge applications [13] within or without the ML context. The vector core has been enhanced to support structured sparsity in pruned ML models. Custom instructions simplify the indexing and fetching of sparse data without ruining the overall efficiency of SparseMM operation. The vector processor receives the appropriate commands from the decode stage of the scalar core and utilizes its own distinct path to the memory subsystem. L1 caches are private to the scalar core.

A systolic array handles the bulk matrix multiplication needed by ML models. It includes both integer [14] and reduced precision floating point datapaths [15]. Pipeline operation is fine-tuned for each type of number format with the goal of reducing the latency of operation. Also, data streamed in the systolic array are appropriately encoded at the edges of the SA [16] with the goal of reducing the associated dynamic power. Also, upon detecting a zero-input value, the pipeline is 'frozen' (amounting to inserting a bubble). In this case, registers are clock gated, and multipliers are data gated and bypassed since the result of the multiplication is known a priori to be equal to zero.

Data transfer to/from the Systolic Array are performed by buffers placed at the borders of the SA. Data transfer to/from these buffers should be scheduled before actual computation begins. The orchestration of data transfer is currently under development.

3 References

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.

[2] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11 976–11 986.

[3] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing," IEEE Computational Intelligence Magazine, vol. 13, no. 3, pp. 55 – 75, 2018.

[4] K. Tateno, F. Tombari, I. Laina, and N. Navab, "CNN-SLAM: Real-Time Dense Monocular SLAM with Learned Depth Prediction," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6243–6252.

[5] H. T. Kung, "Why systolic architectures?" Computer, vol. 15, no. 1, pp. 37–46, 1982.

[6] U. Evci, T. Gale, J. Menick, P. S. Castro, and E. Elsen, "Rigging the lottery: Making all tickets winners," in International Conference on Machine Learning, Jul. 2020, pp. 2943–2952.

[7] A. Mishra, et al., "Accelerating sparse deep neural networks," arXiv:2104.08378, 2021.

[8] A. Gholami et al., "A survey of quantization methods for efficient neural network inference," arXiv:2103.13630, 2021.

[9] N. P. Jouppi, et al., “Ten lessons from three generations shaped google’s tpuv4i: Industrial product,” in International Symposium on Computer Architecture (ISCA). IEEE, 2021, pp. 1–14.

[10] S. Wang and P. Kanwar, “BFloat16: The secret to high performance on Cloud TPUs,” Google Cloud Blog, vol. 4, 2019.

[11] P. Micikevicius, et al., “Fp8 formats for deep learning,” arXiv preprint arXiv:2209.05433, 2022.

[12] K. Patsidis, D. Konstantinou, C. Nicopoulos, G. Dimitrakopoulos, "A Low-Cost Synthesizable RISC-V Dual-Issue Processor Core Leveraging the Compressed Instruction Set Extension," in Microprocessors and Microsystems, Elsevier, Sept. 2018.

[13] K. Patsidis, C. Nicopoulos, G. Sirakoulis, G. Dimitrakopoulos, "RISC-V2: a Scalable RISC-V Vector Processor", in IEEE International Symposium on Circuits and Systems (ISCAS), Oct. 2020.

[14] C. Peltekis, D. Filippas, G. Dimitrakopoulos, C. Nicopoulos, and D. Pnevmatikatos, “ArrayFlex: A Systolic Array Architecture with Configurable Transparent Pipelining,” in Design Automation and Test in Europe (DATE), 2023.

[15] D. Filippas, C. Peltekis, G. Dimitrakopoulos, and C. Nicopoulos, “Reduced-Precision Floating-Point Arithmetic in Systolic Arrays with Skewed Pipelines,” in IEEE Intern. Conf. on Artificial Intelligence Circuits and Systems (AICAS), 2023.

[16] C. Peltekis, D. Filippas, G. Dimitrakopoulos, and C. Nicopoulos, “Low-Power Data Streaming in Systolic Arrays with Bus-Invert Coding and Zero-Value Clock Gating,” in Intern. Conf. on Modern Circuits and Systems Technologies (MOCAST), 2023.

EMERGING TECH CONFERENCE 2023
Edge Intelligence

Papers

Session 1.4 | AI - Edge AI and ML

Session Chairs: Kostas Siozos

Light Bulb control with DNN based voice user interface – the journey to design it
Georgios Flamis, George Chardalias, Vin D'Agostino and Suad Jusuf

Enabling Grid Resilience and Efficiency - EDGE Device for Power Grid analysis and Asset Monitoring
Sami Hammal, Vagelis Alifragkis, Pavlos Psimadas and Nikolaos-Antonios Livanos

Federated transformers for non-intrusive load monitoring in heat pumps
Stylianos Kanoutas, Athanasios Bachoumis, Michael Birbas and Alexios Birbas

AI-assisted Serious Games: Dialogue Management with Generative AI
Eleni Panopoulou, Davide Aversa and Stavros Vassos

Transforming the Path Towards Automation of Monitoring and Management for Edge Computing
Georgios Samaras, Marinela Mertiri, Maria-Evgenia Xezonaki, Nikos Psaromanolakis, Vasileios Theodorou and Theodoros Bozios

Intelligence Functions Placement in B5G / 6G wireless networks
Vasiliki Lamprousi, Sokratis Barmpounakis, Vera Stavroulaki and Panagiotis Demestichas

HETiA
EXPANDING THE DIGITAL FRONTIERS

ARISTOTLE UNIVERSITY OF THESSALONIKI

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 89 – 92

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Light Bulb control with DNN based voice user interface – the journey to design it

George Chardalias[1], Georgios Flamis[1], Vincent D'Agostino[2] and Suad Jusuf[1]

[1] *Renesas Electronics*
[2] *D'Agostino Industries group*
georgios.chardalias.ks@renesas.com, georgios.flamis.zc@renesas.com, vindag@dagostinoindustries.com, suad.jusuf@renesas.com

Abstract

The purpose of a design journey is to show, in a demonstrative way, how to incorporate the DNN based voice user interface technology into a design. The base designs are by definition very simple to highlight the new technology and how it can be incorporated easily. Distinct lines are drawn between the application and the technology so the reader can more generalize how to incorporate it in their own design.

1 Introduction

Our goal with this Design Journey is to create a control for an LED-based light bulb based on the Renesas RA6/4/2 device family running Cyberon DSpotter for Voice User Interface (VoiceUI) technology. This technology is a non-connected speech recognition solution; it runs totally on the local MCU. While this limits its vocabulary and natural language support, it also allows for speech control of devices that are not connected to the cloud.

To accomplish this, we will design and build a circuit that can plug into any of the RA-Voice kit boards via the PMOD connector and write software to support the design. The software project accompanying this Design Journey is targeting the RA6 Voice Kit.

2 The Voice UI

One of the major differences between a connected, natural language solution, and a locally hosted speech recognition solution, is that a non-connected speech recognition system requires a stricter vocabulary set and syntax. This is because the interface is expecting certain words or phrases and cannot process natural language which can reorder words and still have the same meaning. This is the tradeoff for being able to run the speech recognition on an M-class device with a small memory footprint. Multiple language support (Cyberon supports 44 at the time of publication of this document) can be achieved by loading multiple language sets and switching between them at runtime.

For our lightbulb, we have created a system that, depending on the desired outcome, will take two or three commands to complete a task. In graphic form, it looks like the following drawing:

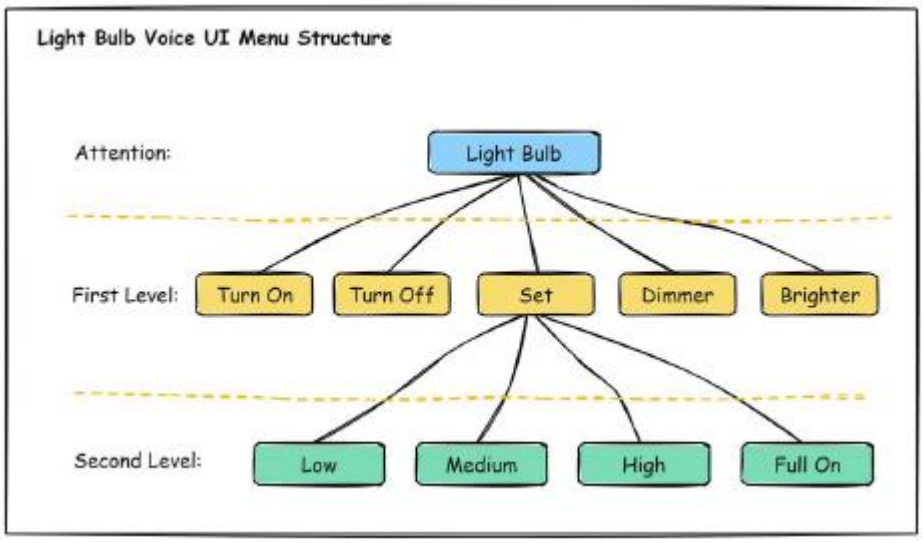


Figure 1 Voice Command options

As can be seen above, an attention word is first used to let the system know a command is coming. There are four commands ("Turn On," "Turn Off," "Brighter," and "Dimmer") that require no more information and complete the command. If the command word "Set" is used, then one of four preset levels is required next, to complete the command. "Dimmer" and "Brighter" traverse the presets one level at a time, and "on" and "off" move between extinguished and the last used preset level. Let's look at what that means:

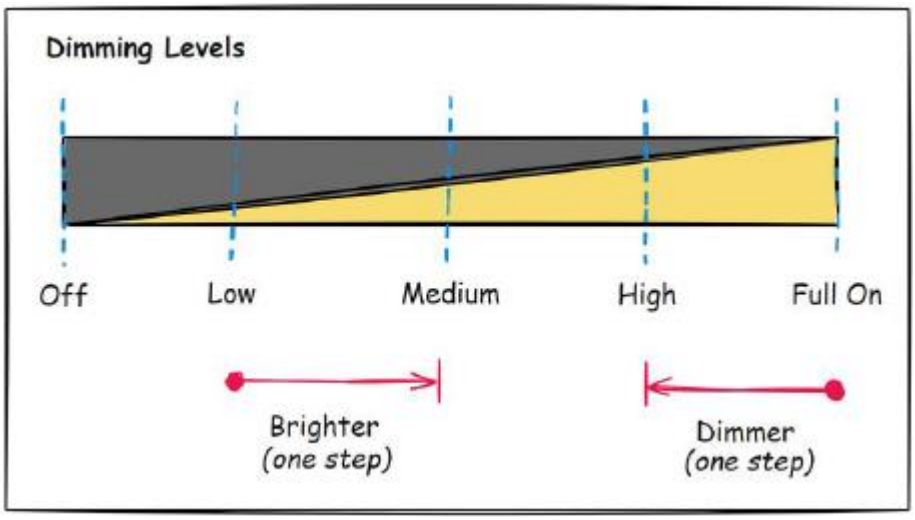


Figure 2: Light dimming levels

This defined command structure is easily created using the Cyberon DSpotter Modelling Tool and tested for recognition with their powerful command testing tools.

Unlike a "trained voice" solution, Cyberon has "pre-trained" the system for units of language sound, called phonemes. Each language has a set of phonemes. English, for example, has between 42 and 45 phonemes. Italian, by contrast, has 32-36. The DSpotter Modelling tool breaks down command

text into these phonemes, which are pre-categorized. There is no need to take thousands of voice samples to train the system as you would with traditional speech recognition.

3 Hardware design

The hardware design for the bulb is relatively simple because the RA devices have timers with PWM outputs which can drive transistor gates directly. The use of transistors is driven by two things: firstly, the GPIO pins of any processor cannot handle the currents required to drive an LED at illumination levels, secondly, the forward voltage drop of white LEDs is very close to the 3.3V Vcc used by the processor. Typically, the LEDs are driven with a higher voltage. The processor pins cannot tolerate this higher voltage, so to isolate the I/O pins from both this voltage and current requirement, we can use small MOSFETs to drive the LEDs.

Since the PMOD connector has ground and 3.3V, we must create a higher voltage. For this, we are using a Renesas part on our plug-in board, the ISL9111A buck-boost controller, with the output set to 5.25VDC. The following illustration shows the overall design of the LED plug-in board.

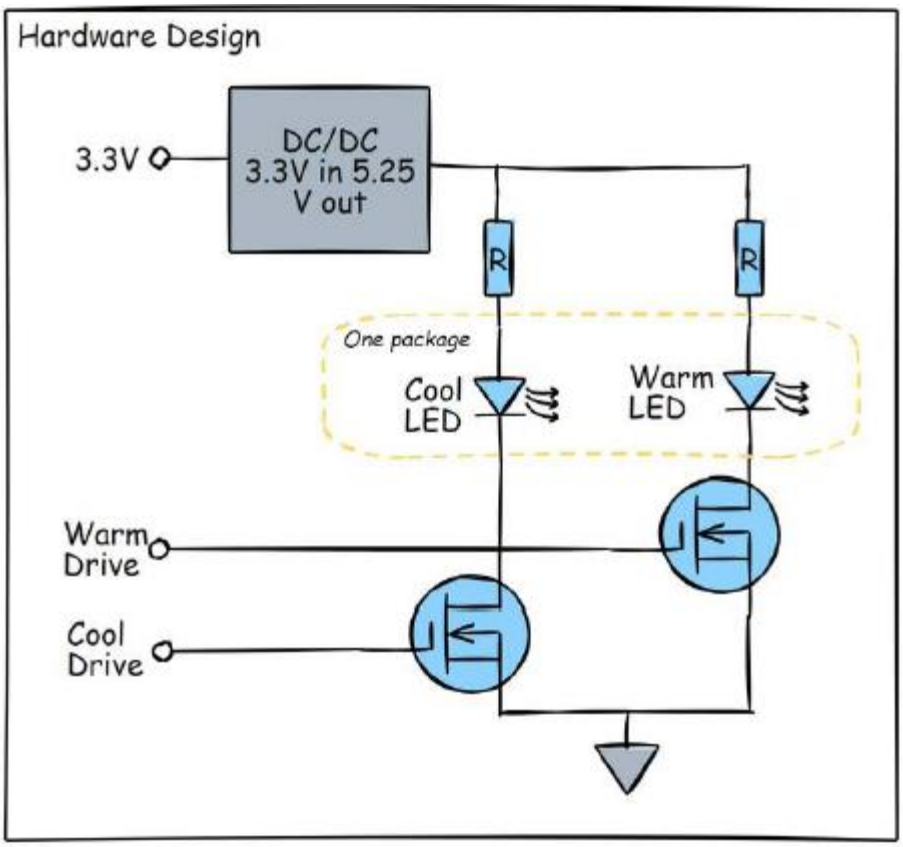


Figure 3: Drawing of the PWM driver

We will drive the two transistor gates with the output from a PWM running with a carrier frequency of 4kHz. This is well above the optical detection of the eye – even in the periphery where we are more sensitive, and well below where the switching transitions occupy a high percentage of overall time (this, in the end, is what makes transistors run hot – not what percentage of time they are on, but the percentage of time they spend in the linear region during switching).

The hardware for the processor and its peripherals is the one supplied with the voice kit. It has a processor, two digital and two analog mics, and a PMOD connector. Our LED board plugs into the PMOD connector. The other connection is the USB-micro to the PC for debug and serial monitoring (the JLINK-OB debugger creates a VCP along with the debug connection).

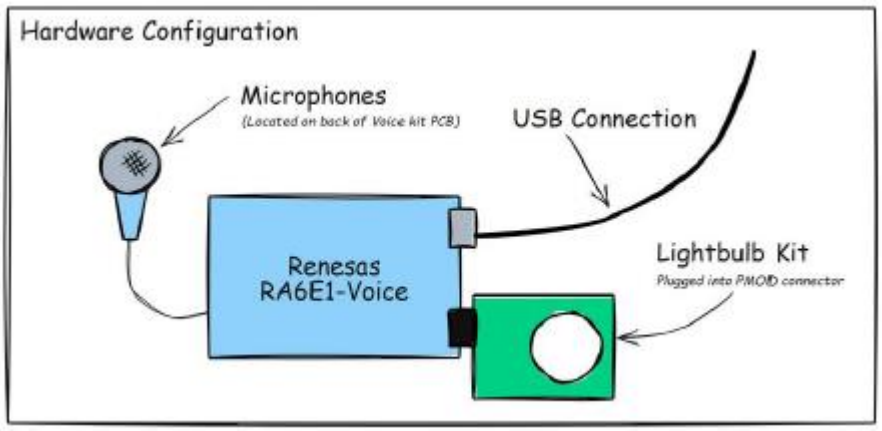


Figure 4: Hardware Configuration

4 Conclusion

Over the previous sections, it has been shown that incorporating the Renesas/Cyberon VoiceUI into products is fast, easy, straightforward, and very maintainable. In this design, the use of a linked hardware solution makes it easy to do two time-dependent tasks together without the overhead of an RTOS to manage the processes.

While this solution generates data for the outside world (the PWMs) the same opportunity is there for data acquisition with digital and analog peripherals.

Renesas has always had a rich set of high-performance peripherals: both digital and mixed mode. Together with the Event Link Controller (ELC) in the RA family, these peripherals can be configured to do many functions independently.

5 References

[1] VOICE-RA6E1 Engineering Manual, Renesas Electronics, 2022

[2] RA6E1 Group Datasheet, Renesas Electronics, 2022

[3] Cyberon Speech Recognition Technology, Renesas Electronics, 2023

[4] Endpoint Embedded Voice User Interface Solution, Renesas Electronics, 2022

[5] Designing Industrial IoT Applications with TinyML, Mouser Electronics, 2023

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 93 – 99

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Enabling Grid Resilience and Efficiency - EDGE Device for Power Grid analysis and Asset Monitoring

Sami Hammal[1], Vagelis Alifragkis[1], Nikolaos Giamarelos[1], PavlosPsimadas[1], Nikolaos-Antonios Livanos[1]

[1] *EMTECH SPACE S.A., Greece*

Abstract

The integration of renewable energy sources (RES) into power distribution grids is crucial for achieving global green energy objectives. However, RES power inverters can adversely impact power quality, increase losses, and lead to power interruptions due to issues such as harmonics, frequency deviation, and grid fault response limitations. To address these challenges, this paper presents an innovative solution combining hardware and software components based on Phasor Measurement Unit (PMU) and EDGE processing technologies that provide the ultimate tools for real-time monitoring of the distribution power grid. The proposed solution, designed for research and innovation activities, features a low-cost, high-performance architecture that facilitates real-time monitoring of secondary distribution substations.

1 Introduction

Today, various devices and equipment are used to monitor the power grid key assets to provide operators with a detailed picture about the state of their grid. Phasor Measurement Units (PMUs) play a significant role in acquiring and digitizing the voltage and current phasors with high sampling frequencies. By using advanced signal processing and correlation techniques on the data received from several PMUs, a high-quality analysis of the dynamic behavior of the electric power grid can be performed. Moreover, EDGE computing paved the way for de-centralized operation in distribution power grid application.

This integrated solution described is based on an open PMU architecture with EDGE processing capability (Nikolaos-Antonios I. Livanos, 2023) and it offers significant advantages in enhancing grid resilience, efficiency, and eventually CO2 reduction. By leveraging the EDGE processing capability of the PMU platform, real-time asset monitoring and diagnostics are achieved. The device offers continuous monitoring of key parameters and enables the detection of grid anomalies, facilitating fast and automated fault detection, isolation, and recovery. Moreover, the solution supports state estimation, power quality monitoring, and dynamic events analysis, among other data-driven valueadded services.

The hardware device is equipped with a Digital Signal Processor, a high-end processor capable of running a Linux OS, and a phasor measurement unit. It is deployed at secondary distribution substations to enable comprehensive analysis of the grid's performance and to facilitate proactive

maintenance strategies. The use of EDGE computing capabilities and cloud technologies allows for scalable deployment and efficient management of the system. A pilot implementation of the prototype design has been successfully deployed in multiple medium-to-low-voltage substations across Cyprus.

This paper showcases the successful implementation of the solution, highlighting its potential to revolutionize power grid analysis and diagnostics. The combination of hardware, software, and advanced monitoring capabilities demonstrates the effectiveness of the solution in achieving the goals of grid resilience, efficiency, power theft detection, and CO2 reduction. The results obtained from the pilot demonstration phase validate the performance and value of the proposed system, paving the way for future research, development, and widespread adoption in the energy sector.

2 System Overview

The overall system consists of the following four major components:

1. Secondary Transformer Monitoring (STM) devices, which serve as a data acquisition point for measuring various variables related to the grid and the transformer.
2. The central application (SCADA), which takes on the role of managing and coordinating the operations of these edge devices. It oversees their operations, retrieves the acquired data from each device, stores them in the central database, and handles tasks that demand more computational resources.
3. The graphical user interface is a web-based environment that provides real-time monitoring representation of the system using graphs, time plot elements, alert notifications as well as the option to view past stored data.
4. The AI-based algorithms are fed with the updated data and executed iteratively to provide more accurate results. These algorithms are mainly specialized for power theft detection, but other applications are also supported, like load prediction, islanding, power flow, and optimal power flow.

Figure 1 illustrates the top-level design of the overall system, providing a visual representation of its high-level structure and functionality.

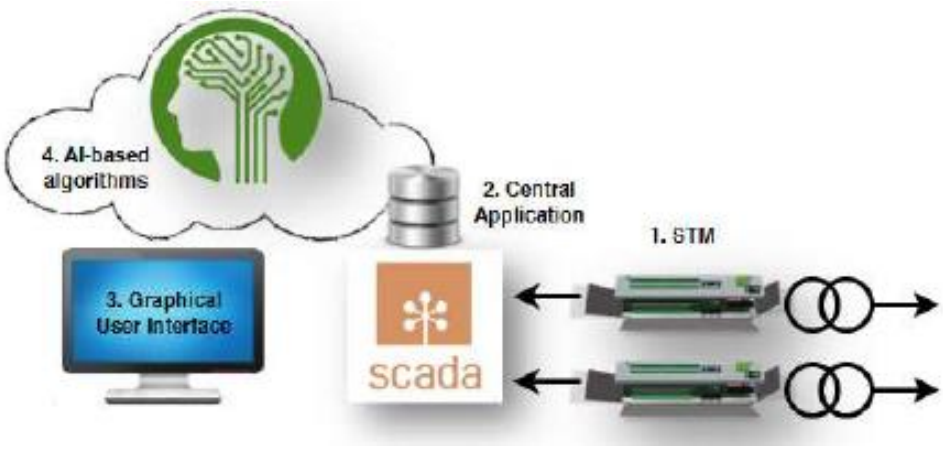


Figure 1: Top-level logical diagram of the solution

3 Hardware Description

The STM device supports easy installation through non-invasive and non-disruptive means. It is divided into two parts: the primary part housing the main components, and the secondary part housing external sensors mounted on the transformer. Figure 2 illustrates the logical structure of the STM device, providing a visual representation of its components and their relations. Specifically, the following components comprise the hardware device:

1 **Gateway (GW):** this is essentially the module that provides the EDGE capability, which is based on the OSD3358 System-in-Package (SiP) developed by Octavo Systems (Octavo systems; Rev, I.N.D. 2022). It can run a Linux operating system and it is responsible for collecting the measurements from the PMU and the RTU, storing them into its database, handling requests from the central application, and synchronizing all the data between the device and the server.

2 **Phasor Measurement Unit (PMU):** this includes an analog front-end that interfaces with the transformer to acquire the voltage and the current measurements. It is comprised of an analog-todigital converter (ADC) for signal sampling and a Digital Signal Processor (DSP, based on TI's low-power C55x fixed-point DSP TMS320C5517) (Texas Instruments. TMS320C5517 Fixed- Point Digital Signal Processor) that executes its own firmware regarding data acquisition, storage, and processing.

3 **Remote Terminal Unit (RTU):** this includes a simple 8-bit microcontroller running its own firmware for handling the communication between the external sensor and the gateway via the RS-485 interface, regarding data transfer and control.

4 **Power supply:** this includes two AC/DC converters to distribute power to the internal components. One converter supplies power to the 4G modem (12V DC), and the other powers the remaining components (5V DC). Additionally, a UPS is utilized to ensure uninterrupted operation during power outages.

5 **4G Modem:** configured with a static IP SIM card, this is used for remote communication and control of the STM device from the central application. It also establishes a local network in the installed substation, to provide seamless connectivity with external devices when it is necessary.

6 **GPS:** this is utilized for enabling synchronized data sampling across multiple STM devices by using the pulse-per-second (PPS) signal.

7 **External Sensor:** this is the secondary part, which includes a set of board sensors for measuring the temperature, humidity, magnetic field, and acoustic noise. It utilizes a microcontroller that interfaces with each sensor, and it is responsible for collecting the measurements and handling the data requests from the RTU via the RS-485 interface.

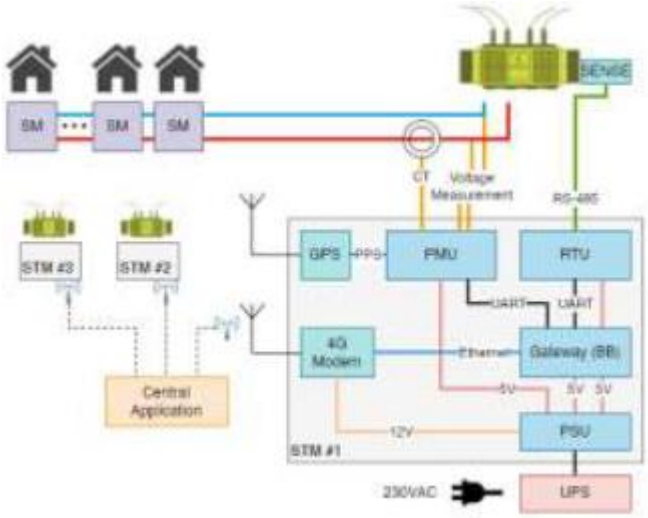


Figure 2: Logical structure of the STM device.

4 Application Results

One of the outcomes of the application is the demonstration of power theft detection. Specifically, two power theft detection methods were utilized, both based on analyzing the collected measurements and identifying discrepancies in technical losses. The first method is done by examining the statistical characteristics of technical losses for a given secondary substation within a certain period to identify potential instances of power theft. Essentially, a set of measurements from both the STM device and the smart meters must be made available to perform the analysis.

The calculation of technical losses is performed using the equation (1) outlined below:

$$E_{STM} = \sum E_{SMi} + E_{TL} \qquad (1)$$

The calculated energy by the STM device (E_{STM}) is equal to the sum of each smart meter energy (E_{SMi}) plus the energy of the technical losses (E_{TL}). Given that E_{STM} and E_{SMi} are known quantities, equation (1) can be solved to calculate the energy of the technical losses.

In general, the energy of the technical losses remains around the same average value, but even if it changes, it happens due to factors unrelated to power theft (e.g., seasonal weather, equipment wearing, etc.). However, if a sudden difference is observed, then this may be considered a potential instance of power theft, in which case further investigation must be performed to confirm or rule out the possibility.

Figure 3 below shows a plot depicting the energy measurements from the STM device, the smart meters, and the technical losses. Two distinct periods are highlighted where abrupt changes in the technical losses' values were detected. These discrepancies indicate the potential occurrence of power theft and showcase the capability of the method to detect such events.

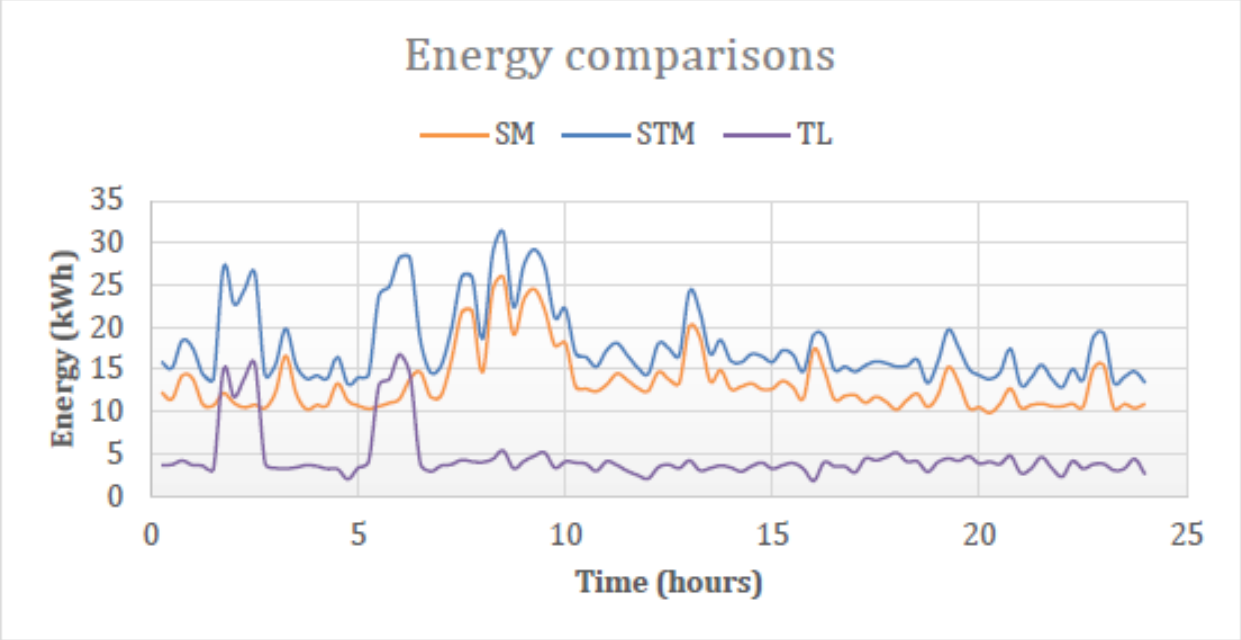


Figure 3: Energy measurements comparisons showing a case of power theft.

The other method to power theft detection involves using the same historical consumption data from the smart meters and comparing it to the values obtained through sequential power flow simulations. In these simulations, different consumers are selected as the slack bus each time. If there is a difference between the measured consumption from a smart meter and the simulated consumption, then this indicates the possibility of an electricity theft event occurring at that specific position in the grid. To perform the power flow method the Pandapower Python library is utilized, which is commonly used for power system modelling, analysis, and optimization (Thurner, et al., 2018), and it has been successfully applied in multiple grid studies (Menke, 2018) (Scheidler, Thurner, & Braun, 2018) (Thurner, Scheidler, Probst, & Braun, 2017) (Wang, 2017).

5 Pilot Operation

The implementation of the solution has been successfully carried out as part of the "Technorevmeta" project during the technology demonstration phase. This collaborative effort involved the Electricity Authority of Cyprus (EAC), the Frederick University in Cyprus, and the Photos Photiades Group, a beverage company, which allowed us to conduct a pilot demonstration at one of their production facilities.

The devices have been in operation for more than a year, continuously measuring the necessary parameters. The central application has been successfully deployed, enabling constant monitoring of communication with the edge devices. It efficiently retrieves data from them every minute and stores it in the central database.

Figure 4 shows its deployment in the secondary substation adjacent to the transformer, and Figure 5 shows the external sensor mounted on the transformer.

Figure 4: STM device installed at the pilot location

Figure 5: External sensor installed on the transformer.

6 Conclusion

In the context of digitizing electrical grids, various challenges emerge, including the need for effective asset monitoring, especially for transformers, and the detection of power thefts. This paper outlines the development of a comprehensive solution that addresses these challenges, making use of the processing capabilities and real-time communication facilitated by the Secondary Transformer Monitoring (STM) EDGE devices installed on secondary distribution transformers. The STM devices

play a crucial role in enabling computationally intensive operations, allowing for the analysis of measured data through AI models. The primary objective of this analysis is to identify anomalies in energy generation and consumption that may indicate instances of power theft. A pilot demonstration has already been conducted, capturing, and storing valuable data to train the AI models. The initial results obtained for power theft detection show great promise, and as the training process continues and more data becomes available, these models are expected to exhibit even more significant improvements in the coming months.

7 Acknowledgement

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T2EDK-02497).

8 References

[1] Menke, J.-H. &. (2018). Heuristic Monitoring Method for Sparsely Measured Distribution Grids. International Journal of Electrical Power & Energy Systems.

[2] Nikolaos-Antonios I. Livanos, S. H. (2023). OpenEdgePMU: An Open PMU Architecture with Edge Processing for Future Resilient Smart Grids. Energies.

[3] Octavo systems; Rev, I.N.D. 2022. (n.d.). Retrieved from https://octavosystems.com/octavosystems.com/wp-content/uploads/2017/08/OSD335x-SMDatasheet.pdf

[4] Scheidler, A., Thurner, L., & Braun, M. (2018). Heuristic Optimization for Automated Distribution System Planning in Network Integration Studies. IET Renewable Power Generation.

[5] Texas Instruments. TMS320C5517 Fixed-Point Digital Signal Processor. (n.d.). Retrieved from https://www.ti.com/lit/ds/symlink/tms320c5517.pdf

[6] Thurner, L., Scheidler, A., Probst, A., & Braun, M. (2017). Heuristic Optimization for Network Restoration and Expansion in Compliance with the Single Contingency Policy. IET Generation, Transmission and Distribution.

[7] Thurner, L., Scheidler, A., Schäfer, F., Menke, J.-H., Dollichon, J., Meier, F., . . . Braun, M. (2018). Pandapower—An Open-Source Python Tool for Convenient Modeling, Analysis, and Optimization of Electric Power Systems. IEEE.

[8] Wang, H. &.-v. (2017). Reactive Power Coordination Strategies with Distributed Generators in Distribution Networks. 1st International Conference on Large-Scale grid Integration of Renewable Energy in India.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 100 – 106

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Federated transformers for non-intrusive load monitoring in heat pumps[*****]

Stylianos Kanoutas, Athanasios Bachoumis, Michael Birbas, and Alexios Birbas

University of Patras, Department of Electrical and Computer Engineering. Patras, Greece
ece8346@upnet.gr, {abachoumis, mbirbas, birbas}@ece.upatras.gr

Abstract

Edge intelligence, i.e., the execution of Machine Learning (ML) algorithms in computing resources at the edge, provides unprecedented benefits for applications in different verticals regarding data privacy, bandwidth, costs, and latency. Non-Intrusive Load Monitoring (NILM) is an application in the smart grid technology domain that could benefit from the advancements in edge intelligence to ensure consumer data privacy and decrease implementation costs. This paper proposes a federated learning-based transformer architecture for the NILM of energy-intensive residential devices, i.e., Heat Pumps (HP). We evaluate the architecture on an open-source dataset, showcasing that the performance does not deteriorate significantly compared to the centralized and is robust against the distribution shifts between the training and inference datasets and the increasing heterogeneity level between clients in the training data.

1 Introduction

Non-intrusive load monitoring (NILM) is a key technology that enables smart grids to monitor and disaggregate individual appliance energy consumption [1]. By providing granular energy consumption data, NILM can help ensure system stability, enable flexibility services, and support demand-side management, improving smart grids' overall efficiency and sustainability. NILM is an enabler for local aggregators to unlock the real-time demand-side flexibility potential of energy-intensive devices, by avoiding the capital expenses of installing new meters. Heat pumps (HP) are controllable loads that provide heating and cooling services, assisting the electricity grid's supply and demand balancing. Moreover, HP can be controlled through demand response programs to shift their electricity consumption to off-peak hours, reducing the strain on the grid during peak periods. However, some HP may connect to the main electrical meter of the building; thus, their consumption cannot be measured directly.

Privacy concerns are raised about end-consumer data in the digitalized era of power grids, where edge intelligence is an emerging topic, i.e., the deployment of ML algorithms and models on devices located at the edge, such as smart meters and gateways. To tackle this issue, a decentralized ML technique, named Federated learning (FL), enables edge devices to collaboratively train a shared

[*****] This research has been partially financed by the EU Horizon 2020 project ENERMAN (GA: 958478).

model while keeping the training data locally. Lately, FL has been applied in smart grids [2]. Employing FL for NILM can address the privacy concerns associated with collecting and centralizing fine-grained energy consumption data, while not significantly decreasing the accuracy compared to centralized learning paradigms. Numerous NILM techniques have been developed, including probabilistic procedures (e.g., Hidden Markov Model), ML techniques such as support vector machines [1], and deep learning models [3, 4]. In addition, works have emerged employing FL to collaborative train of shared NILM models [5, 6].

In that direction, this paper contributes to the literature by applying for the first time an FL-based architecture, using Transformers as a model for the NILM of HP. It evaluates the architecture on an open-source dataset by comparing the accuracy versus other learning architectures, i.e., centralized and personalized learning. Finally, this work investigates the architecture's performance in unseen data while encountering for distribution shift between the training and test sets and heterogeneity among clients' data.

2 Problem statement and proposed solution

2.1. Problem Details

The primary goal of NILM algorithms is to deduce the consumption patterns of household appliances by extracting patterns from the aggregated power consumption data at a smart meter level. This procedure involves breaking down the overall power signal of a household into its constituent parts, namely the power signal of each appliance. This is expressed by:

$$P(t) = \sum_{i=1}^{N} p_i\,(t) \; + \; p_{noise}(t) \tag{1}$$

where N represents the total amount of appliances, and i denotes the index for the i_{th} appliance. At any given time t, the total power consumption $P(t)$ equals the combined power usage of all appliances N, represented by pi. The $+\, p_{noise}$term denotes the noise. Another valuable purpose of the NILM is to classify the state of the *i-th* appliance except for the latter. We assume that the HP display two states (on/off), and for this categorization, a threshold is set to discriminate between those states. Additionally, we set a minimum on- and minimum off-time (3 minutes each) to avoid wrong predictions due to momentary inaccurate data from sensing.

2.2. Federated Transformers

Our study uses a Federated Transformer for NILM application in HP. The employed model architecture is the ELECTRIcity [7]. The model comprises two transformer layers with two attention heads and a hidden size of 64. In the encoding layer, a 1D convolutional layer is implemented for feature extraction, followed by a squared average pooling layer. For the decoding side, a de-convolutional layer is employed. Additionally, we apply a 10% dropout probability to all layers. In contrast with the original training process of ELECTRIcity, we discarded the pre-training unsupervised stage and, consequently, the generator-discriminator implementation, as we tried to keep our algorithm as lightweight as possible. In the supervised training process, each model gets the aggregated data for the smart meter and the HP signal as an input. The loss function L is defined:

$$L = \frac{1}{N}\sum_{i=1}^{N} \; (\hat{y}_i - \; y_i)^2 D_{KL}\left(\text{softmax}\left(\frac{\hat{y}}{\tau}\right) || \text{softmax}\left(\frac{y}{\tau}\right)\right) + \frac{1}{N}\sum_{i=1}^{N} log(1 + exp(\hat{s_i}\, s_i)) + \frac{\lambda}{N}\sum_{i \in o} |\hat{y}_i - \; y_i| \tag{2}$$

where λ represents a hyperparameter that governs the influence of the absolute error derived from the collection O of inaccurately predicted samples and time points when the appliance's state is active. During the inference, the input of the federated model is the smart meter's data and the output is the prediction of the HP consumption. The loss function considers the appliance's true status and the predicted consumption signal's on-off status, denoted as *s*.

Federated transformers for NILM Kanoutas, Bachoumis, M. Birbas and A. Birbas

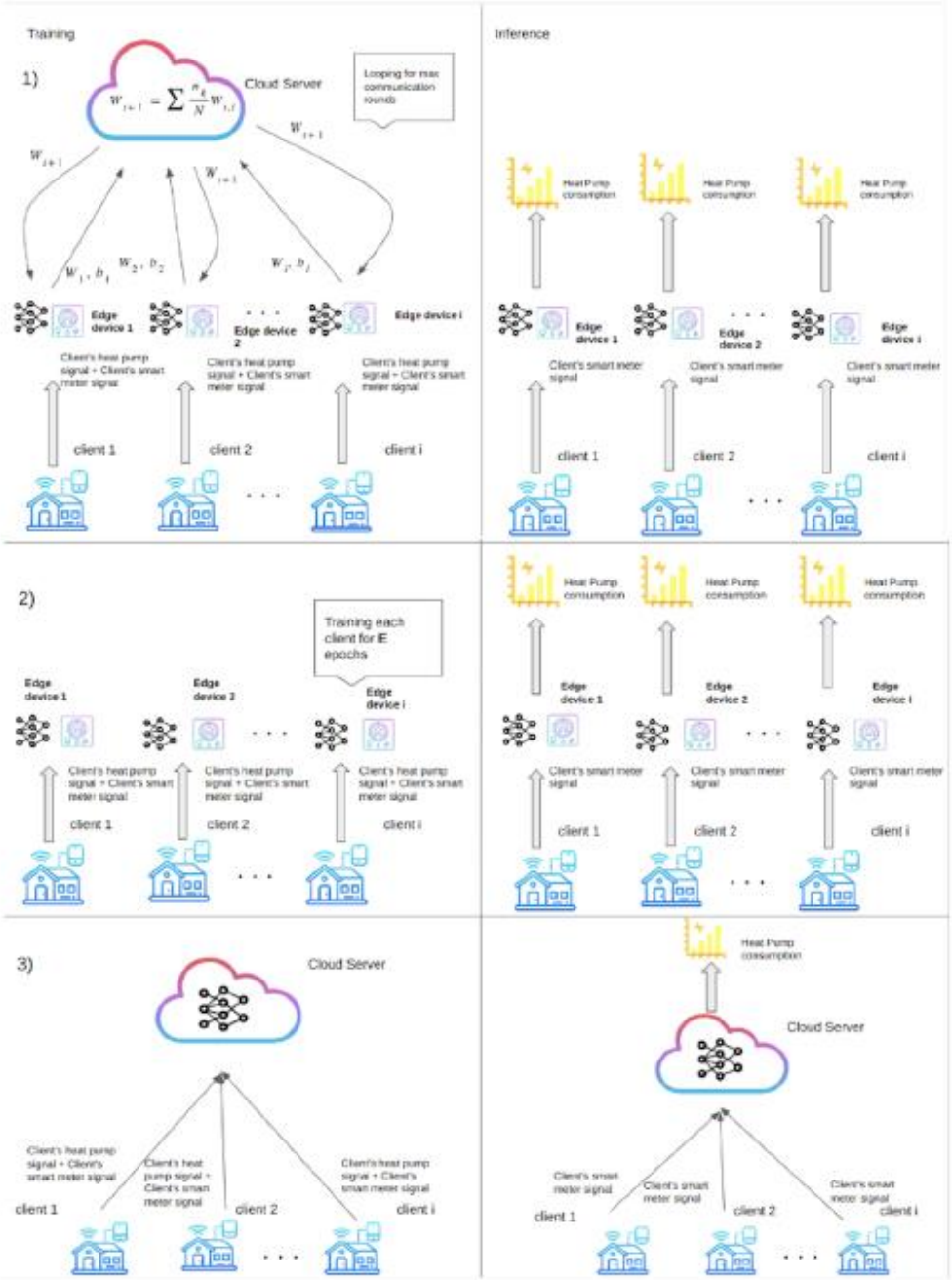


Figure 1: Different learning architectures for training and inference modes. Upper: FL-based, Mid:

Personalized, and Lower: Centralized aggregated.

For the FL, we applied the FedAvg algorithm. Particularly, all clients are equipped with an edge device, e.g., a computationally capable smart meter. Each edge device trains a local model, with the specific household's data. We set the number of inner epochs to 1 and the max communication rounds to 49. We utilized the Adam optimizer with a batch size of 64 and set the λ and τ values equal to 1 and 0.1, respectively. The deployed architecture is illustrated in Fig. 1.

3 Algorithm evaluation case study

We utilized *WPuQ*, an open-source dataset, for training our model and evaluating its performance [8]. The dataset comprises electricity measurements of total household consumption and HP consumption from 38 single-family houses located in Northern Germany with resolution ranging from 10 seconds to 60 minutes. We used the 1-minute resolution of the active power data from 23 of the 38 houses. We excluded houses with limited data available from the analysis and those equipped with photovoltaic (PV) systems to use them further to validate the framework in unseen data. Data preprocessing takes place so that they comply with the following properties: Maximum limit (the upper limit that we set to protect our model from spikes that may distort it) is set to 12 kW and 9 kW for the aggregate signal and HP signal, respectively. The "on threshold", i.e., the value defining the on/off status, is set to 100 W. Finally, we also set the minimum off/on values equal to 3 min, i.e., the minimum time the HP signature signal should be greater or lower than the "on threshold" for the model to capture the on or the off status, respectively. The training set includes the data from 2018 and 2019, where we split it into 90% for the training and 10% for the validation set. The model was tested on out-of-sample data from the year 2020. In addition, the training set was split into windows of 480 samples with a stride of 120 samples.

We evaluate the performance of our architecture, by setting two learning architectures, i.e., personalized and centralized learning (mid and lower plots in Fig. 1). For each test case, we utilize 23 clients of the WPuQ dataset, either by aggregating the clients' HP data in the server (centralized case) or by developing personalized models for each client separately at the edge without any weight-sharing between the clients and the server (personalized case).

To assess model performance, we employ both regression and classification metrics. The regression metrics evaluate the model's ability to predict HP consumption, while the classification metrics evaluate the model's ability to predict the on-off status of the appliance. Regarding the former, we employ:

$$MRE = \frac{1}{man(Y)} \sum_{i=1}^{N} |\hat{y}_i \, y_i| \quad (3) \qquad MAE = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i \, y_i| \quad (4)$$

where y_i and $\hat{y}_i$ denote the ground truth and the model prediction, respectively. For the latter, we use:

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (5)$$

where TP, TN, FP, FN denote True Positive, True Negative, False Positive, and False Negative, respectively. Besides those metrics, we also utilize two statistical approaches: (i) The Kolmogorov-Smirnov (KS) test to measure the level of non-independent and identically distributed (non-IID) data between all the client's data in a specific training scenario and (ii) the Kullback-Leibler (KL) divergence to measure the distribution shift between the training and test sets.

4 Results analysis

4.1. Test cases comparison

In our study, we evaluate the federated implementation by comparing the results to 2 other test cases, the personalized and the centralized ones. For the training and inference stage of the test cases, we utilize 23 clients from the WPuQ dataset. On those 23 clients, the mean metrics comparison is displayed in Fig. 2. Regarding accuracy, F1-score, and MAE approaches, the results of the FL are close to those of the centralized architecture. This indicates that our proposed privacy-preserving method has no significant deficit in performance compared with the centralized and personalized architectures, with the centralized one having the lowest MAE. The personalized case performs slightly better than the federated one due to the individualized training conducted for each model, specifically tailored to each edge device.

We also apply the federated global model to unseen clients (clients without historical data). Unlike the training clients, the data from the unseen clients include the PV production signal [PV capacity: 0.75 kW (SFH26) and 4 kW (SFH33)], which is aggregated with the consumption signal. The test results [MAE & MRE & ACCURACY & F1-SCORE ‖ SFH26: 143.66 & 0.93 & 0.93 & 0.71 ‖ SFH33: 185.45 & 0.9 & 0.83 & 0.62] reveal that our FL model exhibits satisfactory performance, especially in accuracy when applied to unseen clients. We observe a decrease in efficiency, specifically for the SFH33 client, which contains a larger PV. This suggests that there could be a decrease in the performance as the PV size increases. However, further research is required to draw a secure conclusion.

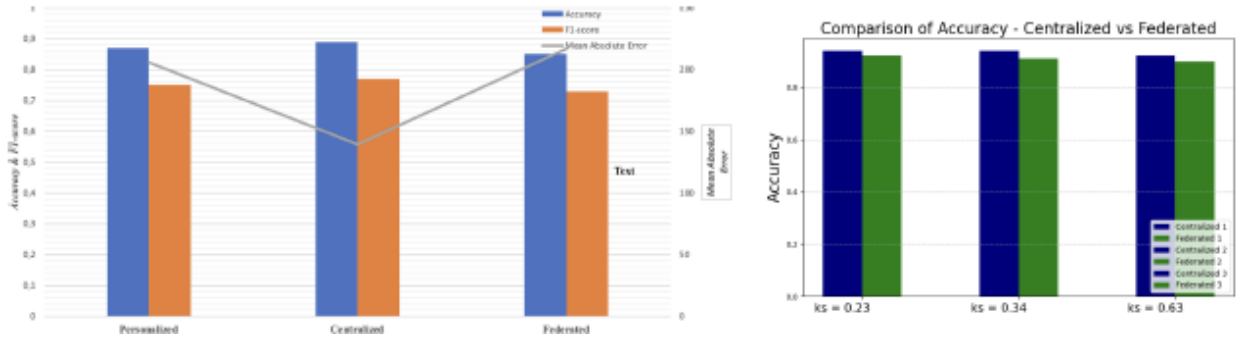


Figure 2: Left: Comparison for the different learning architectures, Right: Impact of non-IID data on testing accuracy between centralized model and federated model.

4.2. Impact of distribution shift on algorithm performance

Fig. 3 presents the relationship between each client's individual inference results and the distribution shift level between its training and test sets. It demonstrates that the model's architecture exhibits no significant decline in the performance associated with an increase in distribution shift until the KL-divergence surpasses a threshold of 4. At this point, there is a notable drop in the model efficiency. Therefore, it can be deduced that the federated transformer architecture is fairly robust to distribution shifts.

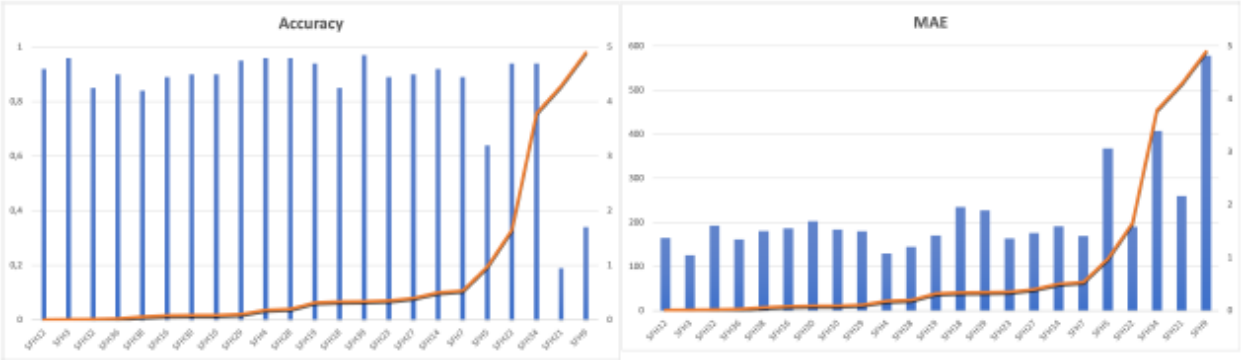


Figure 3: Impact of distribution shift level (KL-divergence with gold line) on federated clients testing accuracy (left plot) and MAE (right plot).

4.3. Impact of non-IID data on algorithm's performance

Furthermore, we estimate the impact of non-IID data on our model's performance. Non-IID data refers to situations where the training and test data distributions significantly differ among clients or households. The FL sensitivity in scenarios involving imbalanced or non-IID data is well-known. Fortunately, this limitation is effectively addressed by integrating the FL algorithm with the transformer architecture, thereby mitigating the effects of non-IID data [9]. Fig. 2 depicts the effect of different levels of non-IID data on the FL algorithm's accuracy. We tested the FL transformer architecture in 3 different scenarios with different non-IID data levels, i.e., 0.23 (first pair of bars), 0.34 (second pair of bars), and 0.63 (third pair of bars). In each of those scenarios, we have at least 8 clients. The results showcase the robustness of our model in handling non-IID data, as evidenced by the absence of any accuracy degradation while increasing the non-IID data level when compared with the accuracy of the centralized architecture.

5 Conclusions

The privacy-by-design principle is the backbone of any solution that ingests the raw consumer's data. In this work, we showcased the application of federated transformers for the NILM of HP, exhibiting performance similar to centralized and personalized cases. We demonstrated the advantages of transformer architecture in dealing with heterogeneous data and its robustness toward distribution shifts. Further experiments will be conducted to validate our method in different datasets while making adaptations in the transformer to increase accuracy.

6 References

[1] Weicong Kong, Zhao Yang Dong, David J Hill, Jin Ma, JH Zhao, and FJ Luo. A hierarchical hidden markov model framework for home appliance modeling. IEEE Transactions on Smart Grid, 9(4):3079–3090, 2016.

[2] Dimitris Mourtzis, John Angelopoulos, and Nikos Panopoulos. Integration of federated learning to smart grid for efficient and secure energy distribution. In Proceedings of the Changeable, Agile, Reconfigurable and Virtual Production Conference and the World Mass Customization & Personalization Conference, pages 477–486. Springer, 2023.

[3] Weicong Kong, Zhao Yang Dong, Bo Wang, Junhua Zhao, and Jie Huang. A practical solution for non-intrusive type ii load monitoring based on deep learning and post-processing. IEEE Transactions on Smart Grid, 11(1):148–160, 2019.

[4] Akriti Verma, Adnan Anwar, MA Mahmud, Mohiuddin Ahmed, and Abbas Kouzani. A comprehensive review on the nilm algorithms for energy disaggregation. arXiv preprint arXiv:2102.12578, 2021.

[5] Shuang Dai, Fanlin Meng, Qian Wang, and Xizhong Chen. Federatednilm: A distributed and privacy-preserving framework for non-intrusive load monitoring based on federated deep learning. In 2023 International Joint Conference on Neural Networks (IJCNN), pages 01–08. IEEE, 2023.

[6] Yu Zhang, Guoming Tang, Qianyi Huang, Yi Wang, Kui Wu, Keping Yu, and Xun Shao. Fednilm: Applying federated learning to nilm applications at the edge. IEEE Transactions on Green Communications and Networking, 2022.

[7] Stavros Sykiotis, Maria Kaselimi, Anastasios Doulamis, and Nikolaos Doulamis. Electricity: An efficient transformer for non-intrusive load monitoring. Sensors, 22(8):2926, 2022.

[8] Marlon Schlemminger, Tobias Ohrdes, Elisabeth Schneider, and Michael Knoop. Dataset on electrical single-family house and heat pump load profiles in germany. Scientific Data, 9(1):56, 2022.

[9] Liangqiong Qu, Yuyin Zhou, Paul Pu Liang, Yingda Xia, Feifei Wang, Ehsan Adeli, Li Fei-Fei, and Daniel Rubin. Rethinking architecture design for tackling data heterogeneity in federated learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10061–10071, 2022.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 107 – 110

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

AI-assisted Serious Games: Dialogue Management with Generative AI

Eleni Panopoulou*, Davide Aversa, Stavros Vassos

Helvia Technologies, Evrystheos 2, Athens 11854, Greece
** Corresponding author. Tel.: +30-694-612-5580; E-mail address: eleni.panopoulou@helvia.io*

Abstract

Cutting-edge technologies such as Virtual Reality (VR) and Artificial Intelligence (AI) have helped elevate the digital storytelling and gaming experience. At the same time, powerful new Large Language Models (LLMs), such as ChatGPT by OpenAI, have taken over the Information Technology (IT) world promising amplification of the user experience in multiple contexts. This paper sets out to explore the utilisation of ChatGPT for managing the dialogue between players and avatars in serious games. Our work is performed in the context of the LAW-GAME project that develops a social AI-powered VR game platform targeting specifically the training needs of Law Enforcement Agencies. Specifically, we focus on one of the platform's gaming modes, namely the Police Interrogation, which aims at a realistic interactive dialogue between a human player (the police officer) and an AI-assisted non-player character (the suspect).

Keywords: Generative AI; ChatGPT; Serious games; Dialogue management; Storytelling

1 Introduction

Serious games have become increasingly popular in recent years as they offer an engaging and interactive way to learn, and can support cognitive, behavioural, affective, motivational, physiological, and social learning outcomes [1]. Enhanced with cutting-edge technologies such as 3D VR worlds and AI agents, serious games can offer training on realistic scenarios and complex situations that can be conducted collaboratively across geographical and professional boundaries [2]. Hence, serious games have seen a rise especially for Law Enforcement Agencies (LEAs) such as police forces, military and first responders.

At the same time, and especially since ChatGPT's public release in November 2022, generative AI (GAI) and LLMs skyrocketed in prominence and investments. Having the potential to disrupt all aspects of our digital lives [3], GAI sparked a race among big IT companies and smaller start-ups as they seek to exploit this emerging technology for new innovations. As regards the field of serious games, GAI holds significant potential for advancing the modelling of player behaviours [4]. LLMs' capacity to generate coherent and meaningful responses can facilitate realistic, believable Non-Playing Characters (NPCs), which may lead to engaging and authentic learning experiences [5]. However, there is still minimal evidence of GAI applications in serious games.

The objective of this paper is to provide initial insights into how GAI can enhance NPC realistic

responses to elevate user experience in a VR game setting. Specifically, we explore the utilisation of OpenAI's GPT-4 model (the most recent update to ChatGPT) for managing the dialogue in LAW-GAME's police interrogation game.

2 The LAW-GAME project

LAW-GAME (https://lawgame-project.eu/) aims to train LEAs in theory and real-life practice through gamification technologies in a safe and controlled virtual environment. To achieve this, the project develops a fully configurable, user-centred, social VR game platform embedding four highly immersive and attractive "gaming modes" as follows:

- *The CSI game*. Trainees examine the crime scene, identify and analyse all kinds of evidence (e.g., bullet holes, fingerprints) in order to solve the case.
- *The Police Interview game*. Trainees practise their interrogation and negotiation skills through conversational interaction with an artificial agent (AI-assisted 3D avatar) assuming the role of the suspect / perpetrator.
- *The Terrorist Attack game*. Trainees practise in managing extreme, time-critical situations in a more effective and efficient manner, reducing, or preventing terrorist threats.
- *The Car Accident game*. Trainees virtually collect scene's data (e.g., tire marks, damage to fixed objects) in order to decipher what happened before, during and after the collision.

3 The Interrogation game

Focusing on the Police Interview: Interrogation game, the game's basic elements are as follows:

- The game takes place inside a police interrogation room.
- There are two main characters: the police officer who interrogates the suspect (i.e., the player); and the suspect (i.e., a NPC controlled by the game's AI engine).
- The game has a specific storyline, i.e., a description of the incident and the reasons for the interrogation, and a set of evidence, gathered by the police prior to the interrogation.
- The player may freely ask any question to the suspect, i.e., the game provides no guidance as to what or when to ask, leaning more towards a realistic setting for testing the skills of a trainee rather than a learning tool.

4 Challenges in dialogue interaction

As discussed in a previous article [6], there are two main challenges in implementing a realistic, interactive dialogue between the LEA officer (player) and the suspect (AI-assisted NPC) in the Interrogation game.

First, the game engine that controls the suspect-NPC responses should be able to understand the officer's questions and generate a relevant reply, i.e., according to the provided scenario. Furthermore, the game engine should also provide a suitable response to questions that are marginally relevant or even completely irrelevant to the storyline.

Second, the game engine should be able to differentiate suspect's responses based on the context of the dialogue, i.e., to be able to decide whether it should respond with a lie or the truth, and to be able

to decide to which direction the storyline should unfold based on the player's previous questions. For example, the suspect should be able to respond differently to the same question being asked following another relevant question; or to the same question being asked after having admitted the truth of an evidence or having confessed the crime.

5 Dialogue management utilising GPT-4

GPT-4 is phenomenally capable of addressing the first challenge. When provided with the storyline, the evidence and relevant instructions, it can easily get into the role of the suspect's character. It can generate responses that are relevant, realistic and within the storyline. Moreover, it can creatively respond to relevant questions that were not foreseen in the storyline (e.g., invent an occupation for the suspect when none is provided). Finally, it can address questions that are completely irrelevant to the storyline while staying in character, e.g. pointing out that this is irrelevant to the interrogation case.

However, GPT-4 does not address the second challenge with similar success. Although it is capable of replying according to context, it cannot trustfully follow instructions on what it should reveal or not and when it should or should not lie, and it cannot trustfully follow the provided reasoning for managing the unfolding of alternate storylines. In fact, it tends to admit the truth much easier than it should for the suspect's role, and it tends to be "too talkative", giving away more information than asked.

This result is probably expected due to the generative nature and purpose of the GPT-4 model. Nonetheless, the fact remains that it cannot be trusted to manage a human-NPC dialogue and to control the storyline in the way we need it to. More so, since our application is a serious game for an actual training situation that targets player assessment and consistency / comparability of evaluations.

Thus, it becomes evident that a hybrid approach is needed. The game engine needs to retain control of the dialogue interaction and guide the unfolding of the storyline, while utilising GPT-4 for the aspects that it does well. For this hybrid approach, we differentiate between two types of questions:

I. questions that should get "one" answer (these answers do not drive the storytelling forward); and

II. questions that should get alternate answers depending on the context (these act as decision points for the unfolding of the storytelling).

All the information needed to reply to type I questions is provided to GPT-4 to generate relevant answers. However, for the type II questions, GPT-4 is instructed to pass them on to be handled by our conversational engine. In our engine we utilise a more traditional, yet controllable approach inspired by work in automated planning [7]. Specifically, we employ an action-driven framework in which a particular interaction (i.e., a question-and-answer pair) can activate a series of pre-conditions and post-conditions that impact the subsequent interactions. Thus, by monitoring the current state of the dialogue, our engine can decide on how to unfold the storytelling and can provide the most relevant answer to type II questions each time.

Initial tests of the hybrid approach are promising, since it provides a good control of the gameplay (through the handling of the type II questions) and, at the same time, a realistic experience for the user (through the utilisation of generative AI). Our next steps are to test the solution with the end users and

refine our approach.

6 Acknowledgement

Part of this work has been carried out in the scope of the LAW-GAME Project, which has received funding from the European Commission under the H2020 programme (Grant Agreement No. 101021714). The authors acknowledge support and contributions from all partners of the project.

7 References

[1] Hamari J, Koivisto J, Sarsa H. (2014). Does Gamification Work? – A Literature Review of Empirical Studies on gamification. In 47th Hawaii International Conference on System Sciences, 2014; pp. 3025-3034

[2] Mystakidis S. (2022). Metaverse. Encyclopedia, 2(1), pp. 486-497

[3] Mondal S, Das S, Vrana V.G. (2023) How to Bell the Cat? A Theoretical Review of Generative Artificial Intelligence towards Digital Disruption in All Walks of Life. Technologies 2023; 11(2), 44

[4] Pérez J, Castro M, López G. (2023). Serious Games and AI: Challenges and Opportunities for Computational Social Science. IEEE Access, vol. 11, pp. 62051-62061

[5] Westera W, Prada R, Mascarenhas S, et al. (2020) Artificial intelligence moving serious gaming: Presenting reusable game AI components. Educ Inf Technol 25, pp. 351–380

[6] Panopoulou E, Vassos S. (2023) AI-assisted Serious Games: Interrogating an Avatar in Virtual Reality, EDGE2022, under publication

[7] Petrick R, Bacchus F. (2004). Extending the Knowledge-Based Approach to Planning with Incomplete Information and Sensing. In Proceedings of the Conference on Principles of Knowledge Representation and Reasoning, 2004; pp. 613-622

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 111 – 117

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Transforming the Path Towards Automation of Monitoring and Management for Edge Computing

Georgios Samaras[1], Marinela Mertiri[1], Maria-Evgenia Xezonaki[1], Nikolaos Psaromanolakis[1], Vasileios Theodorou[1], and Theodoros Bozios[1]

Intracom Telecom, Greece
{gsamaras, marmert, maxez, nikpsarom, theovas, tmpo}@intracom-telecom.com

Abstract

The breakthrough of Artificial Intelligence (AI) has revolutionized Machine Learning (ML), particularly in the form of Transformer models such as Chat Generative Pre-training Transformer (ChatGPT), achieving state-of-the-art (SoTA) performance in various do mains, including Edge Computing. This paper introduces the TANDEM Smart Resource Monitoring and Management approach, which addresses the challenges of resource monitoring and management at the edge. TANDEM leverages AI/ML mechanisms to enable efficient service distribution, user-centered operation, and diverse edge applications support. It proposes novel forms of dynamic resource monitoring and management, utilizing Trans former models for accurate system usage prediction. Furthermore, TANDEM provides an AutoML framework and its AutoTinyML extension, which enhances IoT applications with powerful ML services. The proposed architecture and approach contribute to the intersection of AutoML and Transformer models with Edge Computing. Extensive experimental evaluations and analysis demonstrate the effectiveness and potential of our approach.

1 Introduction

The breakthrough of AI has sparked a surge in ML and Deep Learning (DL) applications and services, driving remarkable growth. Transformer ML models have gained immense popularity, mainly because of the Chat Generative Pre-training Transformer (ChatGPT), attracting significant attention to Automated ML (AutoML). These advanced models achieve SoTA performance across various domains, e.g. edge computing [4]. The widespread adoption of AI/ML will greatly benefit the Edge Computing domain.

Edge network infrastructure capacity poses limitations compared to centralized cloud re sources, requiring resource monitoring and management mechanisms for efficient Life Cycle Management (LCM) and service distribution. Recent advances in processing and AI/ML capabilities of smart devices offer exploitation opportunities, but challenges remain due to varying technology support and ongoing standardization efforts (e.g. ETSI Multi-access Edge Computing (MEC)). TANDEM addresses these

challenges with a novel edge platform architecture, user-centered operations, and support mechanisms. It enables various users to develop, exe cute, and manage edge services without infrastructure complexities, leveraging reusable services. TANDEM explores innovative aspects of edge computing, including service and application com position across edge nodes, incorporating AI/ML mechanisms and online monitoring functions. It automates service chain execution, adheres to standards, and encompasses business aspects such as pricing, Service Level Agreements (SLAs), and service/product life cycles.

Developing real-time prediction models for resource monitoring & management in edge network infrastructures is a complex challenge. Existing systems rely on simplistic heuristics, limiting their effectiveness in diverse applications and environments [6]. However, recent advancements in AI/ML offer the potential for improved models that outperform generic heuristics. By leveraging data-driven techniques and extracting relevant input data, customized ML models can be created. This approach addresses the need for efficient resource allocation and improved performance in modern applications while capitalizing on the growing volume of IoT generated Big Data. The proposed transformer models, incorporating a monitoring module for enhanced extraction of time dependencies, provide a solution for accurate system usage prediction.

This paper proposes novel forms of dynamic resource monitoring & management for edge network infrastructures to enable the automated selection of distributed and diverse resources. Additionally, we propose an AutoML framework and its AutoTinyML extension to enable automated online monitoring at the edge. The intention of this work is to provide an in-depth analysis and extensive experimentation on the intersection of AutoML and Transformer DL models with Edge AI- an area not yet adequately explored. In this sense, the contribution of this work is two-fold:

- A novel architectural element is presented, named Smart Resource Monitoring & Management, which is equipped with intelligent and data-driven prediction capabilities, able to provide adequate resource monitoring and management, to enhance coordination and delivery across stakeholders on the edge continuum;
- Transformer ML models are designed to draw patterns across resource types and determine suitable edge network infrastructure resources;

The rest of the article is organized as follows. Section 2 presents the related work. Section 3 presents the TANDEM and Smart Resource Monitoring & Management architecture. Section 4 discusses the evaluation results, while concluding remarks are presented in Section 5.

2 Related work

In the context of cloud, fog, and IoT environments, Edge Intelligence (EI), or Edge AI, moves AI frontiers from the cloud to the network edge to fully unlock the potential of big data at its production source and facilitate efficient management in various domains. [2] offers insights into the challenges in resource management for edge computing environments and lists several re source management techniques, covering aspects such as resource allocation, task offloading and load balancing. [3] presents a taxonomy of different ML-based resource management techniques in edge computing and analyzes their strengths, limitations, and potential applications. [9] introduces a dynamic resource management approach that allocates resources across cloud and edge environments, ensuring

efficient resources utilization and meeting SLAs.

Growing research interest has been recently devoted to the use of ML transformer models at the edge, although research work is still at an early stage. [1] studies various deployment methods for transformer models at the edge and explores relevant tooling options and practical approaches to illustrate the utilization of these models effectively. [7] examines the use of transformer models in resource-constrained edge environments with a specific focus on ultra-low power devices. [5] introduces Transformers on the Edge (Edgeformers) and demonstrates their effectiveness in capturing contextual information in edge computing scenarios, showcasing their potential for enhancing representation learning tasks in distributed edge environments.

Based on the above, in this work, a mechanism building on transformer models is proposed to facilitate the management and pattern identification of edge computing infrastructure resources. For the monitoring realization, a proposed AutoML pipeline and its AutoTinyML extension are used, showing the improved capabilities of transformer models in edge computing environments.

3 Architecture

The TANDEM platform offers support for Platform-as-a-Service (PaaS) edge and serverless computing models, enabling users to create complex service/function chains by combining existing services with custom ones. It focuses on edge computing environments like smart domains, connected vehicles and Industry 4.0, addressing challenges such as multi-protocol and multi data type support, device interconnection and management. The platform employs reusable and custom services, utilizing edge resources, connected devices, and auto-scaling/auto-healing mechanisms to facilitate users with easy applications development through its API.

The TANDEM architecture, shown in Figure 1, incorporates elements from the European Telecommunications Standards Institute (ETSI) Network Functions Virtualisation (NFV) reference architecture and the ETSI MEC architecture while introducing modules to support various usage scenarios. Its key feature is the ability to create more complex services and applications by integrating services and devices from different edge hosts or the cloud. The architecture is based on microservices, with edge nodes organized in edge clouds forming clusters. The architecture consists of three levels: the System Level, the Edge Computing Level and the Device Level. The System Level coordinates resources, devices, and services across all edge nodes via the Edge Orchestrator, while also providing user management, pricing policies, and billing functions via the Service Orchestrator. The Edge Level encompasses the TANDEM Edge Platform, IoT support services and customer-installed services. Finally, the Device Level includes all devices connected to TANDEM, which can send data and receive commands or microservices.

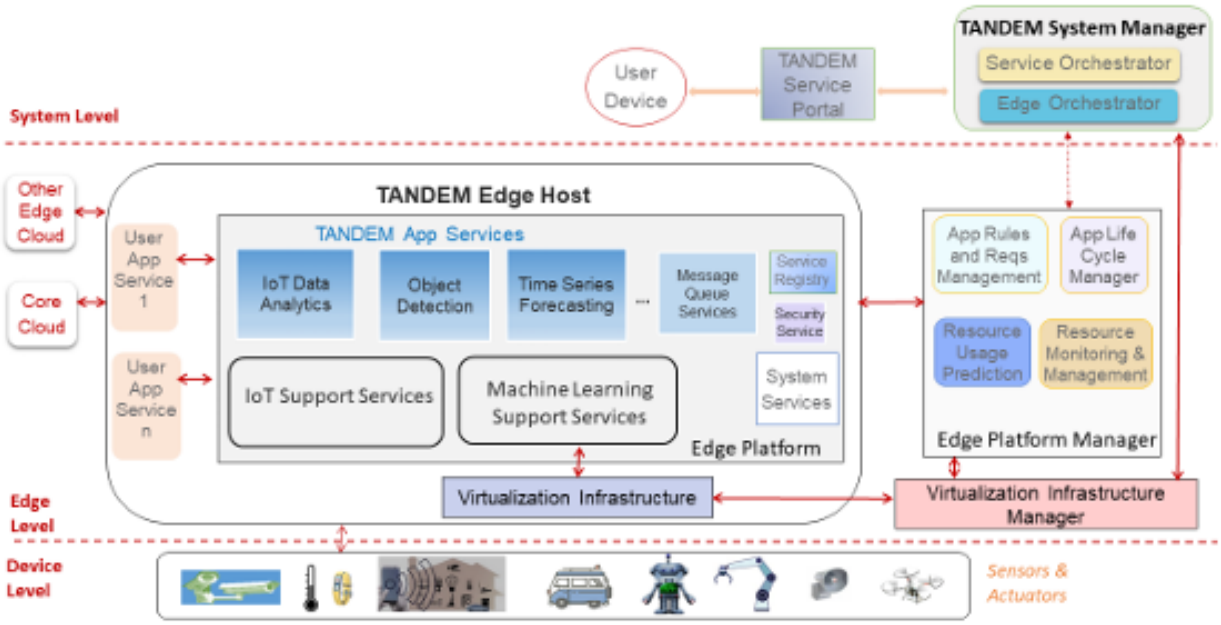


Figure 1: The architectural components and core technological stack of TANDEM

3.1. Smart Resource Monitoring & Management

In this work, we introduce a new subsystem into this architecture, named Smart Resource Monitoring & Management (SRMM), which performs zero-touch monitoring, management, and life cycle functions for one edge node or a cluster of edge hosts (edge cloud). It utilizes powerful AutoML mechanisms for accurate resource usage prediction, enabling efficient resource management and Quality of Service (QoS)

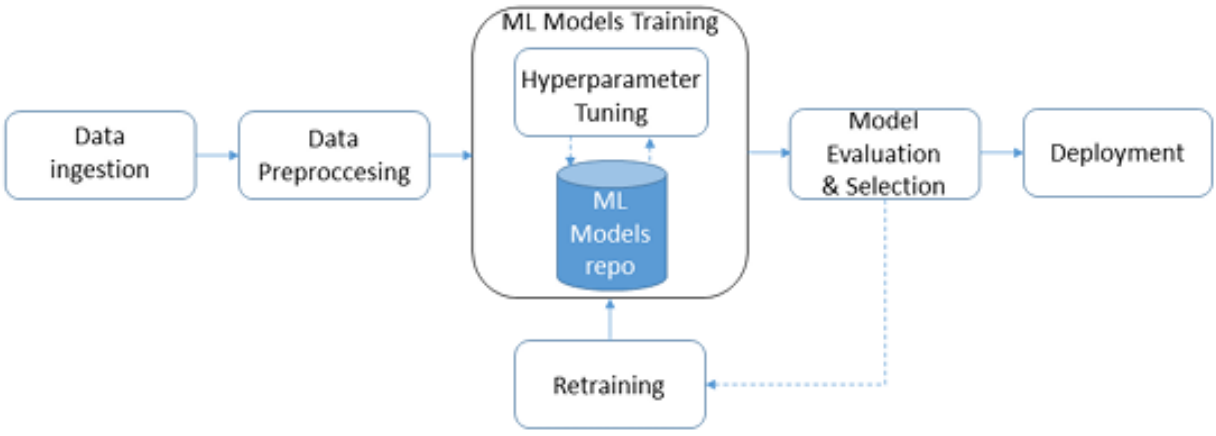


Figure 2: Smart Resource Monitoring & Management AutoML pipeline.

monitoring. Precise predictions are essential for optimal decision-making, which considers both existing and new services for intelligent resource allocation based on predicted resource usage, resource characteristics and user service requirements.

SRMM's AutoML pipeline ingests data and deploys a high-performing ML model for Time Series Forecasting to predict resource usage. The pipeline stages are data pre-processing, i.e. cleaning & transformations, models training in parallel and hyperparameter tuning, model evaluation & selection and an optional retraining stage in case no model passes the evaluation phase, as shown in Figure 2. Moreover, AutoTinyML extends the SRMM's AutoML pipeline with model compression capabilities (e.g. post-training model pruning and quantization). The extension follows the same philosophy with the AutoML pipeline, but comprises an evaluation stage that checks for both model accuracy and size- particularly useful for Edge AI.

Our ML model suite contains novel transformers (Temporal Fusion Transformer (TFT) & Informer), MLP-based (Neural basis expansion analysis for interpretable time series forecasting (N-BEATS) & Neural Hierarchical Interpolation for Time Series Forecasting (N-HiTS)), popular RNN-based models (Gated Recurrent Unit (GRU) & Long Short-Term Memory (LSTM)) and a lightweight regression technique; Support Vector Regression (SVR).

We incorporate MLOps methodologies for the LCM of ML models. Over time, ML model accuracy can degrade due to data drift [8]. Periodic evaluation compares the accuracy to predefined empirical thresholds, triggering SRMM's AutoML pipeline with the newly- unseen by the ML models during the training phase- stored data as input.

The proactive auto-scaling/auto-healing mechanisms of TANDEM rely on SRMM's monitored metrics, where an empirical threshold is defined for each metric. In case SRMM's prediction exceed the set threshold, its management service is notified and automatically performs the required service adaptation. SRMM decides what the adjustment will be based on the metric that is predicted to exceed the threshold and some predefined rules, e.g. a configurable scaling factor. For example, if the metric is service related, then SRMM may decide to scale-out the service itself by deploying a new instance.

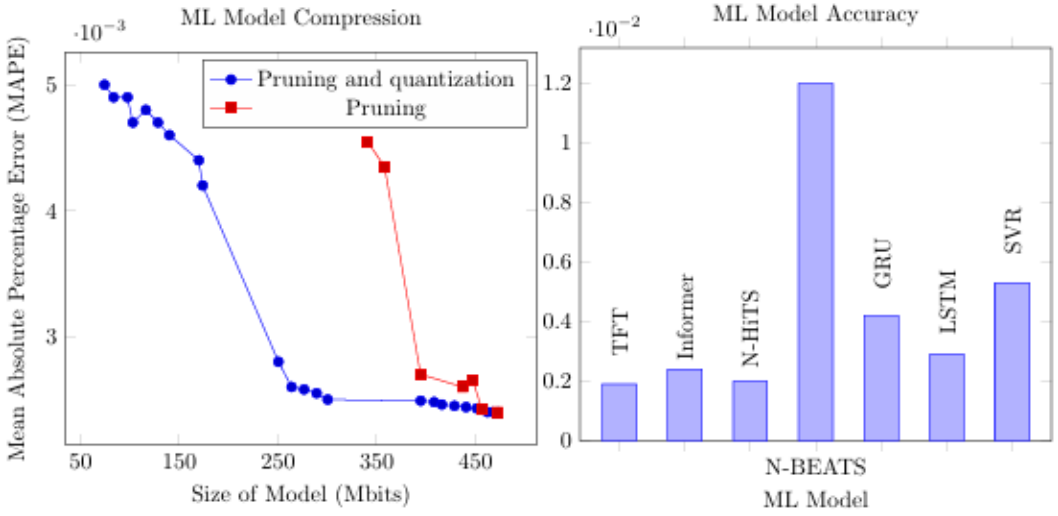


Figure 3: Left: Model accuracy after size pruning. Right: Resource usage prediction accuracy.

4 Performance Evaluation

This section details the thorough evaluation of the SRMM subsystem, focusing on the ML models accuracy performance and the effectiveness of its AutoTinyML extension.

4.1. Resource Usage Prediction Accuracy

Extensive experimentation shows the ML models', presented in section 3.1, robust accuracy. We collected data of four resource usage metrics, namely Memory Consumption, CPU Con sumption, Received Throughput and Transmitted Throughput. The resulted data set contains 8 days of 1 minute-based measurements.

We downsampled the data set and then split it in train, validation and test data sets. The train set and validation set are used internally by the SRMM's AutoML pipeline to automatically train and fine-tune (via 5-fold cross-validation) the ML models. Notice that for evaluation purposes we configured the pipeline to output all the trained models, and not just the best performing model; in other words we disabled the model selection phase. The test set is used to perform time series forecasting. Our results are shown in Figure 3, where it is evident that the transformer models (TFT and Informer) are performing better than RNN-based models (GRU and LSTM), and that they are comparable to the N-HiTS MLP-based model. SVR achieves lower accuracy than the aforementioned models, but outperforms N-BEATS, the other MLP-based model, which has the lowest accuracy in this experimentation setting.

4.2. AutoTinyML Model Compression

The AutoTinyML extension of the SRMM's AutoML pipeline takes into account size and accuracy of the ML model. Resource-constrained environments advocate for smaller models and can tolerate some minimal loss of accuracy. ML model pruning reduces the size and complexity of an ML model by removing unnecessary parameters, connections, or structures, thereby improving efficiency and reducing resource requirements without significant loss in performance. ML model quantization involves the process of reducing the precision of numerical values, typically from floating-point to fixed-point representation, in order to reduce memory footprint, improve computational efficiency, and enable deployment on resource-constrained devices. Our AutoTinyML extension prunes a configurable amount of trainable layer weights and quantizes the ML model post-training, by reducing data type precision. In this evaluation setting, the data set described in 4.1 was used. We tuned the data type precision from float32 to int8, including the corresponding unsigned and quantized types, in combination with applying polynomial decay for tensor sparsity from 50% up to 80%, i.e. from 50% up to 80% zeros in weights. Figure 3 shows the trade-off between size and accuracy, and depicts that model compression managed to reduce size by 6 times while the accuracy difference is significantly lower.

5 Conclusion

Enabling automated monitoring and management of edge computing resources via AI/ML is a growing demand to realize the envisioned capabilities of edge continuum. Towards such a goal, this paper provides insights into the design and implementation of the Smart Resource Monitoring and Management subsystem for the TANDEM approach, which offers a comprehensive solution for

addressing resource monitoring and management challenges in edge computing. By integrating AI/ML mechanisms, TANDEM optimizes resource allocation, improves service distribution, and supports a wide range of edge applications. Experimental evaluations confirm its effectiveness and potential for practical deployment.

6 Acknowledgement

This work is supported by the TANDEM project, co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH CREATE- INNOVATE (project code: T2ΕΛΚ-02825).

7 References

[1] Cassie Breviu. How to operationalize transformer models on the edge. QCon Plus ’22, 2022.

[2] Cheol-Ho Hong and Blesson Varghese. Resource management in fog/edge computing: A survey on architectures, infrastructure, and algorithms. ACM Computing Surveys, pages 1–37, 2019.

[3] Sundas Iftikhar et al. AI-based fog and edge computing: A systematic review, taxonomy and future directions. Internet of Things, 2023.

[4] Juan J. et al. Performance evaluation of state-of-the-art edge computing devices for DNN inference. IECON ’20, 2020.

[5] Bowen Jin, Yu Zhang, Yu Meng, and Jiawei Han. Edgeformers: Graph-empowered transformers for representation learning on textual-edge networks, 2023.

[6] Anshul Makkar. Scope and performance of credit-2 scheduler. Xen-Summit, 06 2016.

[7] Francesco Bianco Morghet, Daniele Jahier Pagliari, Alessio Burrello, et al. Application of trans formers to edge-computing in ultra-low power devices, 2021.

[8] Georgios Samaras et al. Qmp: A cloud-native mlops automation platform for zero-touch service assurance in 5g systems. In MeditCom, 2022.

[9] Shashank Shekhar et al. Dynamic resource management across cloud-edge resources for performance-sensitive applications. CCGRID ’17, 2017.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 118 – 124

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Intelligence Functions Placement in B5G / 6G wireless networks

Vasiliki Lamprousi[1] Sokratis Barmpounakis[1] Vera Stavroulaki[1] and Panagiotis Demestichas[1]

[1]*WINGS ICT Solutions, Greece*
vlabrousi@wings-ict-solutions.eu, sbarmpounakis@wings-ict-solutions.eu,
veras@wings-ict-solutions.eu, pdemest@wings-ict-solutions.eu

Abstract

The optimal computation, resource, and storage positioning for succeeding the best performance of complex systems, populated by multiple mobile devices with respect to mobile users and data protection, is of great interest in "Beyond the 5th Generation" (B5G) / 6th Generation (6G) networks in academia as well as in industry sectors. In this paper, an intelligence functions placement algorithm is proposed for optimally allocating the functionality to the various network/compute nodes as part of the intelligence distribution decision-making functional component of the smart connectivity platform envisioned in the H2020 project DEDICAT 6G. This algorithm can overcome a possible increase of network latency or a possible unavailability of used edge node and can be applicable to many use cases including the ones having robot edge nodes (warehouse environment). The theoretical estimations show that the proposed algorithm can significantly reduce the power consumption compared to baseline placement algorithms.

1 Introduction

The revolution of Industry 4.0 changed the way companies produce, enhance, and deliver their products by combining physical assets and advanced technologies such as Artificial Intelligence (AI), Internet of Things (IoT), robots, 3D printing, cloud computing, etc. The upcoming revolution of Industry 5.0 aims to achieve better industrial production by enabling effective collaboration and interaction among machines and humans (Adel, 2022). To support these technologies and applications, B5G/6G networks will offer ultra-fast, highly adaptive, and dependable platforms. These networks will also enable sophisticated management algorithms that will optimize various performance metrics, such as latency, energy efficiency, and resource allocation to improve the efficiency and resilience of advanced, industrial processes.

The H2020 project DEDICAT 6G (H2020-ICT-52, 2023) has the overall objective to transform B5G networks into a smart connectivity platform that is highly adaptive, ultra-fast, and dependable/resilient for supporting securely innovative, human-centric applications. Within this project a novel intelligence functions placement algorithm was developed for optimally allocating the functionality to the various network/compute nodes including robotic units (DEDICAT6G, 2021) and is presented in this paper.

The contribution of this paper is as follows: (i) This paper studies the intelligence functions placement problem in a smart warehousing environment where there are multiple robotic units, edge/cloud servers, towards energy efficiency and minimum communication cost (related to transmission delay) among others. (ii) A novel metaheuristic algorithm is proposed for addressing this problem scalable also in large scale experimentations. (iii) The preliminary results show that the proposed algorithm can obtain an important reduction of the power consumption compared to baseline placement algorithms.

The rest of the paper is structured as follows. Section 2 presents the related work according to the state of the art in the research field. Section 3 provides the description of the intelligence functions placement problem as well as the proposed solution. Section 4 shows the results obtained till now and Section 5 concludes.

2 Related work

The problem of the optimal placement of services, tasks, computation, to the compute nodes of the system of interest has been widely researched in B5G/6G networks. In the literature these problems are solved with various optimization, metaheuristic, and Machine Learning (ML) algorithms. The authors of (L. Yala, 2018) use a genetic metaheuristic algorithm to solve an optimization problem that aims to place Virtual Network Functions (VNFs) for ultra-reliable and Low Latency Communications (uRLLC) services in a way that minimizes latency and maximizes service availability. Their algorithm can find near-optimal solutions faster than an exact algorithm of a Mixed Integer Programming (MIP) solver. The authors of (J. Kim, 2020) present a deep Q-network-based algorithm for placing Cloud-native networking functions (CNFs) on edge clouds in an efficient way, considering the cost of launching and operating CNFs, the back-haul control traffic overhead, and the number of served requests at each time. The simulation results showed that their algorithm can adapt to changes in service demand and reduce the cost per hour.

Moreover, energy efficiency is a key objective for 6G networks, hence there are many placement optimization algorithms proposed to achieve it. The (J. Li, 2023) considers the deployment of AI tasks on edge computing nodes that can access a cloud server, and jointly optimize the resource allocation, offloading decision, computing time, energy consumption and inference accuracy of each node, with the development of an algorithm which breaks down the problem into simpler subproblems based on the alternating direction multiplier method. Similarly, (A. Mesodiakaki, 2022) addresses the joint problem of user association, traffic routing and VNFs placement to maximize the energy efficiency and user acceptance ratio of the mobile network by introducing an energy-efficient real-time heuristic called ONE that uses online convex optimization techniques to solve the problem in a distributed manner. Furthermore, (Taneja A, 2022) proposes an energy-efficient algorithm for virtual machine placement optimization in cloud data centers that reduces the total power consumption and resource wastage by employing a heuristic resource usage factor model with reward and penalty mechanisms. However, there are limited studies that consider additionally the challenges of robot-based compute node and robot related functionality which our study includes.

3 Problem description / Solution approach

A set of n Functional Entities (FEs) e.g., tasks, jobs, services, is assumed for approaching the intelligence functions placement problem, with specific computation and functional requirements.

The possible communications/dependencies between FEs are represented by a functional graph (Directed Acyclic Graph-DAG), where each node corresponds to a FE and each edge connects interacting FEs and it is weighted according to the amount of data transferred between FEs. Additionally, each edge has a maximum acceptable transmission delay (threshold).

Also, a set of m Hosting Entities (HEs) e.g., edge nodes, core nodes, robotic units, end user devices, is assumed, having some capabilities. These are the maximum available CPU, and memory resources, the battery level if applicable and the set of FEs that each HE can support (e.g., a robotic unit can support object recognition if camera is available but cannot support grasping an object if robotic arm is not available). Also, a system layout graph is considered, consisting of the available HEs and the communicational channels among them with varied capacity link each.

The aim is to allocate the FEs to HEs by ensuring efficiency of the system with low energy consumption and latency. Therefore, the objective is to find the optimal cost allocation that meets the performance requirements. The Objective Function (OF) which is minimized $OF = w_1 B + w_2 P + w_3 D$, consists of the term $B = \sum_{j=1}^{m} y_j b_j$ which denotes the cost related to the battery level of utilized HEs (takes higher values when battery is low, and close to zero values when it is fully charged or when the HE is not battery-powered). y_j is a binary decision variable that indicates if HE j is utilized or not and b_j is a cost related to the battery level of HE j . The $P = \sum_{j=1}^{m} \left(W_{max}^{j} - W_{idle}^{j}\right). Ucpu_j(x_{i,j}) + W_{idle}^{j}$ term denotes the power consumption cost which is modelled based on (Alharbe, Aljohani, & Rakrouki, 2022), where $Ucpu_j(x_{i,j})$ is the CPU utilization rate on the HE j in terms of the decision variable $x_{i,j}$ which indicates if FE i is assigned on HE j. $W_{max}^{j} - W_{idle}^{j}$ are the power consumption when the HE j is fully loaded and idle, respectively. Finally, $D = \sum_{j=1}^{m} \sum_{j\prime=1}^{m} \sum_{i=1}^{n} \sum_{i\prime=1, i\prime\neq i}^{n} z_{i,i\prime,j,j\prime} \frac{k_{i,i\prime}}{cap_{j,j\prime}}$ denotes the cost (transmission delay) imposed by the communication among HEs related to the amount of data transferred ($k_{i,i\prime}$) between FEs and the maximum capacity link ($cap_{j,j\prime}$). The binary decision variable $z_{i,i\prime,j,j\prime}$ shows if HEs j and j' are communicating due to communicating FEs assigned on them. Each term of the OF is normalized and is weighted (w_1, w_2, w_3) depending on the use case.

The constraints of the problem are that each FE can be allocated to only one HE ($\sum_{j=1}^{m} x_{i,j} = 1, \forall\, i = \{1, \ldots n\}$). The available resources (e.g., CPU, memory) of the HEs should be respected ($\sum_{i=1}^{n} x_{i,j} * cpuFE_i \leq cpuFE_j$ and $\sum_{i=1}^{n} [x_{i,j} * memFE_i \leq memHE_j \; \forall\, j = \{1, \ldots m\}$). The functional requirements of the FEs (e.g., camera, wheels) should be respected ($x_{i,j} \leq t_{i,j}, \forall\, i = \{1, \ldots, n\}, j = \{1, \ldots m\}\, where\, t_{i,j} \leq$ is 1 if FE i can be assigned on HE j in terms of functionality types that can be supported by the HE and 0 otherwise). Also, the maximum transmission delay between two interacting FEs should be respected $z_{i,i\prime,j,j\prime} \cdot \frac{k_{i,i\prime}}{cap_{j,j\prime}} \leq l_{i,i\prime}, \forall\, i, i' \in \{1, \ldots, n\}, j, j' \in 1, \ldots, m, where\, i \neq i', j \neq j', where\, l_{i,i\prime}$ is the maximum acceptable transmission delay between FEs i and i').

An example of the intelligence functions placement algorithm utilization is shown in Figure 1. Two robots with different functionalities and one server are depicted having some FEs assigned on them. An increased network latency is detected, and the algorithm is triggered. The output of the algorithm is the reallocation of the FEs to the available HEs to overcome this issue.

The intelligence functions placement problem was solved with the development of a metaheuristic genetic algorithm based on the genetic algorithm paradigm. The algorithm initializes a population of

chromosomes, possible solutions, each consisting of a series of HEs representing the "proposed" HE for each FE (to be placed). Then, the fitness/objective function (OF) is applied to each chromosome obtaining a fitness score each. Over the course of a number of generations defined by the dynamic stopping criterion utilized (for succeeding convergence in a satisfying execution time), a population of chromosomes evolves, and operators like parent selection, crossover, and mutation improve the population's overall fitness. In this metaheuristic tournament selection, one-point crossover and reserve sequence mutation proposed in (Otman Abdoun, 2012) was utilized among others. Crossover and mutation operators occur with a predefined probability. The crossover and mutation rate (predefined occurrence probability) utilized was 0.8 and 0.15, respectively.

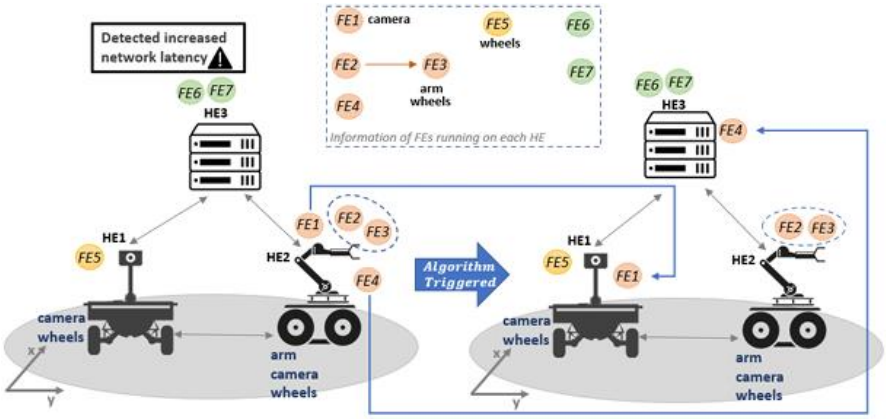


Figure 1: Schematical representation of a case of intelligence functions placement algorithm utilization.

4 Results

A performance testing of the proposed model was performed comparing to a MIP Python solver output

	score		execution time (s)	
# FEs	MIP solver	Proposed algorithm	MIP solver	Proposed algorithm
2	0.1638	0.1638	0.011	0.551
4	0.1697	0.1718	0.030	0.684
8	0.1770	0.1939	1.316	1.371
12	0.1821	0.1822	200	1.494
16	0.1972	0.2081	200	4.675
20	0.2040	0.2278	200	6.128
24	0.2264	0.2457	200	14.675
28	0.2553	0.2935	200	16.624
32	0.2769	0.3101	200	22.307
36	0.3146	0.3317	200	23.730
40	-	0.3517	-	32.850
44	-	0.3787	-	49.427

Table 1: Performance testing (score and execution time) of the proposed intelligence functions placement algorithm compared with the output of the MIP solver with fixed HE-schema and increasing number of FEs.

(Mitchell S, 2011) by measuring the scores (the final minimum obtained value of the objective function OF) and the execution time having a fixed HE-schema of 43 HEs and increasing number of FEs (see Table 1). The levels of the available CPU of the HEs are assumed {2000, 2600, 3000} MIPs, the levels of the available memory are {2048, 4096, 8192} MB, the levels of power consumption when fully loaded and when idle are {260, 360, 460} W and {70, 100, 170} W respectively. The links between HEs have capacity 3.3 – 20 Mbps. The FEs have {500, 750, 1500} MIPS levels of required CPU, {256, 512, 2048} MB levels of required memory, 2-10 MB levels of data transferred and maximum available delay between interacting FEs, 0.2-1.4 s. An execution time limit of 200s was imposed on the MIP Python solver because it was computationally intractable from over 12 FEs. The results showed that the proposed metaheuristic algorithm had close to optimum scores within significantly less time than MIP model, and when exceeding 36 FEs, MIP solver could not find feasible solution within 200s.

Moreover, Figure 2 shows the theoretical measurements of the percentage reduction of power consumption when using the proposed metaheuristic intelligence functions placement algorithm and when using the random feasible placement, which respects the systems constraints, and the round robin placement as baselines. For these measurements the same fixed HE-schema was assumed of 43 HEs as the one assumed for the performance testing. The mean value of ten measurements was obtained in each case. The power consumption was calculated by the P term of the OF described in Section 3. As it is observed in Figure 2, as the number of FEs increases, the power consumption gains increase, until a critical point (25 FEs) and then decreases. The proposed algorithm provides higher power consumption gains when there is sufficient availability of computational resources on the HEs, thus solution space is wider and can easier reach an optimum state. From these estimations is shown that our algorithm can reach up to a 35%-44% reduction of power consumption compared to random feasible placement or round-robin placement, respectively.

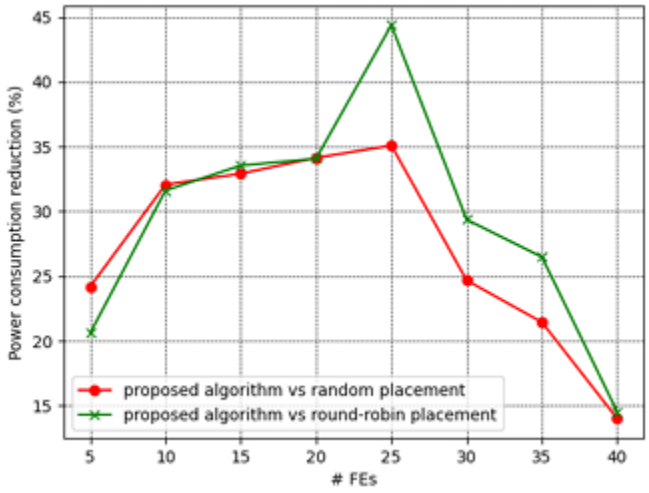


Figure 2. Percentage reduction of power consumption when the intelligence functions placement algorithm is used compared to a random feasible placement and round-robin placement (baselines).

The above scenarios and results were reactive approaches of functions placement. A proactive approach which could predict the future states of the FEs and resources, could identify the possible increase of network latency or possible unavailability of used HE. The prediction of the behavior of

services or components can increase efficiency and reduce the operational and maintenance costs of the system by scheduling any needed placement in advance. The discussed prediction of upcoming critical situations can be succeeded by monitoring data from selected services or components, and train AI/ML models for time series forecasting. The output future state of the system can be used by the developed intelligence functions placement mechanism to determine whether pre-emptive actions are necessary to prevent upcoming critical events. This work is ongoing and will soon generate important results.

5 Conclusion

In this paper, the intelligence functions placement problem for distributing the intelligence/functionality to the various network nodes (edge, core nodes, robotic units etc.) is studied, taking into consideration the transmission delay and power consumption among others. The description of the problem is provided along with the description of the MIP and genetic algorithm implemented for approaching this problem. Additionally, some main results are provided with a discussion on the proactive and dynamical placement of the intelligence functions.

6 Acknowledgement

This work was supported by the European Union H2020 Project DEDICAT 6G under grant no. 101016499. The contents of this publication are the sole responsibility of the authors and do not in any way reflect the views of the EU.

7 References

[1] Abdoun O, Jaafar A, Chakir T (2012). Analyzing the performance of mutation operators to solve the travelling salesman problem. Neural Evolut Comput Int J Emerg Sci 2(1):61–77

[2] Adel A (2022). Future of industry 5.0 in society: human-centric solutions, challenges and prospective research areas. J Cloud Comput; 11:40. https://doi.org/10.1186/s13677-022-00314-5.

[3] L. Yala, P. A. Frangoudis, and A. Ksentini (Dec 2018). Latency and Availability Driven VNF Placement in a MEC-NFV Environment, in 2018 IEEE Global Communications Conference (GLOBECOM), pp. 1–7

[4] J. Kim, J. Lee, T. Kim, and S. Pack (2020). Deep Reinforcement Learning based Cloud-native Network Function Placement in Private 5G Networks, in 2020 IEEE Globecom Workshops (GC Wkshps), Taipei, Taiwan, pp. 1–6

[5] J. Li, F. Lin, L. Yang and D. Huang, (2023 May-June). AI Service Placement for Multi-Access Edge Intelligence Systems in 6G, in IEEE Transactions on Network Science and Engineering, vol. 10, no. 3, pp. 1405-1416, doi: 10.1109/TNSE.2022.3228815

[6] A. Mesodiakaki, M. Gatzianas, G. Kalfas, C. Vagionas, R. Maximidis and N. Pleros, (2022). ONE: Online Energy-efficient User Association, VNF Placement and Traffic Routing in 6G HetNets, 2022 IEEE Globecom Workshops (GC Wkshps), Rio de Janeiro, Brazil, pp. 304-309, doi: 10.1109/GCWkshps56602.2022.10008742

[7] Taneja A, Saluja N, Taneja N, Alqahtani A, Elmagzoub MA, Shaikh A, Koundal D. (2022). Power Optimization Model for Energy Sustainability in 6G Wireless Networks. Sustainability. 14(12):7310. https://doi.org/10.3390/su14127310

[8] Alharbe, N.; Aljohani, A.; Rakrouki, M.A. (2022). A Fuzzy Grouping Genetic Algorithm for Solving a Real-World Virtual Machine Placement Problem in a Healthcare-Cloud. Algorithms, 15, 128. https://doi.org/10.3390/a15040128

[9] Mitchell S, O'Sullivan M, Dunning (2011). Pulp: A linear programming toolkit for python. Accessed May 1, 2013, https://code.google.com/p/pulp-or/

[10] DEDICAT 6G D3.1 First Release of Mechanisms for Dynamic Distribution of Intelligence, Deliverable D3.1, 2021

[11] H2020-ICT-52 project DEDICAT 6G: Dynamic coverage Extension and Distributed Intelligence for human Centric Applications with assured security, privacy, and Trust: from 5G to 6G, Project website available at https://dedicat6g.eu/. Accessed July 11, 2023

EMERGING TECH CONFERENCE 2023
Edge Intelligence

Papers

Session 1.5 | Circuits and Systems

Session Chairs: **Vasilis PAVLIDIS**

Analysis of SQRL strengths and weaknesses compared to other authentication mechanisms
Dimitrios Simeonidis

Fault Detection in Analog Circuits by utilizing the Current Supply Transients
Vassilis Vassios, Argirios Hatzopoulos, Dimitrios Papakostas and Ioannis Intzes

Aging alleviation technique for 8T IMC SRAMs
Helen-Maria Dounavi and Yiorgos Tsiatouhas

Reduction of large-scale RLCk models via low-rank balanced truncation
Christos Giamouzis, Dimitrios Garyfallou, Anastasis Vagenas and Nestor Evmorfopoulos

MORCIC: Model Order Reduction Techniques for Electromagnetic Models of Integrated Circuits
Dimitrios Garyfallou, Athanasios Stefanou, Christos Giamouzis, Moschos Antoniadis, Georgios Chararas, Konstantinos Chatzis, Dimitris Samaras, Rafaela Themeli, Anastasios Michailidis, Vasiliki Gogolou, Nikos Zachos, Nestor Evmorfopoulos, Thomas Noulis, Vasilis F. Pavlidis, Alkiviadis Hatzopoulos, Elpida Chatzineofytou and Yiannis Moisiadis

Design and implementation of a compact RISC-V based Machine Learning accelerator on Low End FPGA
Manolis Galetakis, Stavros Kalapothas, Georgios Flamis, Paris Kitsos and Fotis Plessas

HETiA
EXPANDING THE DIGITAL FRONTIERS

ARISTOTLE UNIVERSITY OF THESSALONIKI

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 125 – 130

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Analysis of SQRL: A Comparative Study with Traditional Authentication Mechanisms

Dimitrios Simeonidis

Department of Informatics-University of Economics - Varna,9002 Varna, 77
Kniaz Boris I, Blvd. Bulgaria

Abstract

This paper presents a comprehensive analysis of the Secure Quick Reliable Login (SQRL) authentication system, comparing its strengths and weaknesses with traditional authentication methods. SQRL is a modern approach designed to address the inherent security and usability issues associated with traditional username/password-based authentication. We delve into SQRL's security features, usability, and its potential to replace conventional authentication methods. The analysis highlights the challenges and advantages SQRL brings to the authentication landscape, emphasizing its suitability for various scenarios.

1 Introduction

Authentication is a fundamental element of digital security, serving as the primary method for individuals to access online services. Historically, traditional username/password authentication has been the default approach for this purpose. However, this conventional method comes with inherent vulnerabilities that have increasingly made it a target for cyberattacks.

One of the most pressing issues with traditional authentication is the widespread problem of password reuse. Users tend to employ the same passwords across multiple services, creating a significant security risk. In the event of a data breach on one platform, malicious actors can potentially gain access to a user's accounts on various other services, compounding the damage.

Additionally, traditional password-based systems are plagued by weak password choices. Many users opt for easily guessable passwords or fail to update them regularly, leaving their accounts vulnerable to brute-force attacks. Furthermore, the rise of sophisticated phishing attacks has made it even more challenging to protect login credentials.

Secure Quick Reliable Login (SQRL) represents a promising alternative to traditional authentication methods. It aims to address the weaknesses of conventional systems while prioritizing both security and user-friendliness. SQRL offers several notable advantages:

Security: SQRL employs advanced cryptographic techniques, making it highly resistant to interception and unauthorized access. By eliminating the need for passwords, it mitigates the risks associated with password reuse and weak passwords.

Reliability: SQRL is designed to be highly reliable, with mechanisms in place for account

recovery and protection in case of device theft or loss.

Ease of Use: SQRL provides a user-friendly experience by enabling login through a simple QR code scan. This eliminates the need for users to remember and type complex passwords.

Integration: While SQRL represents a departure from traditional authentication, it has the potential for seamless integration into various online services, offering a more secure authentication method.

However, SQRL is not without its own set of challenges and limitations. It relies on additional applications for secure key storage and lacks inherent protection against DNS spoofing. Compatibility issues may arise on devices with limited resources, and its widespread adoption depends on service providers incorporating SQRL into their platforms.

In conclusion, SQRL presents a compelling alternative to traditional username/password authentication. By addressing the vulnerabilities of traditional methods while prioritizing security and user-friendliness, SQRL has the potential to enhance the overall landscape of digital authentication. However, its successful integration and adoption will require collaboration between technology providers, service operators, and end-users to ensure a more secure and reliable online experience.

2 SQRL vs. Passwords

Traditional passwords present numerous challenges, including the need for users to remember multiple complex passwords and the risk of password reuse. SQRL offers several advantages in this context:

Strengths

Security: SQRL adopts Trust No One (TNO) principles and proven cryptographic techniques, making it resilient against interception.

Phishing Prevention: SQRL's unique identity creation for each web domain makes it challenging for phishing attacks to succeed.

Usability: SQRL simplifies account creation, eliminating the need for email addresses and complex passwords.

3 SQRL vs. Password Managers

Password managers provide a centralized repository for passwords and offer some similarities to SQRL:

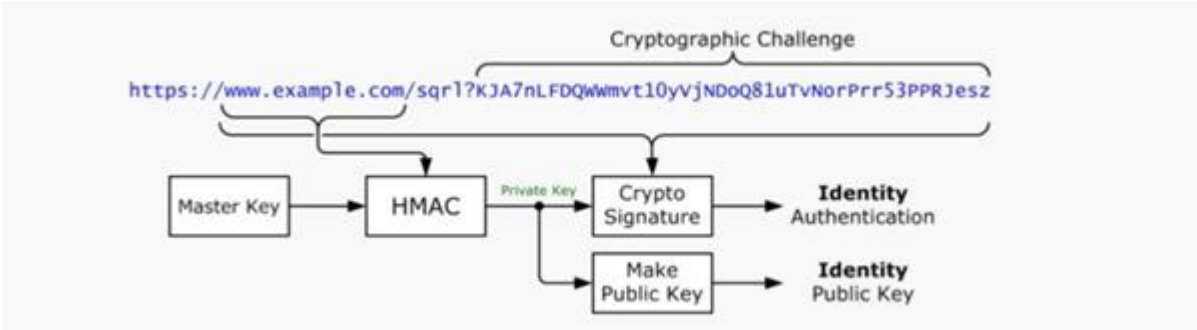
Strengths:

Security: Both SQRL and password managers prioritize security and encryption.

Reliability: Password managers offer reliable access to stored credentials. SQRL is exactly where every user needs it. If stopped working or damaged, rescue can be usedcode to restore the SQRL ID.

- ✓ in case of theft of the device, the SQRL ID of the user, it is protected cryptographically, making it almost impossible to used by third parties.

- ✓ It is possible to replace SQRL ID in order to the potential attacker is excluded from the various accounts of the user.
- ✓ There is the possibility to transfer SQRL ID when changing the user's device to ensure its continuity interoperability.



Weaknesses:

Usability on Untrusted Devices: Password managers may expose passwords when used on untrusted devices, unlike SQRL's QR code scanning.

Phishing Vulnerability: Like traditional passwords, password managers are susceptible to phishing attacks.

Dependency on Passwords: Password managers rely on passwords to function, hindering the goal of password-free authentication.

4 SQRL vs. Client Certificates

Client certificates offer an alternative to passwords but come with their own set of challenges:

Strengths:

Security: Client certificates can provide strong security when properly implemented.

Weaknesses:

Usability: Client certificates can be complex to set up and transfer between devices.

Dependency on Certification Authorities: Trust in third-party Certification Authorities (CAs) is required for client certificates.

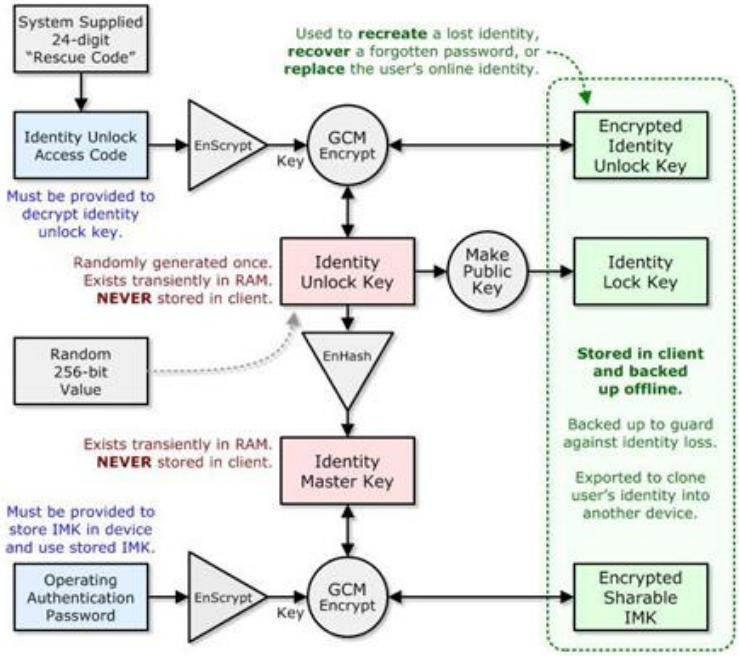
Phishing Vulnerability: Similar to passwords, client certificates can be vulnerable to phishing on untrusted devices.

No Email Addresses: SQRL's anonymity feature eliminates the need for email addresses, enhancing privacy.

Phishing Prevention: SQRL's identity creation based on domain names helps prevent phishing attacks. Although the SQRL identity connection system is not promoted as a solution against phishing but as it turns out, the control architecture SQRL identity presents significant opportunities to prevent such kind of attacks.

When using SQRL, users do not recognize and authenticate themselves with a username and password. On the other hand, their unique identity comes from the Secret Master Key and the site's full domain name. Given that SQRL creates a unique user identity for each web domain, the identity of the user for a phishing site such as ebbay.com or ebay.cn or anything other than the original site that the user he thinks visiting would be useless to any intruder.

This means that the SQRL connection link provided by a website that is malicious, must be correct and authentic. In our example it should be. "ebay.com" is why this string domain name is used in the creation of the identity by which ebay.com knows the user. This means that the SQRL application of the user will connect directly to the original website and not to the fake phishing website.



Effectiveness: SQRL's time-consuming account creation deters attackers.

Ease of Learning: SQRL simplifies the authentication process, enhancing usability.

Reliability and Recovery: SQRL offers reliable authentication and the ability to recover from identity theft.

Integration Potential: SQRL's potential to replace passwords entirely benefits websites and users.

5 Limitations and Challenges of SQRL Authentication

Secure Quick Reliable Login (SQRL) is a robust authentication method, but it has its share of limitations and challenges that must be considered for a complete evaluation.

Dependency on Additional Applications

One notable limitation of SQRL is its reliance on additional applications, especially for securely storing the Master Key. Although SQRL emphasizes security, it necessitates users to store their Master

Key on a device, typically a smartphone. However, this dependency can be problematic on devices with limited RAM or storage capacity. Storing cryptographic keys demands memory resources, and older or budget smartphones may not meet these requirements. Consequently, SQRL's effectiveness may be constrained by the user's choice of hardware, potentially excluding those with less advanced devices from adopting this authentication method.

Vulnerability to DNS Spoofing

SQRL faces a critical challenge related to its susceptibility to DNS spoofing. DNS spoofing involves manipulating the Domain Name System (DNS) to redirect users to fraudulent websites. SQRL, in its basic form, lacks inherent mechanisms to detect DNS spoofing, making it susceptible to such attacks. If a user unknowingly accesses a spoofed website, their SQRL authentication could be compromised, leading to unauthorized access to their accounts. This vulnerability underscores the importance of implementing additional security measures, such as secure DNS protocols, alongside SQRL to mitigate the risks associated with DNS spoofing.

Device Compatibility and Resource Limitations

SQRL's effectiveness is contingent on the user's choice of devices. While modern smartphones are well-equipped to handle the cryptographic demands of SQRL, older or less powerful devices may struggle due to limited RAM and storage. This can create a barrier to entry for users who do not possess the latest hardware, potentially limiting the widespread adoption of SQRL.

Lack of Built-in DNS Protection

SQRL, as a standalone authentication method, does not offer built-in protection against DNS spoofing. Users and service providers must take additional steps, such as implementing secure DNS protocols, to safeguard against this type of attack. This places the onus on both parties to ensure the security of the authentication process.

In conclusion, SQRL presents a compelling alternative to traditional authentication methods, prioritizing security and user privacy. However, its limitations, including dependence on additional applications, vulnerability to DNS spoofing, and compatibility issues with resource constrained devices, should be carefully considered. To maximize its effectiveness, users and service providers must implement additional security measures and address these challenges proactively. As technology evolves, addressing these limitations will be crucial for SQRL's continued success in the authentication landscape.

6 Conclusion

During this work, SQRL technology was analyzed in order to see whether it can replace traditional SFA and 2FA Identity Methods.

It was found that SQRL technology is a technology for secure authentication over the internet, using modern devices such as Smartphones. It provides several advantages over traditional means of authentication. The analysis shows that the main vulnerability is mainly caused by the way of Use and the errors of the user, commonly the human factor.

Currently there is not a plethora of SQRL implementations and its integration in several applications. Due to this, the impact of application errors cannot be determined. However, history shows that many

points of vulnerability come from the field of authentication. Security checks must be performed on both applications and server applications. Devices infected with malware used for authentication enable identity theft. Practice shows that malware is persistent. This justifies the need for as safe an environment as possible. The proposed solutions are a never-ending process. SQRL supports identity recovery with Rescue Code in case the user's identity has been exposed or stolen. Current technology does not have an automated "lock" and "change of identity" process, and the processes must be performed by users themselves on all websites they visit.

7 References

[1] SQRL Translations. CrowdIn.com. [Accessed: September 30, 2023] [URL: https://www.crowdin.com/project/sqrl]

[2] Gibson, Steve. "Secure Quick Reliable Login: A highly secure, comprehensive, easy-to use replacement for usernames, passwords, reminders, one-time-code authenticators... and everything else." GRC.com. Gibson Research Corporation. [Accessed: March 7, 2021] [URL: https://www.grc.com/sqrl/sqrl.htm]

[3] Gibson, Steve. "SQRL Q&A #176 (Transcript)." Security Now!. Gibson Research Corporation. [Accessed: October 16, 2013] [URL: https://www.grc.com/sn/sn-176.htm]

[4] Babioch, Karol. "Security Analysis and Implementation of the SQRL Authentication Scheme (BSc)." IT Security, Department of Informatics, Technical University of Munich. [Accessed: March 18, 2015]

[5] Gibson, Steve. "How SQRL Can Thwart Phishing Attacks." GRC.com. Gibson Research Corporation. [Accessed: March 7, 2021] [URL: https://www.grc.com/sqrl/phishing.htm]

[6] "Secure QR Login." Drupal.org. https://www.drupal.org/project/sqrl] [Accessed: December 2022] [URL:

[7] Persson, Daniël. "SQRL Login – WordPress plugin." WordPress.org. [Accessed: November 2019] [URL: https://www.bestpractices.dev/pt-BR/projects/3269]

[8] Sylvester, Paul. "SQRL implementations on Android and it works!" Paul's Tech Talk. [Accessed: December 2014] [URL: "SQRL implementations on Android and it works!"]

[9] Lambert, Patrick. "SQRL: A new method of authentication with QR codes." Tech Republic. [Accessed: 2013] [URL: "SQRL: A new method of authentication with QR codes"]

[10] Holmlund, Daniel. "Authentication Without Passwords Implementing SQRL." 2014 HTML5 Developer Developer Conference. Silicon Valley International Game Developers Association. [Accessed: January 3, 2014] [URL: Authentication Without Passwords Implementing SQRL]

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 131 – 137

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Fault Detection in Analog Circuits by utilizing the Current Supply Transients

Vassilios D. Vassios[1] ,Argirios.T.Hatzopoulos[1] , Ioannis G. Intzes[1] and Dimitrios K. Papakostas[1]

[1]*Department of Information and Electronic Engineering International Hellenic University Thessaloniki, Greece*
vassios@yahoo.com, ahatz@ihu.gr, intzes@iee.ihu.gr, dpapakos@ihu.gr

Abstract

An improved method in fault detection in Analog Circuits is presented in this paper. The CUTs are subjected to a short undervoltage, and the transient current is measured. From the current measurement of the power supply during this time the mean and fundamental rms values are calculated. From these measurements for both the mean and the fundamental values three points of interest are extracted. The maximum/minimum over/undershooting values, the relaxation point values, and the damping point values. These twelve points of interest comprise the signature for comparison for all the subsequent CUTs. The algorithm is tested both in simulation and with actual measurements with actual CUTs.

1 Introduction

Fault detection in analog and mixed signal circuits is a continuous field of research and development. A vast variety of techniques and algorithms have been proposed over the years. Fault detection is a practice the IC manufacturers and electronic devices industries use to achieve higher quality and reliability for their customers. Also, these techniques help to improve the repair times of machines that use these electronic devices[1].

Numerous methodologies and algorithms are proposed for fault identification, which include fault dictionary methods [2], IPS measurement [3], the rms approach[4] and many others. Several different signatures can be created by using several metrics, including magnitude/phase [5], spectral analysis [6], wavelet decomposition [7], and a multitude of mathematical techniques such as statistical analysis (standard deviation) [8], topological analysis (Malahanobis distance) [9]. For this paper, the Under/Over Voltage (UOV) algorithm is presented. In this algorithm the IPS is measured, and the moving average/fundamental rms algorithms are used to produce the signature fields and calculate the region of interest in the waveforms without performing any complex or time-consuming calculations.

2 Review of the Algorithm

Analog and mixed signal circuits continue to function properly even if they experience a small fluctuation in their power supply line. The Under/Over Voltage algorithm relies on this property. The

proposed algorithm processes the power supply's Current and calculates the overshooting, undershooting and damping values of the supply current's waveforms in respect to the time they occurred.

The algorithm consists of several steps, which are analyzed in more detail in the paragraph below.

Step 1) Wait for the CUT to stabilize in its normal operating mode.

Step 2) The CUT suffers a small under-voltage in its power supply line (Negative Step).

Step 3) The power supply current to the CUT is measured and the mean, true rms and fundamental rms values are calculated.

Step 4) The maximum/minimum value for the current over/undershooting and the damping point for both the Mean and fundamental rms are calculated along with the time they occurred.

Step 5) The CUT is driven back to its normal voltage supply values (Positive Step).

Step 6) Steps 3 and 4 are repeated one more time for the Positive Step and all the corresponding current and time value pairs are calculated.

Step 7) The twelve different current value/time pairs, six the mean values and six for fundamental rms values, comprise the signature library.

Step 8) Using Monte Carlo simulation, steps 1 through 7 are repeated to calculate the variance σ and generate a detection area of the typical current/time values that can be compared to all other measurements.

The measurements on actual CUTs in Step 8 are substituted for the Monte Carlo simulations by the actual experimental data.

3 Implementation of the Algorithm

A small and simple CE amplifier, as shown in Fig.1. The circuit's behavior operation was first simulated with the PSpice module from Altium Designer 20 and then the actual circuit was designed and constructed for verification or disproval of the simulation results. The circuit is stimulated with a sinusoidal signal that has a 1kHz frequency and an amplitude of 10mV(p-p) to drive it. In the experimental setting the supply current is measured with the use of a current sense amplifier by measuring the differential voltage drop along a shunt resistor[11]. Additionally, a digital to analog converter is used to generate the sinusoidal signal from a microcontroller. Step 2 changes the circuit's power consumption state, or the I_{PS}, by reducing the supply voltage by 10% from its nominal value. These actions are depicted in Fig. 2. The waveforms are separated into two sections. The transition from 100% to 90% of the power-supply voltage constitutes the first section and will be referred as the Negative Step (NS), and the transition from 90% to 100% of the power-supply voltage constitutes the second section and will be referred as the Positivity Step (PS). The time duration for both steps is defined as the time between transitions and is extracted from trial-and-error simulations. The same algorithm was used to process the two sections separately.

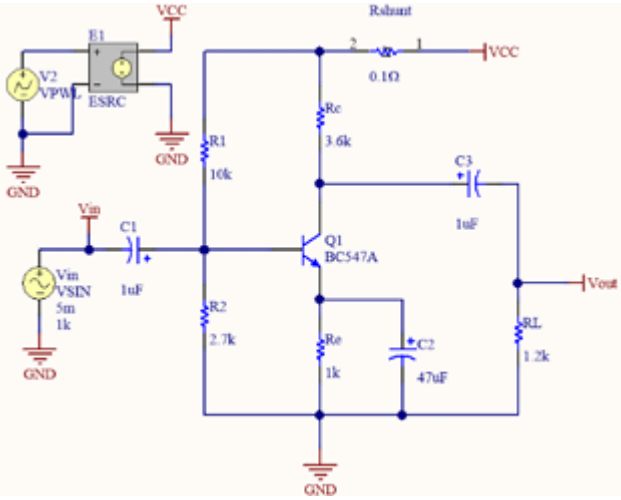


Figure 1. Schematic Description of the CE amplifier

From the author's previous work [22], it became apparent that the mean and the true rms calculations have very close values which forces the algorithm to be unable to recognize a significant number of Faults. In that scope, the "fundamental rms" is used in the current work. These similarities and differences between the mean, true rms and fundamental rms are shown in Fig.3.

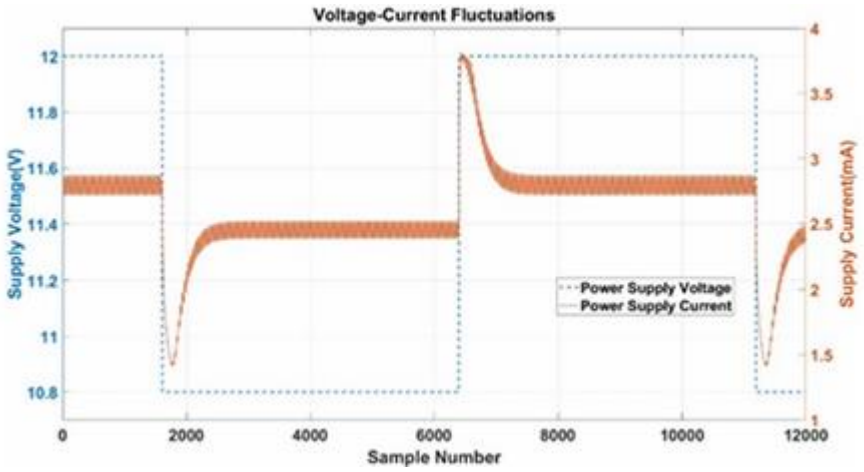


Figure 2. Voltage and Current Supply Waveforms

For the fundamental rms calculation, a slightly different mathematical formula is used. The formula for the true rms is given by equation (1). The supply current is comprised of two components, the DC and AC, which are denoted in equation (2).

$$I_{rms} = \sqrt{\frac{1}{T}\int_0^T I_{PS}(t)^2 dt} \qquad (1) \qquad \text{where } I_{ps}(t) = I_{DC} + I_{AC}(t) \qquad (2)$$

I_{DC} is the DC component and $I_{AC}(t)$ is a periodical and symmetrical around zero, time dependent function. By replacing the current, IPS(t) in equation (1) it takes the form,

$$I_{AC(rms)} = \sqrt{I_{rms}^2 - I_{DC}^2} \qquad (3)$$

From this point on, the value of $I_{AC(rms)}$ will be denoted as fundamental rms and it represents the AC component of the power supply current IPS. In Step 3, using the moving mean/rms technique, the mean and rms values for each step are calculated. The signal shown in Fig 2 is decomposed to the signals in Fig 3 with the use of equation (3) and moving mean and fundamental rms which represent the DC and the AC components of the Negative Step section of the IPS current.

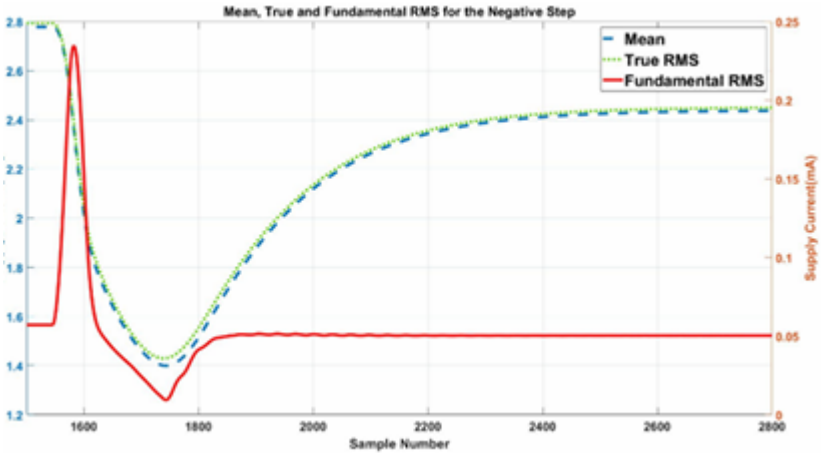


Figure 3. Mean, True and Fundamental RMS component of the Negative Step of the IPS current.

In Step 4, the values for the maximum overshooting and relaxation points along with the current/time pairs for each local point are calculated. To determine the current/time values of the relaxation point the angle of the slope of the signal is used. This angle is set to be 1o which is very close to stabilization line.

For the calculation of the damping point, a straight line is created which passes through both maximum and relaxation points. The gradient of the straight line is calculated. The damping point is defined to be the first point after the maximum overshooting point that has the same gradient value. The Points and the respected regions of interest are shown in Fig 4.

In Steps 5 and 6 the process of Steps 1 through 4 is repeated but for the transition of the power supply voltage from 90% to 100% of the nominal value (Positive Step), which produces similar signals.

The twelve current/time pairs from the mean, fundamental rms local maximum minimum, damping points and relaxation points for both positive and negative step are used to produce the signature in Step 7. The set of signatures that make up the training set are determined by the standard deviation created in Step 8 using the Monte Carlo simulations with a uniform distribution and a 10% value tolerance for every component of the circuit. A region where all measurements are thought to belong to a good/working CUT is created with the use of the standard deviation.

Repeating steps 1 through 7 while injecting errors into the CUT is the subsequent stage. For each CUT a fault introduced and a set of points will be extracted. Each new signature will be evaluated against the No-Fault signature by determining if the current/time pairs of the Faulty CUT signature are inside the detection area. At least one current/time pair needs to be outside the detection region for a faulty CUT to be detected. All twelve current/time pairs must be inside the detection area for the CUT to be classified as good. In all other cases, the CUT is classified as faulty. Hard faults (open and shorts) are inserted to the CUT to test the algorithm.

A short circuit to a specific component is indicated by placing a resistance of $R_p \leq 1\Omega$ in parallel to the corresponding component. To model an open circuit to component, the corresponding component is replaced by a resistance of value $R_p = 10M\Omega$ [10].

4 Experimental Results

The algorithm's overall detectability for the simulations was also employed to a Negative Feedback Amplifier and the results are shown in Table 1. The fundamental rms method managed to detect all injected faults of the CUT. For the experimental setup the algorithm is trained with 60 good circuits to obtain the points and regions of interest. Then the algorithm is tested with 100 CUTs for both test circuits, with a 50%/50% proportion for, Good/Faulty CUTs. The detection region varied from ±1σ to ±6σ. The result was that by decreasing the size of the detection region, the number of false negatives results (good CUTs being recognized as Faulty) increased and, vice-versa by increasing the size of the detection region the number of false positives (Faulty CUTs being recognized as good) increased.

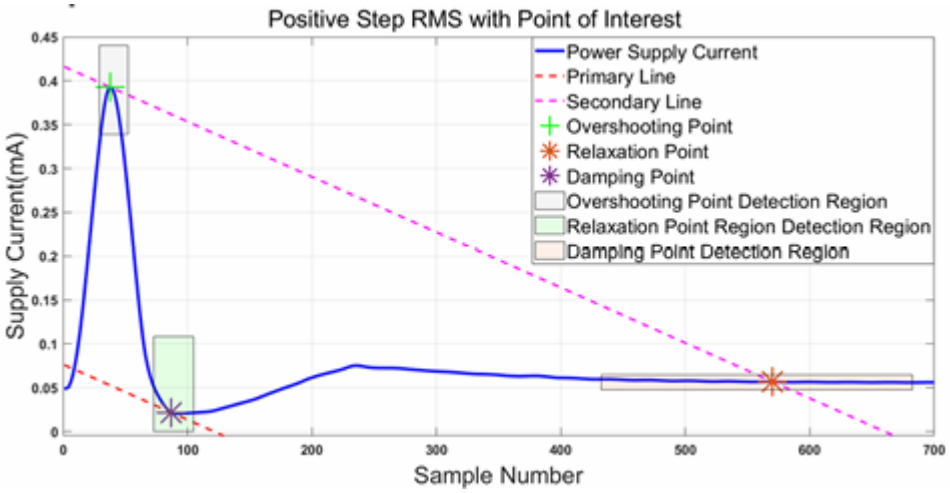


Figure 4. Positive Step with Points and regions of Interest.

Table II shows the detectability, in absolute numbers, of the algorithm in relation to the size of the detection region compared to the author's previous work [11]. It shows that a detection region with size of ±3σ around the mean value of the points of interest yields better detectability and accuracy.

Circuit	Number of Faults	Detected Faults without damping point	Detected Faults with damping point	Detectability(%) without damping point	Detectability(%) with damping point
Common Emitter Amplifier	22	22	22	100	100
Negative Feedback Amplifier	46	46	46	100	100

Table 1: Fault Detection results

Detection Range	True Positive without Damping Point	True Positive with Damping Point	True Negative without Damping Point	True Negative with Damping Point	False Positive without Damping Point	False Positive with Damping Point	False Negative without Damping Point	False Negative with Damping Point
±1σ	37	43	50	50	0	0	13	7
±2σ	41	48	50	50	0	0	9	2
±3σ	50	50	50	50	0	0	0	0
±4σ	50	50	44	49	6	1	0	0
±5σ	50	50	39	43	11	7	0	0
±6σ	50	50	32	38	18	12	0	0

Table 2: Fault Detectability for the Negative Feedback Amplifier

5 Conclusion

This paper presents an enhanced technique for detecting faults in analog circuits. This technique utilizes the fundamental RMS function, which can separate a signal into its basic DC and AC components.

The proposed technique calculates the overshooting/damping/stabilization values of the IPS of the CUT with respect to the time they occur. The algorithm applies the moving mean and moving rms algorithms to the IPS measurements to provide a signal that is smoother and can be analyzed in greater detail. Two circuits made use of the algorithm. A two-stage amplifier with negative feedback and a CE amplifier. 22 hard faults were introduced into the CE amplifier, and 46 hard faults were injected into the two-stage negative feedback amplifier. In comparison to the authors' previous work the algorithm was able to identify 100% of the faults on the CE amplifier and 100% of the injected faults for the negative feedback amplifier and reduced the number of false positives and false negatives in respect with the size of the detection range. The algorithm improved the previous detection range from ±3σ to ±2σ.

Based on the results, the next step is to improve the algorithm so that it can be applied to operational amplifier circuits in all of their operating modes, filters, amplifiers, and comparators. The algorithm at the time is refined so that it can operate both in unipolar and bipolar circuits.

6 References

[1] B.Olleta, H.Jiang, D.Chen, R.L.Geiger" Methods of testing analog and mixed signal using dynamic element matching for source linearization", US Patent Number:7.587.647B2, September 2009.

[2] Milne A, Taylor D, Naylor K, Assessing and comparing fault coverage when testing analogue circuits. In: Proceedings of IEEE conference on circuit devices and systems, vol. 144, 1; 1997. p. 1–4.

[3] R. Kondgunturi, E. Bradley, K. Maggard, C. Stroud "Benchmark circuits for analog and mixed signal testing.", Conference Proceedings - IEEE SOUTHEASTCON (1999) 1999-March 217 220

[4] A.A. Hatzopoulos, "Analog Circuit Testing", Proceedings of the 2017 IEEE 22nd International Mixed-Signals Test Workshop, IMSTW 2017.

[5] D. K. Papakostas, A.A. Hatzopoulos, "A Unified Procedure for Fault Detection of Analog and Mixed-Mode Circuits Using Magnitude and Phase Components of the Power Supply Current

Spectrum", IEEE Transactions on instrumentation and measurement, vol. 57, No. 11, November 2008.

[6] A D. Spyronasios, M G. Dimopoulos, A. A. Hatzopoulos, "Wavelet Analysis for the Detection of Parametric and Catastrophic Faults in Mixed-Signal Circuits", IEEE Transactions on instrumentation and measurement, vol. 60, No. 6, July 2011.

[7] M.F Toner, G.W. Roberts, "A BIST scheme for an SNR test of a sigma-delta ADC," IEEE Int. Proc. Test. Conference 1993.

[8] C. T. Chen ,C. T. Yen, C. H. Wen, C. Y. Yang, C. H. Wu, K. C. Chern, M. Chen, Y. Y. Kuo, C. Y. Lee, J. N. Kao, S. Yi, "CNN-based Stochastic Regression for IDDQ Outlier Identification", Proceedings of the IEEE VLSI Test Symposium, Vol: April 2020.

[9] K. Wang*, Y. Guana, D. Lib, X. Lic, J. Guand,"Research on Fault Diagnosis of Analog Circuit Based on Volterra Theory and Higher-Order Spectrum Analysis", IOP Conf. Series: Materials Science and Engineering 782 (2020) 032096.

[10] P. Kabisatpathy, A. Barua, S. Sinha, "Fault Diagnosis of Analog Integrated Circuits", 1st ed., Springer, 2005, pp 30-31.

[11] V. Vassios, A. Hatzopoulos, D. Papakostas," Improved Fault Detection of Analog Circuits by utilizing the Fundamental RMS of the Supply Current Fluctuation",12th International Conference on Modern Circuits and Systems Technologies (MOCAST23), University of West Athens, Athens, pp. 28-30, June 2023

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 138 – 143

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Aging alleviation technique for 8T IMC SRAMs

Dounavi Helen-Maria[1] and Tsiatouhas Yiorgos[1]

[1] *Dept. of Computer Science and Engineering, VCAS Lab, University of Ioannina, Greece*
edounavi@cse.uoi.gr, tsiatouhas@cse.uoi.gr

Abstract

CMOS Static Random Access Memories (SRAMs) are widely used for In-Memory Computing (IMC) in modern systems to achieve fast and efficient logic and arithmetic computations. However, aging, such as BTI, is a serious threat to the reliability of the SRAMs operations, affecting also significantly the results of the IMC operations in these memories. Hence, it is crucial to develop aging mitigation strategies to maintain the reliability of the memory. The current work proposes an aging alleviation technique for 8T CMOS SRAMs that are frequently used for IMC, by adopting a special purpose and commonly exploited in conventional 8T SRAMs source line, with proper voltage bias on it during non-active periods, as an effective solution to mitigate aging.

1 Introduction

The Von-Neumann architecture, which isolates the memory from the processing unit, is regularly employed in systems dedicated for computing. Yet, data-centric applications such as artificial intelligence, have serious performance limitations due to the requirement for data movements between the memory and the compute cores (Zhiting, L. et al, 2022), (Si, X. et al, 2019). IMC, which realizes data processing within the memory, is an efficient method for reducing the costs related to data transfer. SRAM memory on the other hand, is extensively used in cutting-edge technologies since it is fast, less power-demanding and reliable. As a result, SRAM-based IMC approaches give great potential for speeding up a wide range of applications, which have a significant impact on the future of the computing industry. In SRAM-IMC the complete execution of logical and arithmetic operations is realized within the SRAM. A variety of SRAM-IMC designs in literature propose computations to be carried out directly on the bitlines for the acceleration of the IMC logical operations (Yin, S. et al., 2020), (Jaiswal, A. et al., 2019).

Nevertheless, Bias Temperature Instability (BTI) poses a significant risk to the accuracy of SRAM IMC computations. Timing failures related to transition delays are caused by BTI; an aging phenomenon which causes the absolute threshold voltage value of a transistor under stress to rise over time. nMOS transistors are affected by Positive BTI (PBTI), whereas pMOS transistors are affected by Negative BTI (NBTI). The reliability of the SRAM-IMC is seriously degraded, and under the influence of strong BTI, logic operations carried out within the SRAM can produce false results. The intermediate output value in SRAM-IMC for logic operations is a voltage level, digitized afterwards to the final output. The aging of IMC cell's transistors causes the response voltage levels to change, producing incorrect output

results. To provide sustainable in-memory computations, it is vital to build an aging alleviation strategy. For SRAM-IMC, a variety of memory architectures have been proposed. Numerous works (Jaiswal, A. et al., 2019), (Chang, W. et al., 2021), (Agrawal, A. et al., 2018) have employed the typical 8T SRAM cells for logic IMC operations instead of the common 6T SRAM cells due to read write disturb issues in them. The 8T SRAM IMC structure suggested in (Chang, W. et al., 2021) is used as a case study in the present work. The main concept is the layout modification of the typical 8T SRAM by separating a transistor source line from the ground (Gnd) and properly biasing it, to alleviate the impact of BTI on the transistors within the memory's computation path.

2 Preliminaries

The typical 8T SRAM architecture commonly utilized in contemporary IMC technologies is the main focus of this study. As depicted in Figure 1a, two additional to the typical 6T cell nMOS transistors (MN5 and MN6), create the read path through the extra Read bitline (RBL), dividing the read from the write operation which is performed through the traditional Write bitline (WBL). This modification does not alter the write operation while, for the read operation, the memory cells are accessed through the additional Read wordline (RWL) (set to logic '1'), maintaining the Write wordline (WWL) to low (logic '0'). The RBL is initially pre-charged to Vdd and during the read operation the value stored in the memory cell determines the output RBL voltage level (V_{RBL}). In read mode, when the value stored in the cell is logic 1 ($Q = '1'$) transistors MN5 and MN6 are activated and the VRBL drops through the activated reading path. When the cell's stored value is logic '0' ($Q = '0'$) the RBL remains charged to V_{dd}. Reading can be executed with a substantial voltage swing on the RBL when using the 8T cell's decoupled read port, eliminating completely any read disturb failures (Jaiswal, A. et al., 2019). A part of the memory array with two 8T SRAM cells attached on a RBL is depicted in Figure 1b, for the demonstration of the read paths formed in a column, through the corresponding MN5 and MN6 transistors of the cells. This architecture, exploited by (Chang, W. et al., 2021) for IMC operations, is hereby explored as a case study of an 8T SRAM-IMC scheme (Figure 1b). Initially, a pair of RWLs (such as RWL1 and RWL2) are enabled to perform an IMC logic operation, and then the combined stored value of Cell1 and Cell2 determines the VRBL, within a specified time duration. Then, to carry out various logic operations (OR, AND, NOR, NAND), the read out V_{RBL} is transferred to properly skewed inverters for digitization. In Figure 1c, the VRBL voltage levels are presented for the different stored values stored in Cell1 and Cell2 accordingly (cases 00, 01, 10 and 11 are given). Each properly skewed inverter, treats the V_{RBL} as a logic '1' or '0' respectively, delivering the response of the corresponding logic operation.

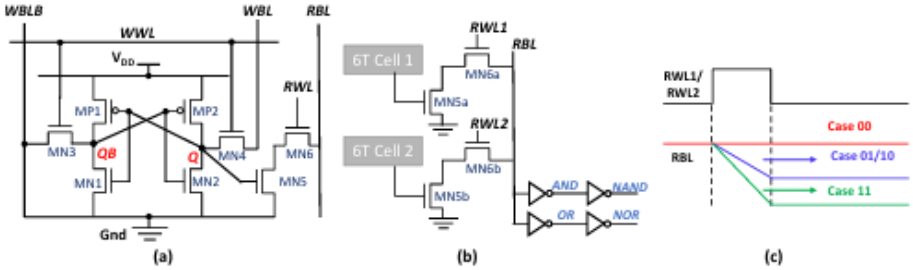


Figure 1: (a) The typical 8T SRAM cell structure, (b) an 8T SRAM-IMC column, with skewed inverters to sense logic operations, (c) the voltage levels of RBL under different cases of stored values in the cells

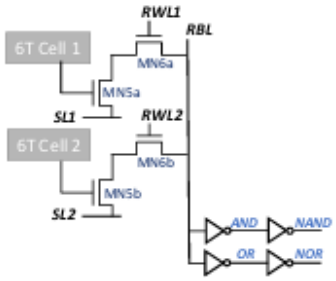


Figure 2: The proposed 8T SRAM aging alleviation modifications

However, SRAM memory cells are prone to transistor aging, such as BTI, raising the likelihood of false IMC responses, which is a serious reliability issue in contemporary technologies. Extreme stress circumstances, such as high temperatures and gate-to-source voltage levels, hasten BTI aging even more. Numerous solutions have been proposed to mitigate the effects of NBTI, nevertheless, with technology evolution, PBTI also poses a serious threat to SRAM-IMC (Jaiswal, A. et al., 2019), (Chen, Y.-G. et al., 2022). To properly address the PBTI-induced issue, (Chang, W. et al., 2021) suggests a modified 8T SRAM IMC architecture with supplemental transistors and the use of an aging detection algorithm to determine the health state of each memory array row. Additional area overhead is required for the implementation of the aging detection scheme. More recently, the authors in (Chen, Y.-G. et al., 2022) suggests adjusting the operating supply voltage during computation to counteract the deteriorating RBL discharge current brought on by the effects of aging. This solution needs both an aging monitoring method and a dynamic voltage frequency scaling mechanism, while it presents an important increase in the circuit's average power consumption.

With a minimal area overhead and a small power consumption penalty, the current work seeks to alleviate PBTI-induced reliability degradation on the nMOS network of the reading/computing path of the 8T SRAM IMC structure presented in Figure 1. The typical 8T SRAM cell is modified with the separation of the MN5 transistor source line from the ground (Gnd), and the application of an appropriate voltage level to it, as explained below. It should be noted that any potential aging of the 6T Cells does not affect the IMC reliability and thus, the aging of MN5 and MN6 transistors is examined.

3 Proposed technique

An 8T CMOS SRAM aging alleviation approach is proposed to achieve accurate IMC results and maintain the memory's reliable operation. Initially, note that the reading operation corresponds only to a brief time interval of the total SRAM's operating time. During this time, the gate of the MN6 transistors (Figure 1) are set to high (logic '1') and the transistors only enter the stress state during this brief phase. Then, the gate is set back to low (logic '0') and the transistors return to the relaxed state for the rest operation of the memory. Hence, aging does not have a significant impact on MN6 transistors. On the contrary, when a logic '1' value is the one stored in a memory cell ($Q = '1'$), the accompanying MN5 transistor constantly experiences extreme DC stress since VGS=Vdd for as long as the cell's stored value is kept the same. Because AC BTI stress involves recovery cycles that minimize the influence on Vt, DC stress is the main reason for transistor's Vt degradation.

In the proposed 8T SRAM architecture, depicted in Figure 2, the source terminals of MN5 transistors are separated from Gnd and act as independent source lines (SL). To reduce the PBTI on MN5, a positive

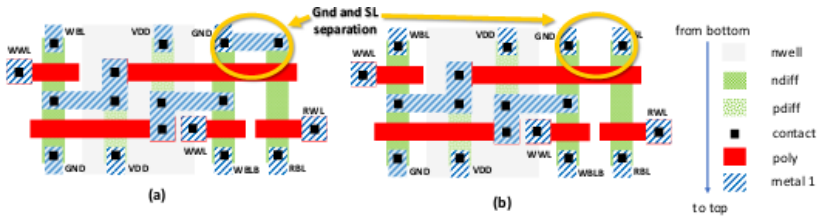


Figure 3. (a) Layout of the standard 8T SRAM cell
(b) Layout for the 8T SRAM cell with separated Gnd and SL

voltage $V_{SL} > 0$ (eg. V_{dd}) is by default applied on the SL lines in all memory rows when they are not activated for any operation (inactive state). As these periods correspond the vast majority of the memory's operational time, it follows that for MN5 transistors $V_{GS} < V_{dd}$ for most of their lifetime. Thus, MN5 transistors experience extreme BTI stress only when the related cell is activated for reading or computation, remaining to a decreased BTI stress or even to a fully relaxed state (when our selection is $V_{SL} = V_{dd}$) during the rest of the cell's prolonged inactive period, reducing the impact of PBTI on it. Note that the stress on MN5 will be negligible as that on MN6, when VSL is equal to or close to V_{dd}

To maintain the memory operation, an SL is turned to Gnd when a reading/computing operation is initiated and the pertinent RWL is set to high. Since the reading path through MN5 leads to Gnd when Q='1', the RBL will be discharged, whereas when Q='0', MN5 will be inactive and the cell will not contribute to the RBL's voltage drop. To examine the additional area cost of the suggested solution, Figure 3a depicts the typical 8T SRAM cell layout and Figure 3b the modified 8T SRAM cell layout of the suggested method, where the SL and Gnd nodes are separated. Note that the SL line is shared by all MN5 transistors of every two consecutive rows. As shown, there is no additional silicon area cost at the memory cell level. However, due to the routing requirements of higher-level power supply metal lines, the exploitation of the SL line as a separate line will result in a 14% area overhead penalty at memory array level. Nevertheless, according to the literature, the adoption of a separate source line, with the same area cost, is a common practice in order to support other special functionality in 8T SRAMs (Jaiswal, A. et al., 2019).

4 Simulation Results

Proper simulations on a memory array structure were conducted for the study of the aging effect on the 8T SRAM IMC. An 8T SRAM array with 256 rows and 64 cells per row, according to the topology of Figure 1a, was designed (with Virtuoso of CADENCE) and simulated (with Spectre) exploiting the 28nm CMOS technology of UMC ($V_{dd} = 0.9V$). Figure 4 illustrates the outcomes of IMC operations for the case where two 'fresh' cells are activated with 01/10 as stored values (Figure 4a) or 11 (Figure 4b) accordingly. Once the RBL voltage level falls below a selected threshold value, the skewed inverters (Figure 1) output the results of the relevant logic operation. Under severe PBTI stress, MN5 transistors' current will not develop an adequate VRBL level, within the IMC time duration, lower than the skewed inverters' transition threshold, resulting in a false result. To assess the impact of the PBTI on the IMC results, different threshold voltage shift (ΔV_t) levels were applied to MN5 transistors, by inserting a DC voltage source of proper polarity at their gate terminals. When only one cell has stored logic '1' (01/10 scenario), Figure 5 shows the transition of the RBL and its final voltage level at IMC mode for the "fresh" and the aged case (MN5 transistor with $\Delta V_t = 50mV$ or 100mV). It also depicts the skewed inverter's transition threshold. For the aged cases, the inverter's transition threshold is violated and a

failure is produced as expected.

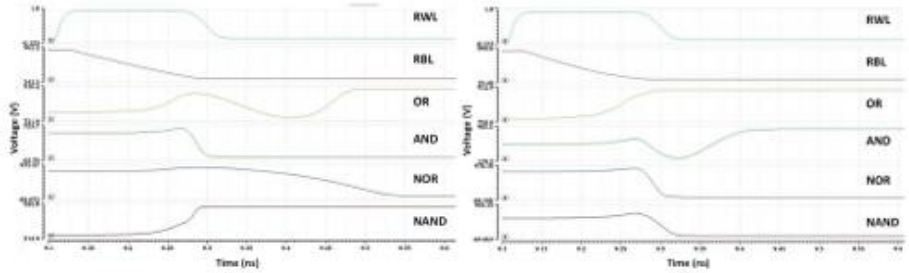


Figure 4. (a) The result of IMC when 01/10 are stored and **(b)** when 11 is stored in the memory cells.

To further accentuate BTI influence, Table 1 presents the required delay time for the generation of the logic result and the V_{RBL} after the NOR and NAND operations are performed, for both MN5 transistors' "fresh" and aged (at $\Delta V_t = 50mV$) state. Since case 00 will not result to an important RBL voltage change in either a fresh or an aged state, it has not been listed in the Table. It is clear that PBTI has a significant impact on IMC results. Biasing the SL at appropriate voltage levels above zero and near V_{dd}, the voltage stress on MN5 transistors is drastically reduced or even eliminated (when $V_{SL} = V_{dd}$) so that the BTI effect is significantly alleviated maintaining the reliability of the memory operation.

The suggested technique guarantees the SRAM's reliable operation with a small but acceptable area overhead penalty. The simulations also revealed that the proposed method will result in a 8.52μW increase of the overall dynamic power consumption in the memory array, over 13.73μW of the original design, during the read/compute operations when $V_{SL} = V_{dd}$, while it is smaller for lower voltage levels (e.g. 2.32μW for VSL=Vdd/2). Note that there is not any additional delay penalty, since the discharging/charging of the SL is achieved during the charging/discharging of the pertinent wordline.

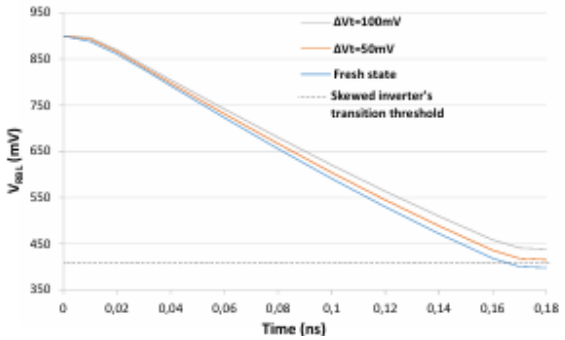


Figure 5: RBL voltage level for different ΔVt (01/10 scenario)

	Operation	Delay (ps)	V_{RBL} (mV)
Fresh state	NOR (Case 01/10)	400	400
	NOR (Case 11)	170	116
	NAND (Case 11)	200	116
ΔVt= 50mV	NOR (Case 01/10)	Failed	416
	NOR (Case 11)	180	126
	NAND (Case 11)	Failed	126

Table 1: Delay time and remaining RBL voltage level for (a) the fresh state and (b) under 50mV of aging

5 Conclusion

In this work, a scheme and a method for the alleviation of PBTI-induced aging in the reading/computing path of 8T IMC SRAMs is proposed. It involves the layout modification of the typical 8T SRAM cell for the insertion of an additional properly biased source line. When the SRAM is active, the source lines of the cells that are accessed are set to the predefined voltage level that is suitable for the corresponding operation. However, during the SRAM's non-active periods, the source line is set to a selected positive voltage level in order to alleviate aging.

The proposed technique does not induce any performance degradation, inserts a small, but acceptable according to the common practice, area overhead and increases the power consumption depending on the used bias voltage. On the other side, it offers an easy to implement solution against the influence of BTI aging on read/computing performance, that ensures the reliable memory operation throughout its lifetime, a crucial issue in high reliability systems. Previously presented techniques, mitigate aging by periodically monitoring the memory and properly adjusting its operation. These approaches affect performance and may not be effective in case of overaged transistors.

6 Acknowledgments

We acknowledge support of this work by the project "Dioni: Computing Infrastructure for Big-Data Processing and Analysis." (MIS No. 5047222) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Program "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

7 References

[1] Agrawal, A. et al. (2018). X-SRAM: Enabling In-Memory Boolean Computations in CMOS Static Random Access Memories. IEEE Trans. Circ. Syst. I (TCAS-I), vol. 65, no. 12, pp. 4219-4232.

[2] Chang, W. et al. (2021). An Aging-Aware CMOS SRAM Structure Design for Boolean Logic In Memory Computing. IEEE Intern. Symp. on Defect and Fault Toler. in VLSI and Nanotech. Syst. (DFT), (pp. 1-4). Athens.

[3] Chen, Y.-G. et al. (2022). Aging-Compromised Computing-In-Memory Dot-Product Calculation Technique Through DVFS. 24th Workshop on Synth. and Syst. Integr. of Mixed Inf. Tech. (SASIMI), (pp. 44-47). Hirosaki.

[4] Jaiswal, A. et al. (2019). 8T SRAM Cell as a Multibit Dot-Product Engine for Beyond Von Neumann Computing. IEEE Trans. on VLSI Syst., vol. 27, no. 11, pp. 2556-2567.

[5] Si, X. et al. (2019). A Dual-Split 6T SRAM-Based Computing-in-Memory Unit-Macro With Fully Parallel Product-Sum Operation for Binarized DNN Edge Processors. IEEE Trans. on Circ. and Syst. I: Regular Papers, vol. 66, no. 11, pp. 4172-4185.

[6] Yin, S. et al. (2020). XNOR-SRAM: In-Memory Computing SRAM Macro for Binary/Ternary Deep Neural Networks. IEEE J. of Solid-State Circ., vol. 55, no. 6, pp. 1733-1743.

[7] Zhiting, L. et al. (2022). A review on SRAM-based computing in-memory: Circuits, functions, and applications. J. Semicond., vol. 43, no. 3, pp. 173–210.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 144 – 150

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Reduction of large-scale RLCk models via low-rank balanced truncation

Christos Giamouzis, Dimitrios Garyfallou, Anastasis Vagenas, and Nestor Evmorfopoulos

Dept. of Electrical and Computer Engineering, University of Thessaly, Volos, Greece
{cgiamouzis, digaryfa, avagenas, nestevmo}@e-ce.uth.gr

Abstract

Model order reduction (MOR) is an important step in the design process of integrated circuits. Specifically, the electromagnetic models extracted from modern complex designs result in a large number of passive elements that introduce limitations in the simulation process. MOR techniques based on balanced truncation (BT) can overcome these limita tions by producing compact reduced-order models (ROMs) that approximate the behavior of the original models at the input/output ports. In this paper, we present a low-rank BT method that exploits the extended Krylov subspace and efficient implementation tech niques for the reduction of large-scale models. Experimental evaluation on a diverse set of analog and mixed-signal circuits with millions of elements indicates that up to ×5.5 smaller ROMs can be produced with similar accuracy to ANSYS RaptorX™ ROMs.

1 Introduction

Electromagnetic model extraction plays a key role in the design and analysis of integrated circuits. The extracted models are simulated to accurately predict the behavior of the passive elements of the design. Model order reduction (MOR) can reduce the complexity of RLCk models with many elements (>1M) and ports (>10), while retaining an accurate approximation of the input and output behavior of the circuit [1, 2]. Therefore, the simulation time of complex systems can be radically decreased by constructing reduced-order models (ROMs) of smaller dimensions that preserve the essential characteristics of the original models.

MOR methods are distinguished into two main categories. Moment matching (MM) techniques [1] are preferred due to their computational efficiency. However, they rely on an ad hoc selection of the number of moments, which correlates the final ROM size with the number of ports. On the other hand, techniques based on balanced truncation (BT) [2] offer reliable bounds for the approximation error and have no fundamental limitation to the number of ports they can handle, resulting in more compact ROMs. Nevertheless, BT applies only to small-scale models since it involves the computationally expensive solution of Lyapunov equations [2].

In this work, appropriate performance improvements are explored to overcome the main drawback of the conventional BT method. To this end, we adopt an efficient low-rank technique based on the extended Krylov subspace (EKS) for solving the Lyapunov equations. The proposed approach can be integrated into industrial extraction tools, such as the ANSYS Rap torX™ [3], to obtain more compact ROMs of large-scale multi-port RLCk models.

2 Background

Consider the modified nodal analysis (MNA) description [4] of an n-node, m-branch (inductive), p-input, and q-output RLCk circuit in the time domain:

$$\begin{pmatrix} G_n & E \\ -E^T & 0 \end{pmatrix}\begin{pmatrix} v(t) \\ i(t) \end{pmatrix} + \begin{pmatrix} G_n & 0 \\ 0 & M \end{pmatrix}\begin{pmatrix} \dot{v}(t) \\ i(t) \end{pmatrix} = \begin{pmatrix} B_i \\ 0 \end{pmatrix} u(t), \;\; y(t) = \begin{pmatrix} L_1 & 0 \end{pmatrix}\begin{pmatrix} v(t) \\ i(t) \end{pmatrix} \tag{1}$$

where $G_n \in \mathbb{R}^{n\times n}$ (node conductance matrix), $C_n \in \mathbb{R}^{n\times n}$ node capacitance matrix), $M \in \mathbb{R}^{m\times n}$ (branch inductance matrix), $E \in \mathbb{R}^{n\times m}$ (node-to-branch incidence matrix), $V \in \mathbb{R}^n$ (vector of node voltages),
$i \in \mathbb{R}^m$ (vector of inductive branch currents), $u \in \mathbb{R}^p$ (vector of input excitations), $B_1 \in \mathbb{R}^{n\times p}$ (input-to-node connectivity matrix), $y \in \mathbb{R}^q$ (vector of output measurements), and $L1_l \in \mathbb{R}^{q\times n}$ (node-to-output connectivity matrix). Moreover, we denote $\dot{v}(t) \equiv \frac{dv(t)}{dt}$and $i(t) \equiv \frac{di(t)}{dt}$. If we now define the model order as $N \equiv n + m$, the state vector as $x(t) \equiv \begin{pmatrix} v(t) \\ i(t) \end{pmatrix}$, and also:

$$G \equiv -\begin{pmatrix} G_n & E \\ -E^T & 0 \end{pmatrix}, \qquad C \equiv \begin{pmatrix} G_n & 0 \\ 0 & M \end{pmatrix}, \qquad B \equiv \begin{pmatrix} B_i \\ 0 \end{pmatrix}, \qquad L \equiv \begin{pmatrix} L_1 & 0 \end{pmatrix}$$

then Eq. (1) can be written in the generalized state-space form, or so-called descriptor form:

$$C\,\frac{dx(t)}{dt} = Gx(t) \; + \; Bu(t), \quad y(t) = Lx(t) \tag{2}$$

The objective of MOR is to produce an equivalent ROM:

$$\tilde{C}\,\frac{d\tilde{x}(t)}{dt} = \tilde{G}\tilde{x}(t) \; + \; \tilde{B}u(t), \quad \tilde{y}(t) = \tilde{L}\tilde{x}(t) \tag{3}$$

Where $\tilde{G}, \tilde{C} \; \in \mathbb{R}^{rxr}$, $\tilde{B} \; \in \mathbb{R}^{rxp}$, $\tilde{L} \in \mathbb{R}^{qxr}$, the reduced order r << N, and the output error is bounded as $\|\tilde{y}(t) - y(t)\|_2 < \varepsilon\|u(t)\|_2$ for given u(t) and small ε. The output error bound can be expressed in the frequency domain as $\|\tilde{y}(s) - y(s)\|_2 < \varepsilon\|u(s)\|_2$ via Plancherel's theorem [5]. If

$$H(s) = L(sC - G)^{-1}B, \qquad \tilde{H}(s) = \tilde{L}(s\tilde{C} - \tilde{G})^{-1}\tilde{B}$$

are the transfer functions of the original model and the ROM, the corresponding output error is:

$$\|\tilde{\mathrm{y}}(\mathrm{s}) - \mathrm{y}(\mathrm{s})\|_2 = \left\|\tilde{\mathrm{H}}(\mathrm{s})\mathrm{u}(\mathrm{s}) - \mathrm{H}(\mathrm{s})\mathrm{u}(\mathrm{s})\right\|_2 \le \left\|\tilde{\mathrm{H}}(\mathrm{s}) - \mathrm{H}(\mathrm{s})\right\|_\infty \|\mathrm{u}(\mathrm{s})\|_2 \tag{4}$$

Where $\|.\|_\infty$ is the $\mathcal{L}_2$ matrix norm or $\mathcal{H}_\infty$ norm of a rational transfer function. Thus, to bound this error, we need to bound the distance between the transfer functions: $\left\|\tilde{\mathrm{H}}(\mathrm{s}) - \mathrm{H}(\mathrm{s})\right\|_\infty < \varepsilon$.

3 MOR by Balanced Truncation

BT relies on the computation of the controllability Gramian P and observability Gramian Q, which are calculated as the solutions of the following Lyapunov matrix equations [2]:

$$(C^{-1}G)P + P(C^{-1}G)^T = -(C^{-1}B)(C^{-1}B)^T, \;\; (C^{-1}G)^TQ + Q(C^{-1}G) = -L^TL \tag{5}$$

The controllability Gramian P characterizes the input-to-state behavior, i.e., the degree to which the states are controllable by the inputs, while the observability Gramian Q characterizes the state-to-output behavior, i.e., the degree to which the states are observable at the outputs. In principle, a ROM can be obtained by eliminating the states that are difficult to reach or observe. However, in the original

state-space coordinates, there are states that are difficult to reach but easy to observe, and vice versa. The process of "balancing" transforms the state vector to a new coordinate system, where for each state, the degree of difficulty is the same for both reaching and observing it. An appropriate transformation Tx(t) exists, leading to the following model:

$$TCT^{-1}\frac{d(Tx(t))}{dt} = TGT^{-1}(Tx(t)_- + TBu(t),\ \ y(t) = LT^{-1}(Tx(t)) \qquad (6)$$

that preserves the transfer function H(s). This renders $P = Q = diag(\sigma_1, \sigma_2, \ldots \sigma_N)$ where σ_i are known as the Hankel singular values (HSVs) of the model and are equal to the square roots of the eigenvalues of product PQ, i.e., $\sigma_i = \sqrt{\lambda_i(PQ)}$. In the above balanced model, the states that are easier to reach and observe correspond to the largest HSVs. If *r* of them are preserved (truncating the *N* − *r* states corresponding to the smallest HSVs), it can be shown that the distance between the original and the reduced-order transfer functions is bounded as:

$$\left\|\tilde{\mathrm{H}}(\mathrm{s}) - \mathrm{H}(\mathrm{s})\right\|_\infty < 2(\sigma_{\mathrm{r}+1} + \sigma_{\mathrm{r}+2} + \ldots + \sigma_{\mathrm{N}}). \qquad (7)$$

The latter is an "a-priori" criterion for selecting the ROM order for a desired output error tolerance ε, which constitutes a significant advantage of BT over MM techniques. The main steps of the BT procedure are summarized in Algorithm 1. The major drawback of BT is the significant

Algorithm 1 MOR by balanced truncation

1: Solve the Lyapunov equations to obtain the Gramian matrices $\mathbf{P}$ and $\mathbf{Q}$ [6]
2: Compute the SVD of the Gramian matrices: $\mathbf{P} = \mathbf{U}_P\mathbf{\Sigma}_P\mathbf{V}_P^T$ and $\mathbf{Q} = \mathbf{U}_Q\mathbf{\Sigma}_Q\mathbf{V}_Q^T$
3: Find the square root of the Gramian matrices: $\mathbf{Z}_P = \mathbf{U}_P\mathbf{\Sigma}_P^{1/2}$ and $\mathbf{Z}_Q = \mathbf{U}_Q\mathbf{\Sigma}_Q^{1/2}$
4: Compute the SVD of the product of the roots: $\mathbf{Z}_Q^T\mathbf{Z}_P = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$
5: Compute transformation matrices: $\mathbf{T}_{(r\times N)} = \mathbf{\Sigma}_{(r\times r)}^{-1/2}\mathbf{U}_{(r\times N)}\mathbf{Z}_Q^T$, $\mathbf{T}_{(N\times r)}^{-1} = \mathbf{Z}_P\mathbf{V}_{(N\times r)}\mathbf{\Sigma}_{(r\times r)}^{-1/2}$
6: Compute ROM: $\tilde{\mathbf{G}} = \mathbf{T}_{(r\times N)}\mathbf{G}\mathbf{T}_{(N\times r)}^{-1}$, $\tilde{\mathbf{C}} = \mathbf{T}_{(r\times N)}\mathbf{C}\mathbf{T}_{(N\times r)}^{-1}$, $\tilde{\mathbf{B}} = \mathbf{T}_{(r\times N)}\mathbf{B}$, $\tilde{\mathbf{L}} = \mathbf{L}\mathbf{T}_{(N\times r)}^{-1}$

computational and memory cost for deriving the ROM, which hinders the applicability to large scale models (with *N* over a few thousand states). This is because the operations involved (e.g., the solution of Lyapunov equations and the singular value decomposition [SVD]) are computationally expensive with a complexity of $O(N^3)$. Moreover, they are applied on dense matrices, since the Gramians **P, Q** are dense even if the system matrices **C, G ,B**, L are sparse.

However, the products $(C^{-1}B)(C^{-1}B)^T$ *and* L^TL have low numerical order compared to *N*, as *p,q* << *N*, resulting in low-rank Gramian matrices that can be approximated using low-rank techniques. This greatly reduces the complexity and memory requirements of the solution of the Lyapunov equations and the SVD analysis, which are now of order *k* instead of full order *N*.

3.1. Low-rank BT MOR

The essence of low-rank BT MOR is to iteratively project the Lyapunov equations of Eq. (5) onto a lower-dimensional Krylov subspace and then solve the resulting small-scale equations to obtain low-rank approximate solutions of Eq. (5). More specifically, if $K \in \mathbb{R}^{Nxk}$ $(k \ll N)$ is a projection matrix whose columns span the *k*-dimensional Krylov subspace:

$$\Re_k(G_C, B_C) = span\{B_C, G_CB_C, G_C^2B_C, \ldots, G_C^{k-1}B_C\}$$

Where $G_C \equiv C^{-1}G, B_c \equiv C^{-1}B$, then the projected Lyapunov equation (for the controllability Gramian

P) onto $\Re_k(G_C, B_C)$ is:

$$(\mathrm{K^T G_C K})\mathrm{X} + \mathrm{X}(\mathrm{K^T G_C K})^{\mathrm{T}} = +\mathrm{K^T B_C B_C^T K} \tag{8}$$

(the same holds true for the observability Gramian **Q** with G_C^T, $\mathrm{L^T}$ in place of $\mathrm{G_C}$, $\mathrm{B_C}$). The solution $X \in \mathbb{R}^{kxk}$ of Eq. (8) can be back-projected to the N-dimensional space to give an approximate solution **P = KXK^T** for the original large-scale Eq. (5), and a low-rank factor $Z \in \mathbb{R}^{Nxk}$ of P can be obtained as **Z = KUΣ$^{1/2}$**, where **[U, Σ, V] = *SVD*(X)**.

Although the projection process is independent of the subspace selection, its effectiveness is critically dependent on the chosen subspace. The convergence to the final solution can be accelerated by enriching the standard Krylov subspace $\Re_k(G_C, B_C)$ with information from the subspace $\Re_k(G_C^{-1}, B_C)$, which corresponds to the inverse matrix G_C^{-1}, leading to the EKS [7, 8]:

$$\Re_{\mathrm{k}}^{\mathrm{C}}(\mathrm{G_C}, \mathrm{B_C}) = span\{\mathrm{B_C}, \mathrm{G_C^{-1}B_C}, \mathrm{G_C B_C}, \mathrm{G_C^{-2}B_C}, \mathrm{G_C^2 B_C} \ldots, \mathrm{G_C^{-(k-1)}B_C}, \mathrm{G_C^{k-1}B_C}\} \tag{9}$$

The EKS method (EKSM) starts with the pair $\{\mathrm{B_C}, \mathrm{G_C^{-1}B_C}\}$ and generates an extended subspace $\Re_{\mathrm{k}}^{\mathrm{C}}(\mathrm{G_C}, \mathrm{B_C})$ of increasing dimension, solving the projected Lyapunov Eq. (8) in each iteration, until a sufficiently accurate approximation of the solution of Eq. (5) is obtained. The complete EKSM is presented in Algorithm 2. Below are some efficient implementation details:

- **Matrix inversion by linear solves:** The inputs to Algorithm 2 are not actually $G_C \equiv C^{-1}G$ *or* $G_C^T \equiv (C^{-1}G)^T$ but the system matrices **G**, **C** or **G^T**, **C^T**, since the (generally dense) inverse matrices are only needed in products with p vectors (in step 2) and 2*pj* vectors (in steps 4 and 11 of each iteration). These can be implemented as linear solves **CY = R** and **GY =R** (or ***C^{T}Y=R, G^{T}Y=R***) by any direct or iterative algorithm like [9].
- **Handling of sparse/dense matrices**: Note that matrix **M** of Eq. (1) is highly dense, as it generally includes a huge number of mutual inductances. To effectively handle the sparse (***C$_n$***) and dense (**M**) blocks of matrix **C**, we use efficient data structures and numerical techniques. For example, for linear solves and matrix-vector products, we employ parallel CPU-optimized methods for sparse matrices and leverage GPU-accelerated techniques [10] for dense matrices.
- **Solution of the small-scale Lyapunov equations**: To solve the small-scale (*2pj* × *2pj*) Lyapunov equations in step 5 of each iteration, we employ the Bartels-Stewart algorithm [6].
- **Convergence criterion**: An appropriate stopping criterion is the residual of Eq. (5) with the approximate solution **P = KXK^T** to reach a certain threshold in magnitude, i.e.,

$$\frac{\|G_C K^{(j)} X K^{(j)T} + K^{(j)} X K^{(j)T} G_C + B_C B_C^T\|}{\|B_C B_C^T\|} \leq tol \tag{10}$$

However, this criterion equals to $\|R^T M X\| \leq tol$ [11], which can be computed more efficiently. A tolerance of tol = 10^{-10} is typically adequate to obtain an accurate model.

Algorithm 2 Extended Krylov subspace method for low-rank solution of Lyapunov equations

Input: $\mathbf{G}_C \equiv \mathbf{C}^{-1}\mathbf{G}, \mathbf{B}_C \equiv \mathbf{C}^{-1}\mathbf{B}$ (or $\mathbf{G}_C^T, \mathbf{L}^T$)
Output: $\mathbf{Z}$ such that $\mathbf{P} \approx \mathbf{Z}\mathbf{Z}^T$

$j = 1$; $p = size\_col(\mathbf{B}_C)$
$\mathbf{K}^{(j)} = Orth([\mathbf{B}_C, \mathbf{G}_C^{-1}\mathbf{B}_C])$
while $j < maxiter$ **do**
  $\mathbf{A} = \mathbf{K}^{(j)T}\mathbf{G}_C\mathbf{K}^{(j)}$; $\mathbf{R} = \mathbf{K}^{(j)T}\mathbf{B}_C$
  Solve $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^T = -\mathbf{R}\mathbf{R}^T$ for $\mathbf{X} \in \mathbb{R}^{2pj \times 2pj}$
  if converged **then**
    $[\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}] = \mathbf{SVD}(\mathbf{X})$; $\mathbf{Z} = \mathbf{K}^{(j)}\mathbf{U}\mathbf{\Sigma}^{1/2}$
    break
  end if
  $k_1 = 2p(j-1)$; $k_2 = k_1 + p$; $k_3 = 2pj$
  $\mathbf{K}_1 = [\mathbf{G}_C\mathbf{K}^{(j)}(:, k_1+1 : k_2), \mathbf{G}_C^{-1}\mathbf{K}^{(j)}(:, k_2+1 : k_3)]$
  $\mathbf{K}_2 = Orth(\mathbf{K}_1)$ w.r.t. $\mathbf{K}^{(j)}$
  $\mathbf{K}_3 = Orth(\mathbf{K}_2)$
  $\mathbf{K}^{(j+1)} = [\mathbf{K}^{(j)}, \mathbf{K}_3]$
  $j = j + 1$
end while

4 Experimental Evaluation

4.1. Experimental setup

To evaluate EKSM, we used large-scale RLCk models extracted from different circuits using ANSYS RaptorX™ [3]. These circuits consist of many passive elements, including mutual induc tances. The EKSM ROMsare compared against golden ROMs produced by RaptorX™, through S-parameter plotting. The characteristics of the RLCk models are listed in Table 1. All exper iments were executed on a Linux server with a 3.60 GHz 8-thread CPU and 16 GB of memory.

Model	Initial order	#nodes	#resistors	#capacitors	#inductors	#mutual ind.	#ports
VGA_28	95189	57675	155879	169600	37514	126766838	13
Hybrid_56	98024	59210	112338	290572	38814	165802476	5
Wilkinson_56	100888	60703	115117	271293	40185	193641938	4
VCO_13	104367	61264	604072	596846	43103	188436057	4
CSLNA_56	128574	78046	188842	472573	50528	169339965	9
Wilkinson_28	129087	78263	123254	266710	50824	259462454	4
Hybrid_28	134710	75766	128935	283905	53169	264162513	5
LNACASC_28	162881	96876	774427	684662	66005	323090671	11

Table 1: Detailed characteristics of RLCk models

4.2. Experimental results

The efficiency of the EKSM against RaptorX™ is demonstrated in Table 2. The S-parameters plots of Figure 1 indicate that EKSM achieves accuracy close to that of RaptorX™ while producing roughly ×3.1 more compact ROMs. Although EKSM has higher reduction time and memory requirements, they are still reasonable and can be significantly improved in future work.

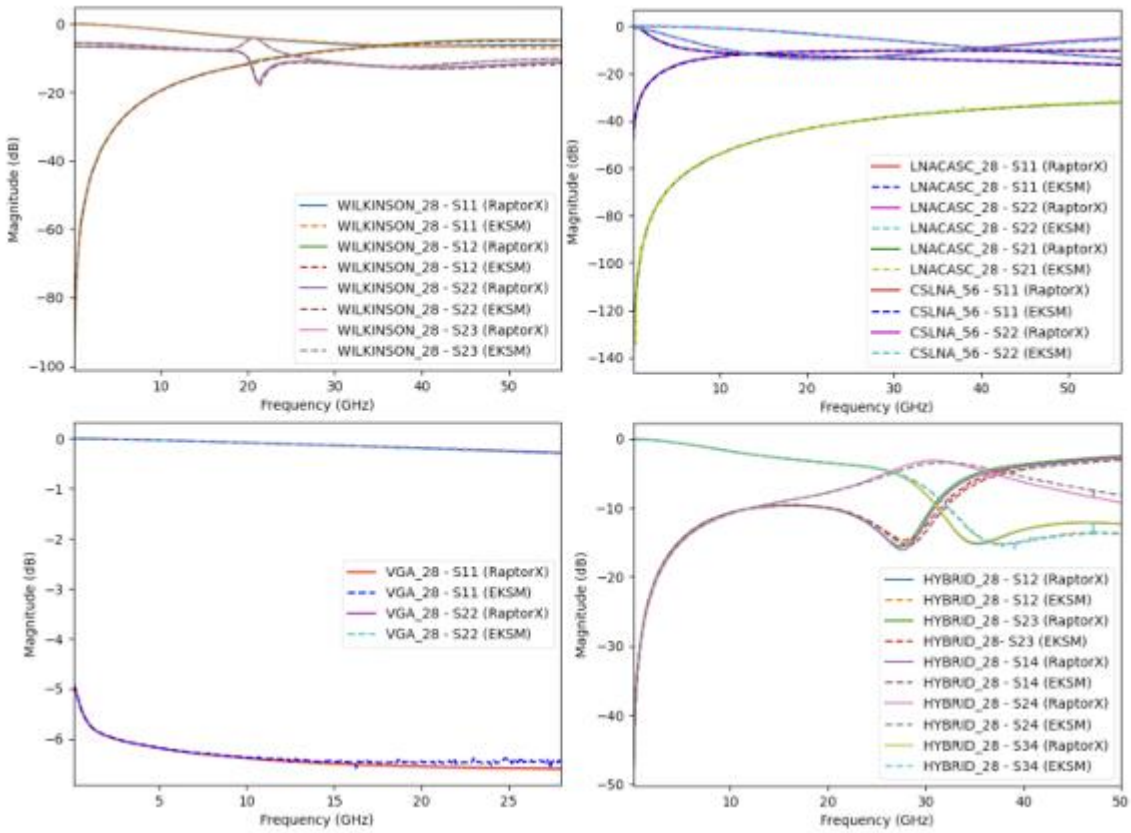


Figure 1: Comparison of accuracy between EKSM and RaptorX™ ROMs.

Model	Initial order	ROM order		Reduction time (s)		Memory (GB)	
		RaptorX™	EKSM	RaptorX™	EKSM	RaptorX™	EKSM
VGA_28	95189	4744	1040	67	1037	5.6	29
Hybrid_56	98024	1267	397	104	613	3.7	44.1
Wilkinson_56	100888	765	320	154	570	3.8	45.1
VCO_13	104367	407	311	119	673	4.6	44.2
CSLNA_56	128574	2172	716	74	1237	4.3	44.1
Wilkinson_28	129087	885	302	205	801	3.9	58.3
Hybrid_28	134710	787	399	217	1032	4	59.2
LNACASC_28	162881	4768	879	373	2866	15.7	73

Table 2: ROM order and MOR performance of EKSM vs RaptorX™

5 Conclusions

Alternative MOR techniques to reduce large-scale RLCk models with accuracy comparable to commercial tools are presented. The proposed low-rank BT method is evaluated across diverse large-scale benchmark circuits by comparing their S-parameters. Experimental results indicate that our approach achieves sufficient accuracy while providing ROMs that are up to ×5.5 smaller than the ROMs obtained by ANSYS RaptorX™.

6 Acknowledgments

This research has been co-financed by the European Regional Development Fund and Greek national funds via the Operational Program ”Competitiveness, Entrepreneurship and Inno vation,” under the call ”RESEARCH-CREATE-INNOVATE” (project code: T2EDK-00609).

7 References

[1] A. Odabasioglu et al., “Prima: Passive reduced-order interconnect macromodeling algorithm,” IEEE Trans. on CAD of Integrated Circuits and Systems, vol. 17, no. 8, pp. 645–654, 1998.

[2] S. Gugercin et al., “A survey of model reduction by balanced truncation and some new results,” International Journal of Control, vol. 77, no. 8, pp. 748–766, 2004.

[3] “Ansys-RaptorX.” [Online]. Available: www.ansys.com/products/semiconductors/ansys-raptorh

[4] C.-W. Ho et al., “The modified nodal approach to network analysis,” IEEE Trans. on Circuits and Systems, vol. 22, no. 6, pp. 504– 509, 1975.

[5] K. Grochenig, Foundations of Time-Frequency Analysis. Birkhauser, 2001.

[6] D. Lathauwer et al., “Computation of the canonical decomposition by means of a simultaneous generalized schur decomposition,” SIAM Journal on Matrix Analysis and Applications, vol. 26, no. 2, pp. 295–327, 2004.

[7] P. Stoikos et al., “The Extended and Asymmetric Extended Krylov Subspace in Moment-Matching Based Order Reduction of Large Circuit Models,” arXiv:2204.02467 [cs.OH], 2022.

[8] C. Chatzigeorgiou et al., “Exploiting Extended Krylov Subspace for the Reduction of Regular and Singular Circuit Models,” in Proc. of the 26th Asia South Pacific Design Automation Conference, pp. 773–778, 2021.

[9] E. Bavier et al., “Amesos2 and Belos: Direct and Iterative Solvers for Large Sparse Linear Sys tems,” Sci. Program., vol. 20, no. 3, p. 241–255, jul 2012.

[10] D. Garyfallou et al., “A Combinatorial Multigrid Preconditioned Iterative Method for Large Scale Circuit Simulation on GPUs,” in Proc. of the 15th International Conference on Synthesis, Model ing, Analysis and Simulation Methods and Applications to Circuit Design, pp. 209–212, 2018.

[11] V. Simoncini, “A new iterative method for solving large-scale lyapunov matrix equations,” SIAM Journal on Scientific Computing, vol. 29, no. 3, pp. 1268–1288, 2007

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 151 – 157

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

MORCIC: Model Order Reduction Techniques for Electromagnetic Models of Integrated Circuits

Dimitrios Garyfallou[1], Athanasios Stefanou[2], Christos Giamouzis[1], Moschos Antoniadis[2], Georgios Chararas[2], Konstantinos Chatzis[2], Dimitris Samaras[2], Rafaela Themeli[2], Anastasios Michailidis[2], Vasiliki Gogolou[2], Nikos Zachos[2], Nestor Evmorfopoulos[1], Thomas Noulis[2], Vasilis F. Pavlidis[2], Alkiviadis Hatzopoulos[2], Elpida Chatzineofytou[3], and Yiannis Moisiadis[3]

[1] *Dept. of Electrical and Computer Engineering, University of Thessaly, Volos, Greece*
{digaryfa, cgiamouzis, nestevmo}@e-ce.uth.gr
[2] *Aristotle University of Thessaloniki, Thessaloniki, Greece*
{tnoul, vpavlid, alkis}@auth.gr
[3] *ANSYS-Hellas, Athens, Greece*
{elpida.chatzineofytou, yiannis.moisiadis}@ansys.com

Abstract

Model order reduction (MOR) is crucial for the design process of integrated circuits. Specifically, the vast amount of passive RLCk elements in electromagnetic models extracted from physical layouts exacerbates the extraction time, the storage requirements, and, most critically, the post-layout simulation time of the analyzed circuits. The MORCIC project aims to overcome this problem by proposing new MOR techniques that perform better than commercial tools. Experimental evaluation on several analog and mixed-signal circuits with millions of elements indicates that the proposed methods lead to ×5.5 smaller ROMs while maintaining similar accuracy compared to golden ROMs provided by ANSYS RaptorX™.

1 Introduction

Electromagnetic model extraction plays a key role in the design and analysis of integrated circuits. The extracted models are simulated to accurately predict the behavior of the passive elements of the design. Model order reduction (MOR) can reduce the complexity of RLCk models with many elements (>1M) and ports (>10), while retaining an accurate approximation of the input and output behavior of the circuit [1, 2]. Therefore, the simulation time of complex systems can be radically decreased by constructing reduced-order models (ROMs) of smaller dimensions that preserve the essential characteristics of the original models.

MOR methods are distinguished into two main categories. Moment matching (MM) techniques [1] are preferred due to their computational efficiency. However, they rely on an ad hoc selection of the number of moments, which correlates the final ROM size with the number of ports. On the other hand, techniques based on balanced truncation (BT) [2] offer reliable bounds for the approximation error and have no fundamental limitation to the number of ports they can handle, resulting in more compact ROMs. Nevertheless, BT applies only to small-scale models since it involves the

computationally expensive solution of Lyapunov equations [2].

In this work, appropriate performance improvements are explored to overcome the main drawback of the conventional BT method. To this end, we adopt an efficient low-rank technique based on the extended Krylov subspace (EKS) for solving the Lyapunov equations. The proposed approach can be integrated into industrial extraction tools, such as the ANSYS Rap torX™ [3], to obtain more compact ROMs of large-scale multi-port RLCk models.

2 Background

Consider the modified nodal analysis (MNA) description [4] of an n-node, m-branch (inductive), p-input, and q-output RLCk circuit in the time domain:

$$\begin{pmatrix} G_n & E \\ -E^T & 0 \end{pmatrix}\begin{pmatrix} v(t) \\ i(t) \end{pmatrix} + \begin{pmatrix} G_n & 0 \\ 0 & M \end{pmatrix}\begin{pmatrix} \dot{v}(t) \\ i(t) \end{pmatrix} = \begin{pmatrix} B_i \\ 0 \end{pmatrix} u(t), \quad y(t) = \begin{pmatrix} L_1 & 0 \end{pmatrix}\begin{pmatrix} v(t) \\ i(t) \end{pmatrix} \tag{1}$$

where $G_n \in \mathbb{R}^{n\times n}$ (node conductance matrix), $C_n \in \mathbb{R}^{n\times n}$ node capacitance matrix), $M \in \mathbb{R}^{m\times n}$ (branch inductance matrix), $E \in \mathbb{R}^{n\times m}$ (node-to-branch incidence matrix), $V \in \mathbb{R}^n$ (vector of node voltages),
$i \in \mathbb{R}^m$ (vector of inductive branch currents), $u \in \mathbb{R}^p$ (vector of input excitations), $B_1 \in \mathbb{R}^{n\times p}$ (input-to-node connectivity matrix), $y \in \mathbb{R}^q$ (vector of output measurements), and $L1_l \in \mathbb{R}^{q\times n}$ (node-to-output connectivity matrix). Moreover, we denote $\dot{v}(t) \equiv \frac{dv(t)}{dt}$and $i(t) \equiv \frac{di(t)}{dt}$. If we now define the model order as $N \equiv n + m$, the state vector as $x(t) \equiv \begin{pmatrix} v(t) \\ i(t) \end{pmatrix}$, and also:

$$G \equiv -\begin{pmatrix} G_n & E \\ -E^T & 0 \end{pmatrix}, \qquad C \equiv \begin{pmatrix} G_n & 0 \\ 0 & M \end{pmatrix}, \qquad B \equiv \begin{pmatrix} B_i \\ 0 \end{pmatrix}, \qquad L \equiv \begin{pmatrix} L_1 & 0 \end{pmatrix}$$

then Eq. (1) can be written in the generalized state-space form, or so-called descriptor form:

$$C\frac{dx(t)}{dt} = Gx(t) + Bu(t), \quad y(t) = Lx(t) \tag{2}$$

The objective of MOR is to produce an equivalent ROM:

$$\tilde{C}\frac{d\tilde{x}(t)}{dt} = \tilde{G}\tilde{x}(t) + \tilde{B}u(t), \quad \tilde{y}(t) = \tilde{L}\tilde{x}(t) \tag{3}$$

Where $\tilde{G}, \tilde{C} \in \mathbb{R}^{rxr}$, $\tilde{B} \in \mathbb{R}^{rxp}$, $\tilde{L} \in \mathbb{R}^{qxr}$, the reduced order r << N, and the output error is bounded as $\|\tilde{y}(t) - y(t)\|_2 < \varepsilon\|u(t)\|_2$ for given u(t) and small ε. The output error bound can be expressed in the frequency domain as $\|\tilde{y}(s) - y(s)\|_2 < \varepsilon\|u(s)\|_2$ via Plancherel's theorem [5]. If

$$H(s) = L(sC - G)^{-1}B, \qquad \tilde{H}(s) = \tilde{L}(s\tilde{C} - \tilde{G})^{-1}\tilde{B}$$

are the transfer functions of the original model and the ROM, the corresponding output error is:

$$\|\tilde{\mathrm{y}}(\mathrm{s}) - \mathrm{y}(\mathrm{s})\|_2 = \left\|\tilde{\mathrm{H}}(\mathrm{s})\mathrm{u}(\mathrm{s}) - \mathrm{H}(\mathrm{s})\mathrm{u}(\mathrm{s})\right\|_2 \leq \left\|\tilde{\mathrm{H}}(\mathrm{s}) - \mathrm{H}(\mathrm{s})\right\|_\infty \|\mathrm{u}(\mathrm{s})\|_2$$

Where $\|.\|_\infty$ is the $\mathcal{L}_2$ matrix norm or $\mathcal{H}_\infty$ norm of a rational transfer function. Thus, to bound this error, we need to bound the distance between the transfer functions: $\left\|\tilde{\mathrm{H}}(\mathrm{s}) - \mathrm{H}(\mathrm{s})\right\|_\infty < \varepsilon$. To achieve this, BT transforms the original model into a ROM with a "balanced" state vector and then truncates the joint controllability-observability singular values of the system (so-called Hankel singular values) that sum up to the given threshold ε, as described in [2].

3 MOR by Balanced Truncation

3.1. Initial BT MOR

BT relies on the computation of the controllability Gramian P and observability Gramian Q, which are calculated as the solutions of the following Lyapunov matrix equations [2]:

$$(C^{-1}G)P + P(C^{-1}G)^T = -(C^{-1}B)(C^{-1}B)^T,\ (C^{-1}G)^T Q + Q(C^{-1}G) = -L^T L \qquad (4)$$

The main steps of the BT procedure are summarized in Algorithm 1. As can be seen, the operations involved (e.g., the solution of Lyapunov equations and the singular value decomposition [SVD]) are computationally expensive with complexity O(N3). Moreover, they are applied on dense matrices, since the Gramians P,Q are dense even if the system matrices C,G,B,L are sparse. Consequently, the significant computational and memory cost for deriving the ROM hinders the applicability of BT to large-scale models (with order N over a few thousand states).

Algorithm 1 MOR by balanced truncation

Input: $\mathbf{G}, \mathbf{C}, \mathbf{B}, \mathbf{L}$
Output: $\tilde{\mathbf{G}}, \tilde{\mathbf{C}}, \tilde{\mathbf{B}}, \tilde{\mathbf{L}}$

1: Solve the Lyapunov equations to obtain the Gramian matrices $\mathbf{P}$ and $\mathbf{Q}$ [6]
2: Compute the SVD of the Gramian matrices: $\mathbf{P} = \mathbf{U}_P \mathbf{\Sigma}_P \mathbf{V}_P^T$ and $\mathbf{Q} = \mathbf{U}_Q \mathbf{\Sigma}_Q \mathbf{V}_Q^T$
3: Find the square root of the Gramian matrices: $\mathbf{Z}_P = \mathbf{U}_P \mathbf{\Sigma}_P^{1/2}$ and $\mathbf{Z}_Q = \mathbf{U}_Q \mathbf{\Sigma}_Q^{1/2}$
4: Compute the SVD of the product of the roots: $\mathbf{Z}_Q^T \mathbf{Z}_P = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$
5: Compute transformation matrices: $\mathbf{T}_{(r\times N)} = \mathbf{\Sigma}_{(r\times r)}^{-1/2} \mathbf{U}_{(r\times N)} \mathbf{Z}_Q^T$, $\mathbf{T}_{(N\times r)}^{-1} = \mathbf{Z}_P \mathbf{V}_{(N\times r)} \mathbf{\Sigma}_{(r\times r)}^{-1/2}$
6: Compute ROM: $\tilde{\mathbf{G}} = \mathbf{T}_{(r\times N)} \mathbf{G} \mathbf{T}_{(N\times r)}^{-1}$, $\tilde{\mathbf{C}} = \mathbf{T}_{(r\times N)} \mathbf{C} \mathbf{T}_{(N\times r)}^{-1}$, $\tilde{\mathbf{B}} = \mathbf{T}_{(r\times N)} \mathbf{B}$, $\tilde{\mathbf{L}} = \mathbf{L} \mathbf{T}_{(N\times r)}^{-1}$

However, the products $(C^{-1}B)(C^{-1}B)^T$ *and* $L^T L$ have low numerical order compared to N, as p, $q << N$, resulting in low-rank Gramian matrices that can be approximated using low-rank techniques. This greatly reduces the complexity and memory requirements of the solution of the Lyapunov equations and the SVD analysis, which are now of order k instead of full order N.

3.2. Low-rank BT MOR

The essence of low-rank BT MOR is to iteratively project the Lyapunov Eq. (4) onto a lowerdimensional Krylov subspace (Kk) [7] and then solve the resulting small-scale equations to obtain low-rank approximate solutions of Eq. (4). In this work, we exploit the EKS to accelerate the convergence to the final solution [8]. The complete EKS method is presented in Algorithm 2.

Algorithm 2 Extended Krylov subspace method for low-rank solution of Lyapunov equations

Input: $\mathbf{G}_C \equiv \mathbf{C}^{-1}\mathbf{G}, \mathbf{B}_C \equiv \mathbf{C}^{-1}\mathbf{B}$ (or $\mathbf{G}_C^T$, $\mathbf{L}^T$)
Output: $\mathbf{Z}$ such that $\mathbf{P} \approx \mathbf{Z}\mathbf{Z}^T$

$j = 1$; $p = size\_col(\mathbf{B}_C)$
$\mathbf{K}^{(j)} = Orth([\mathbf{B}_C, \mathbf{G}_C^{-1}\mathbf{B}_C])$
while $j < maxiter$ **do**
  $\mathbf{A} = \mathbf{K}^{(j)T} \mathbf{G}_C \mathbf{K}^{(j)}$; $\mathbf{R} = \mathbf{K}^{(j)T} \mathbf{B}_C$
  Solve $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^T = -\mathbf{R}\mathbf{R}^T$ for $\mathbf{X} \in \mathbb{R}^{2pj \times 2pj}$
  if converged **then**
    $[\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}] = \mathbf{SVD}(\mathbf{X})$; $\mathbf{Z} = \mathbf{K}^{(j)} \mathbf{U} \mathbf{\Sigma}^{1/2}$
    break
  end if
  $k_1 = 2p(j-1)$; $k_2 = k_1 + p$; $k_3 = 2pj$
  $\mathbf{K}_1 = [\mathbf{G}_C \mathbf{K}^{(j)}(:, k_1 + 1 : k_2), \mathbf{G}_C^{-1} \mathbf{K}^{(j)}(:, k_2 + 1 : k_3)]$
  $\mathbf{K}_2 = Orth(\mathbf{K}_1)$ w.r.t. $\mathbf{K}^{(j)}$
  $\mathbf{K}_3 = Orth(\mathbf{K}_2)$
  $\mathbf{K}^{(j+1)} = [\mathbf{K}^{(j)}, \mathbf{K}_3]$
  $j = j + 1$
end while

4 Experimental Evaluation

To evaluate the proposed MOR methods, we developed an EDA tool that implements the BT algorithms presented in Section 3. As depicted in Figure 1, the only input is a configuration file that defines the path to the MNA matrices along with some parameters. After applying BT MOR, the tool performs DC, transient, and SP analysis, to compare the ROM to the original model. The output includes the S-parameters and MNA matrices of the ROM. The cross-platform MORCIC tool was developed in C++ using the CMake automation sotware. All experiments were executed on a Linux workstation with a 3.60 GHz CPU and 16 GB of memory.

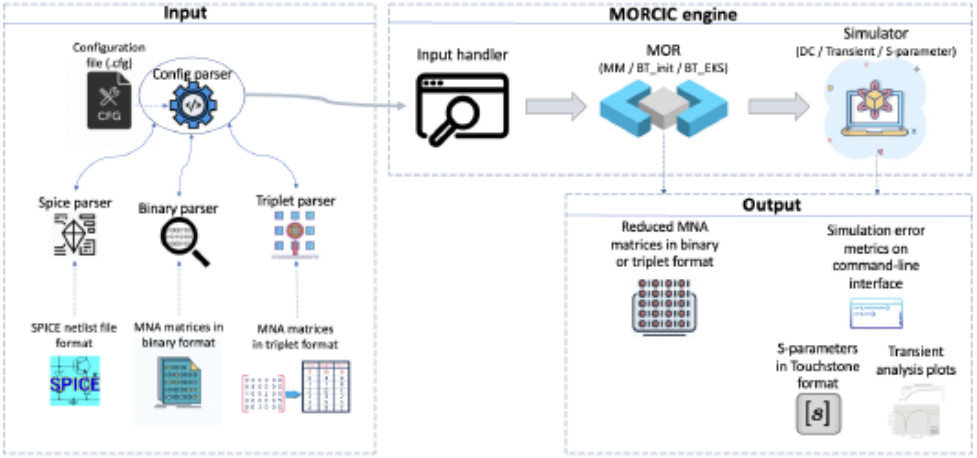


Figure 1: Software architecture of the MORCIC tool.

4.1 Initial BT MOR

For the evaluation of the initial BT MOR method, we used small-scale RC and RLCk models (i.e., real transmission lines) extracted by ANSYS RaptorX™ [3], which are presented in Table 1.

Table 1: Small-scale RC and RLCk models of transmission lines

Model	Initial order	#nodes	#resistors	#capacitors	#inductors	#mutual ind.	#ports
RC_1	48	48	202	273	0	0	2
RC_2	526	526	6667	6872	0	0	6
RLCk_1	5431	3084	2998	1282	2347	136271	2
RLCk_2	21800	12166	34635	31131	9634	23639237	6

As can be seen in Table 2, the mean relative error (MRE) and maximum relative error (MAX RE) for the DC analysis are lower than 0.61% and 0.62%, respectively. As for the SP analysis, MRE remains under 0.73% while MAX RE is below 1.89%. The reduction percentage to achieve accurate results is within 65-87%. However, considering the runtime and memory overhead of the initial method, its application on large-scale models is practically infeasible.

4.2 Low-rank BT MOR

To validate the accuracy and performance of the low-rank BT MOR method, we designed disparate circuits in the GlobalFoundries 22 nm FDSOI technology and extracted the corresponding large-scale RLCk models using RaptorX™ [3]. The choice of the benchmark circuits is driven by their diversity and therefore,

Table 2: Evaluation of ROMs generated by the initial BT MOR against the original models

Model	ROM order	Reduction (%)	DC analysis		SP analysis		Reduction time	Memory (GB)
			MRE (%)	MAX_RE (%)	MRE (%)	MAX_RE (%)		
RC_1	17	64.58	0.54	0.55	0.73	0.78	0.04 s	0.01
RC_2	93	82.32	0.09	0.17	0.32	0.43	2.51 s	0.09
RLCk_1	1131	79.18	0.61	0.62	0.25	1.89	1.12 h.	7.24
RLCk_2	2797	87.17	0.46	0.61	0.08	1.38	6 days	87.11

different metrics are used to describe their behavior. As shown in Table 3, the evaluated designs include Hybrid and Wilkinson couplers as well as typical transceiver blocks like low-noise-amplifiers (LNAs) and oscillators, where the metrics of interest are the reflection coefficients and performance (gain, noise, linearity). In our experiments, we utilized the MORCIC tool to generate ROMs with target accuracy comparable to RaptorX™.

Table 3: Large-scale RLCk models and metrics of interest for the designed circuits

Block/DUT	Initial Order	Ports	Mutual inductors	Simulated Metrics
Hybrid Coupler @28GHz	134710	5	79001243	S-parameters of coupler as:
Hybrid Coupler @56GHz	98024	5	52363149	power splitter & divider
Wilkinson Coupler @28GHz	129087	4	259462454	S-parameters of coupler as:
Wilkinson Coupler @56GHz	100888	4	193641938	power splitter & divider
VGA @28GHz	95189	13	40230583	S-parameters, attenuation
VCO @13GHz	104367	4	70445484	Spectrum, PN, osc. frequency
LNA Common-Source @56GHz	128574	9	72832315	S-parameters, gain, CP1dB,
LNA Cascode @28GHz	162881	11	98585323	IIP3, Noise Figure (NF)

The accuracy evaluation is performed by comparing the ROMs generated by the low-rank BT MOR against the reference ROMs obtained by RaptorX™, as the simulation of the original extracted models (i.e., full RLCk netlists) is infeasible. The evaluated metrics for the Hybrid and Wilkinson couplers at 28 GHz, both operating as power splitters, are demonstrated in Figure 2. As can be seen, the S-parameters of the MORCIC ROMs closely match those of the RaptorX™ ROMs across the frequency range, and most importantly at the frequency of interest. The insertion-loss error is lower than 0.5 dB, while the respective phases differ by less than 2 degrees.

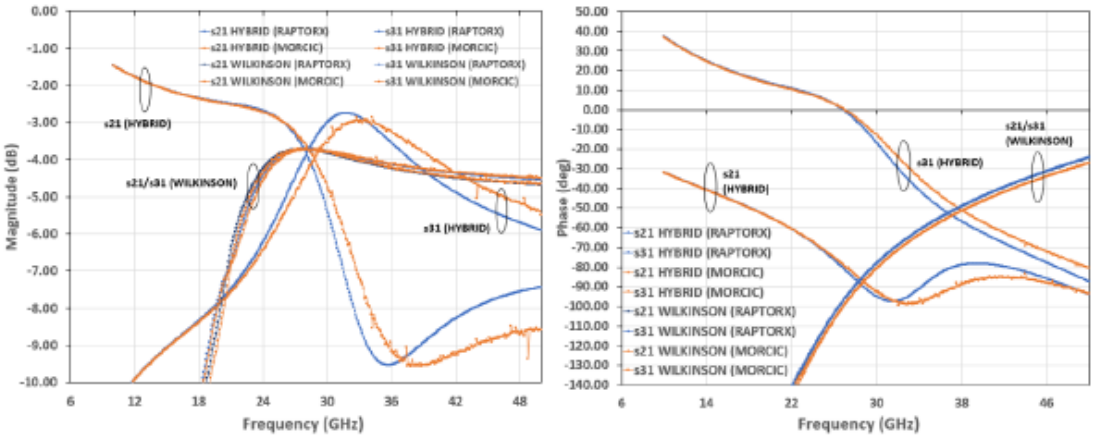


Figure 2: MORCIC vs RaptorX™ ROM accuracy: Hybrid/Wilkinson S-parameters and phases.

The efficiency of the MORCIC tool against RaptorX™ is demonstrated in Table 4. On average, MORCIC

produces ×3.1 more compact ROMs. As the number of ports increases, the advantage of BT is more evident, with the maximum improvement in ROM order reaching ×5.5 for LNACasc 28. Although MORCIC has higher reduction time and memory requirements compared to RaptorX™, they are still reasonable and can be significantly improved in future work.

Table 4: MORCIC vs RaptorX™ ROM order and MOR performance

Model	Initial order	ROM order		Reduction time (s)		Memory (GB)	
		RaptorX™	MORCIC	RaptorX™	MORCIC	RaptorX™	MORCIC
VGA_28	95189	4744	1040	67	1037	5.6	29
Hybrid_56	98024	1267	397	104	613	3.7	44.1
Wilkinson_56	100888	765	320	154	570	3.8	45.1
VCO_13	104367	407	311	119	673	4.6	44.2
LNACS_56	128574	2172	716	74	1237	4.3	40.1
Wilkinson_28	129087	885	302	205	801	3.9	54.3
Hybrid_28	134710	787	399	217	1032	4	53.2
LNACasc_28	162881	4768	879	373	2866	15.7	73

5 Conclusions

In this paper, we present efficient BT MOR techniques to reduce electromagnetic RLCk models. The accuracy of the proposed methods has been evaluated across diverse benchmark circuits, such as the Hybrid/Wilkinson couplers, primarily comparing their S-parameters. Experimental results demonstrate that our low-rank BT MOR approach achieves sufficient accuracy while providing ROMs that are up to ×5.5 smaller than the ROMs obtained by ANSYS RaptorX™.

6 Acknowledgments

This research has been co-financed by the European Regional Development Fund and Greek national funds via the Operational Program ”Competitiveness, Entrepreneurship and Innovation,” under the call ”RESEARCH-CREATE-INNOVATE” (project code: T2EDK-00609).

7 References

[1] A. Odabasioglu et al., “Prima: Passive reduced-order interconnect macromodeling algorithm,” IEEE Trans. on CAD of Integrated Circuits and Systems, vol. 17, no. 8, pp. 645–654, 1998.

[2] S. Gugercin et al., “A survey of model reduction by balanced truncation and some new results,” International Journal of Control, vol. 77, no. 8, pp. 748–766, 2004.

[3] “Ansys-RaptorX.” [Online]. Available: www.ansys.com/products/semiconductors/ansys-raptorh

[4] C.-W. Ho et al., “The modified nodal approach to network analysis,” IEEE Trans. on Circuits and Systems, vol. 22, no. 6, pp. 504– 509, 1975.

[5] K. Grochenig, Foundations of Time-Frequency Analysis. Birkhauser, 2001.

[6] D. Lathauwer et al., “Computation of the canonical decomposition by means of a simultaneous generalized schur decomposition,” SIAM Journal on Matrix Analysis and Applications, vol. 26, no. 2, pp. 295–327, 2004.

[7] V. Simoncini, “A new iterative method for solving large-scale lyapunov matrix equations,” SIAM Journal on Scientific Computing, vol. 29, no. 3, pp. 1268–1288, 2007.

[8] C. Chatzigeorgiou et al., “Exploiting Extended Krylov Subspace for the Reduction of Regular and Singular Circuit Models,” in Proc. of the 26th Asia South Pacific Design Automation Conference, pp. 773–778, 2021.

[9] E. Bavier et al., “Amesos2 and Belos: Direct and Iterative Solvers for Large Sparse Linear Sys tems,” Sci. Program., vol. 20, no. 3, p. 241–255, jul 2012.

[10] D. Garyfallou et al., “A Combinatorial Multigrid Preconditioned Iterative Method for Large Scale Circuit Simulation on GPUs,” in Proc. of the 15th International Conference on Synthesis, Model ing, Analysis and Simulation Methods and Applications to Circuit Design, pp. 209–212, 2018.

[11] V. Simoncini, “A new iterative method for solving large-scale lyapunov matrix equations,” SIAM Journal on Scientific Computing, vol. 29, no. 3, pp. 1268–1288, 2007

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 158 – 165

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Design and implementation of a compact RISC-V based Machine Learning accelerator on Low End FPGA

Manolis Galetakis[1], Stavros Kalapothas[2], Georgios Flamis[2], Paris Kitsos[2], Fotis Plessas[1]

[1] *University of Thessaly, Volos, Greece*
[2] *University of the Peloponnese, Patras, Greece*
egaletakis@uth.gr, s.kalapothas@go.uop.gr, g.flamis@go.uop.gr, kitsos@uop.gr, fplessas@e-ce.uth.gr

Abstract

In recent years, the use of Machine Learning (ML) algorithms on edge devices for common applications such as computer vision, speech recognition etc. is increasing rapidly. With the evolution and the complexity of the latest ML models, inference has become more computationally expensive and therefore, deployment to resource-constraint edge devices is a challenging task. A common technique to address that challenge and increase performance and efficiency, is to offload compute intensive functions from software and run them onto dedicated hardware instead. In this paper we present the complete flow of software-hardware co-design for a lightweight object detection ML model, first time implemented on a RISC-V soft processor. Then we describe software optimizations for the model and a compact hardware accelerator design. We deploy and profile, in terms of cycle count, resource utilization and power consumption of our design on a Xilinx Artix7 35T low-end Field Programmable Gate Array (FPGA) device. Our RISC-V based compact accelerator achieves almost 3.7x inference speedup @ 180 mW power consumption and consumes 5.5K slice LUTs (26.41%) – 4.3K slice registers (10.19%). Therefore, our implementation facilitates object detection for a variety of non-speed demanding embedded applications, at competent speed on small– entry-level FPGA devices.

1 Introduction

Nowadays, accelerator implementations utilizing RISC-V features have emerged rapidly. In our recent article [1] a survey on RISC-V based ML accelerators were conducted emphasizing at the various hardware cores in conjunction with the software frameworks and stacks available to implement FPGA ML inference.

Edge Impulse FOMO (Faster Objects, More Objects) [2], a novel machine learning algorithm which enables object detection on devices with limited hardware resources. FOMO, according to the author, compared to MobileNet SSD or YOLOv5, it can measure and locate objects in an image, as well as monitor numerous objects in real time while using up to 30 times less computing power and memory. The network was trained with a small subset, only two hundred images, of the PASCAL Visual Object Classes 2005 [3] dataset to perform object detection and has achieved slightly over 81% accuracy. As

of this writing the authors cannot find similar implementations of FOMO neural network on a RISC-V based embedded (not SoC-based) FPGA.

When it comes to the deployment of ML accelerators, FPGA devices have emerged as the preferred choice. This is due to its low cost, high degree of reconfigurability, and most importantly, its quick time-to-market. With regard to SoC and software design and implementation, the CFU-Playground [4] is utilized, a full-stack open-source framework that aims to enable rapid prototype development of ML accelerators for FPGA systems. The selection was based on the tool simplicity of design and evaluate FPGA-based hardware accelerators combined with a soft processor, to increase the performance of ML model computation.

In this paper a complete accelerator design is presented using the aforementioned tools. The full design flow is described starting from the network selection and training, following the accelerator design methodology and deployment in a FPGA development platform.

The rest of the paper is organized as follows. In Section 2 a list of RISC-V hardware accelerators is presented. The overall proposed accelerator design architecture is described in Section 3. In Section 4 the experimental results are depicted. Finally, the conclusions and possible improvements are enclosed in Section 5.

2 Related Work

In this section, a diverse array of RISC-V-based hardware accelerator designs, which has been reported in the literature, is presented. The list can be categorized into two types, accelerators implemented as FPGA-based softcores and accelerators available as Application Specific Integrated Circuit (ASIC) designs.

The advantages of using FPGAs for accelerating AI workloads and more specifically Convolutional Neural Networks (CNN), have been extensively studied and reported in [5]–[8] in recent years, such the ones derived from the spatial computing and low-latency nature of FPGAs. RISC-V architecture has also been exploited due to its inherent extensibility with the use of custom ISA extensions, such as vector, or DSP extensions, which can be imported to increase the computation performance in an extended domain of applications. In addition, there are software frameworks that can generate RISC-V synthesizable accelerated cores featuring, for instance, the RV32V, or any other custom ISA extension.

In [9], the authors have demonstrated a 5-stage RV32I barrel processor for matrix-matrix product operations with little less than 7K LUTs utilization in FPGA, in its minimum 8-thread configuration. Also, in [10], a RISC-V software processor based on VexRiscv core and synthesized on a FPGA with 9K LUTs utilization for multiplication operations is presented. In [11], a cluster of RISC-V cores as an accelerator overlay based on the Parallel Ultra Low Power (PULP) platform is implemented on a FPGA with more than 20K LUTs utilization in most configurations. Moreover in [12], a scalable vector processor in RTL synthesizable core files is introduced, that uses novel techniques for high-performance and cost-effective implementation.

Arnold in [13], is a fully programmable RISC-V microcontroller unit (MCU) fabricated in 22nm as a System-on-Chip (SoC) for better performance and energy efficiency. GAP-8 in [14], is a ultra-low power RISC-V SoC for Edge-AI applications with a focus on battery operated scenarios. The Kendryte K210 SoC [15], is a low-cost ASIC with a dual-core 64-bit capable RISC-V that implements the RV64GC ISA and has demonstrated 0.25 TOPS performance on a fixed-point precision CNN model at 400MHz. Lastly, another fabricated and commercially available hardware accelerator is the

MAX78000 MCU [16], which incorporates a 32-bit RISC-V coprocessor. The main role of the RISC-V is to act as a smart direct memory access (DMA) and move data between the CNN and on-chip memory.

3 Proposed Accelerator Design

To profile and analyze the FOMO model "The Step-by-Step Guide to Building an ML Accelerator" from CFU-Playground documentation [17], has been followed. The final trained model has a total of 27 layers, with the majority being convolutional and depthwise convolutional, fifteen and six respectively, and the remaining layers being added, pad and softmax. At first, we compiled the model and run inference (object detection) on 96 by 96 pixel grayscale images. As expected, the majority of time 95.4% is consumed in the two main convolutional layers. The convolution function is presented in Algorithm 1 in the form of pseudocode. The inner loops, lines four to thirteen hereinafter called for simplicity "the inner loop", also described as fetch data from memory – calculate quantized dot product – store result back in memory. On the other hand, the loops in lines one to three hereinafter also called for simplicity "the outer loop" simply takes care that the correct data are fetched from the memory for the algorithm to operate correctly. To further profile this inner loop the RISC-V performance counters, are utilized. Our observations at the cycle count of the outer loop compared to the total spend in the convolution function depict that almost 90% of the time is spent in the inner loop. In conclusion the portion of code, i.e., the inner loop, is the best candidate for optimization and therefore we should focus on it. The same conclusion derives from the analysis of the code for depthwise convolution layer. The next two paragraphs quote a short description of both software and hardware optimization that were applied.

3.1. Software Optimizations

Prior any hardware accelerator has been developed, the ML model can be accelerated by further parameterizing in software.

After inspection of the two convolution functions, it is observed that the possible convolution parameters (padding, filter width, height etc.) are constant. However, these parameters can be replaced, with literal values, directly in the source code.

Algorithm 1 Original 2D Software Convolution in CFU Playground

Parameter: X, Y, Z = *output matrix dimensions*
D, C = *weights matrix dimensions*
Input: I = *image data matrix*, W = *weights matrix*, B = *bias*,
S = *scaling vector*, O = *offsets vectors*
Output: R = *dot product output*

```
1  for x = 0 to X do
2    for y = 0 to Y do
3      for z = 0 to Z do
4        for h = 0 to D do
5          for w = 0 to D do
6            for c = 0 to C do
7              R^z_{x,y} += W^{z,c}_{h,w} * (I^c_{x+h,y+w} + O_i)
8            end for
9          end for
10       end for
11       R^z_{x,y} += B^z
12       R^z_{x,y} *= S^{z-1}
13       R^z_{x,y} += O^z_r
14     end for
15   end for
16 end for
```

Algorithm 1: Convolution Software Function

There are some variables' value checking inside the code that is predetermined for our model for convolution or depth wise convolution layers. Some examples of such checks include the existence of a bias value, or the input data is inside image that both are always true. These checks can also be stripped away. It is also feasible to eliminate loops controlled by parameters with fixed values, such as batches or the width and height of filters, without impacting the outcome. With respect to Algorithm 1 for convolutional layer, in the last loop in lines six to eight, the value of filter depth is always a multiple of four, except the first layer. Taking into consideration that the greatest common factor of these layers is eight and the values of filter and input data are continuous in memory, we can unroll this loop by a factor of eight. The effect of these optimizations is summarized in Fig. 2 for convolution and depthwise convolution (orange column) showing speed up improvement by a factor of 2.3x.

3.2. Hardware Optimizations

Following the software optimizations applied previously, when the last loop unrolled by a factor of eight this equals to eight discrete multiply and accumulate operations in each iteration. In the original code, the algorithm operates on data and filter values stored in byte arrays. But our soft core uses 32-bit wide registers which are contiguous in memory. This means that we can perform single instruction multiple data (SIMD) multiply and accumulate calculations thus having the same results with two discrete operations per iteration. Focusing more on the convolution function from the perspective of computation, the result from the accumulation is multiplied by a quantized multiplier, after bias addition and finally limited between a constant min and max value. All this functionality is realized in hardware with the final accelerator architecture shown in Fig. 1.

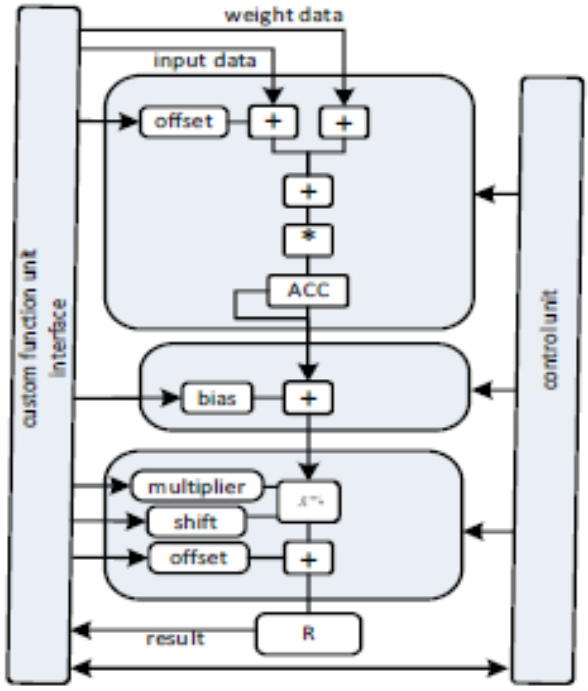


Figure 1: Accelerator Architecture

The accelerator is connected to the soft processor through the custom function unit (CFU) interface using RISC-V custom command a process thoroughly described in the documentation of CFU

Playground Framework. Three units namely mac, bias and quantizer are responsible to implement the SIMD, bias addition and quantizer multiplication – limitation respectively. A simple control unit is used to realize the CFU protocol communication and synchronize units while a simple register is utilized to buffer results. To integrate the hardware accelerator functionality the following changes were made to original convolution function, in Algorithm 1:

1. Two initialization commands were implemented and used to transfer configuration parameters (offset, bias, min and max values) into the appropriate accelerator unit.
2. Multi-byte SIMD accumulation is done using two sequential accumulations commands whose operands are four-byte input and filter values.
3. Bias addition and quantization is realized through a simple read command which also transfers the result back to CPU.

Fig. 2 summarizes the impact of the hardware accelerator showing speed up improvement by a factor of 3.7x.

4 Experimental Results

Our experiments targeted the Arty A7 development board from Digilent [18]. This is primarily because of two reasons a) the availability of the specific board on the lab and b) the ability to measure the power consumption. The Arty board includes circuitry for monitoring the main (5V) and part (only 0.95V) of the FPGA core voltage supplies as well as the current consumed from those supplies. The first represents the total system (board level) power consumption, and the second represents the power consumption of the digital logic and block ram of our design in the FPGA core.

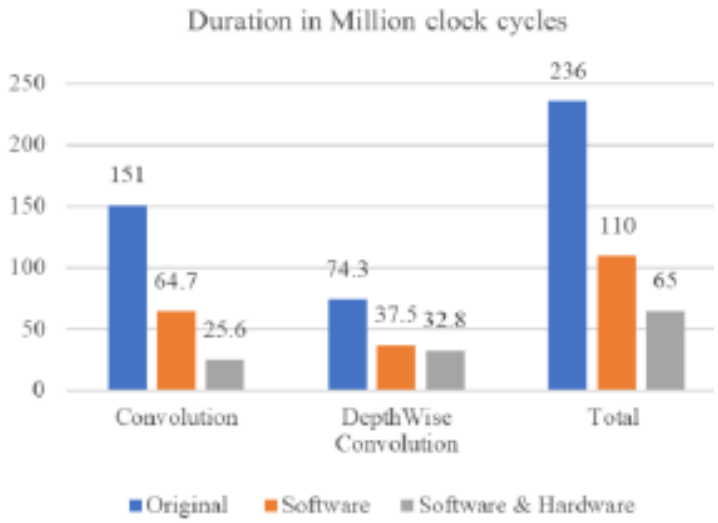


Figure 2: Speed Optimization

The experiment configuration includes the VexRiscv full variant with 8KB data and 8KB instruction cache, complemented with the compact accelerator presented in Section 3 and we measure speed (cycle count), area utilization and power consumption. The results are presented in Fig. 2, Fig. 3 and Fig. 4 respectively. When deemed necessary three discrete experiment configurations were executed for a) original code, b) software only optimizations and c) software and hardware optimizations.

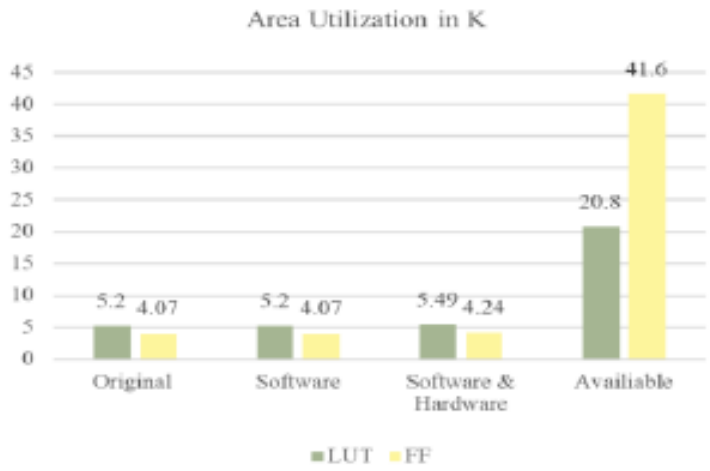


Figure 3: Area Utilization

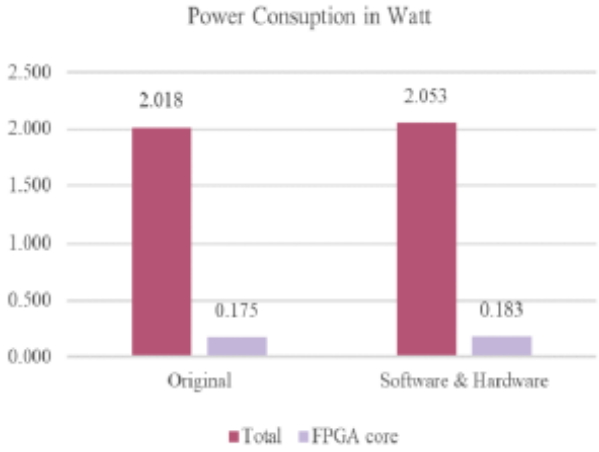


Figure 4: Power Consuption

The results show overall performance improvement by a factor of 3.7x in terms of execution speed. The time for one frame inference taking into consideration that our clock speed is 100MHz reduced from 2.5 seconds to 0.65 seconds. The design consumes 5.5K slice LUTs (26.41%) – 4.3K slice registers (10.19%) thus it can fit in most of the entry level FPGAs of all major manufactures (Intel, Xilinx, Lattice, etc.). The power consumption of the FPGA chip is about 180 mW, which corresponds to the maximum value measured over one hundred images dataset. This result is acceptable if we consider that, according to a thorough study on tinyML conducted on diverse applications and systems in [19], the typical power consumption is between 25mW and 200mW.

5 Conclusions

In this paper we presented a complete hardware-software RISC-V based ML accelerator design using open-source tools. We described a full design workflow, starting from the network selection and training of the model, carried on with the accelerator design methodology and deployment in a FPGA development platform. Our implementation achieved an overall speed increase by a factor of 3.7x compared to running inference natively without the software hardware optimizations available in CFU

illustrating that is feasible to achieve object detection at decent speed on low-resource and entry-level FPGA devices.

This is an exploratory work which is presented as an attempt to exploit the CFU framework on resource-constrained devices and therefore, further work is required to proliferate results. Finally, from the power consumption perspective, we believe that by eliminating the necessity of external memory, or replacing it with internal Block-RAM, will have a huge positive impact.

The source code is available in the form of a CFU Playground project at: https://github.com/ECSAlab/fomo-object-detection.

6 References

[1] *A Survey on RISC-V-Based Machine Learning Ecosystem*. S. Kalapothas, M. Galetakis, G. Flamis, F. Plessas, and P. Kitsos. 2, s.l. : Information, 2023, Vol. 14.

[2] FOMO. *EDGE IMPULSE*. [Online] https://docs.edgeimpulse.com/docs/edge-impulse-studio/learning-blocks/object-detection/fomo-object-detection-for-constrained-devices.

[3] The PASCAL Visual Object Classes. [Online] http://host.robots.ox.ac.uk/pascal/VOC.

[4] *CFU Playground: Full-Stack Open-Source Framework for Tiny Machine Learning (tinyML) Acceleration on FPGAs*. Prakash, S., et al. s.l. : IEEE ISPASS, 2023.

[5] *Best Practices for the Deployment of Edge Inference: The Conclusions to Start Designing*. Flamis, G., Kalapothas, S. and Kitsos, P. 16, s.l. : Electronics, 2021, Vol. 10.

[6] *A systematic literature review on hardware implementation of artificial intelligence algorithms*. Talib, M.A., Majzoub, S., Nasir, Q. s.l. : Supercomput, 2021.

[7] *Domain-specific hardware accelerators*. W.J. Dally, Y. Turakhia, and S. Han. 7, s.l. : Communications of the ACM, 2020, Vol. 63.

[8] *FPGA-Based Accelerators of Deep Learning Networks for Learning and Classification: A Review*. A. Shawahna, S. M. Sait and A. El-Maleh. s.l. : IEEE Access, 2019, Vol. 7.

[9] *RISC-V barrel processor for deep neural network acceleration*. M. AskariHemmat, O. Bilaniuk, S. Wagner, Y. Savaria, and J. P. David. s.l. : In 2021 IEEE International Symposium on Circuits and Systems (ISCAS), 2021.

[10] *Accelerated RISC-V for post-quantum SIKE*. Elkhatib, Rami, B. Koziel, R. Azarderakhsh, and M. M. Kermani. 6, s.l. : IEEE Transactions on Circuits and Systems, Vol. 69.

[11] *A RISC-V-based FPGA Overlay to Simplify Embedded Accelerator Deployment*. G. Bellocchi, A. Capotondi, F. Conti and A. Marongiu. s.l. : 24th Euromicro Conference on Digital System Design (DSD), 2021.

[12] *RISC-V2: A Scalable RISC-V Vector Processor*. K. Patsidis, C. Nicopoulos, G. C. Sirakoulis and G. Dimitrakopoulos. s.l. : IEEE International Symposium on Circuits and Systems (ISCAS), 2020.

[13] *Arnold: An eFPGA-Augmented RISC-V SoC for Flexible and Low-Power IoT End Nodes*. Pasquale Davide Schiavone, Davide Rossi, Alfio Di Mauro, Frank Gurkaynak, Timothy Saxe, Mao Wang, Ket Chong Yap, Luca Benini. 4, s.l. : IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2021, Vol. 29.

[14] *GAP-8: A RISC-V SoC for AI at the Edge of the IoT.* Flamand, E. s.l. : IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP), 2018.

[15] Kendryte home page. [Online] https://www.canaan-creative.com/product/kendryteai.

[16] *Benchmarking the MAX78000 artificial intelligence microcontroller for deep learning applications*. M. Clay, C. Grecos, M. Shirvaikar, and B. Richey. s.l. : Real-Time Image Processing and Deep Learning, 2022, Vol. 12102.

[17] CFU-Playground documentation. [Online] https://cfu-playground.readthedocs.io/en/latest/index.html.

[18] Artix-7. [Online] digilent. https://digilent.com/reference/programmable-logic/arty-a7/start.

[19] *A Comprehensive Survey on TinyML*. Y. Abadade, A. Temouden, H. Bamoumen, N. Benamar, Y. Chtouki and A. S. Hafid,. s.l. : IEEE Access, 2023, Vol. 11.

EMERGING TECH CONFERENCE 2023
Edge Intelligence

Posters

The Smart Fridge Project
Aikaterini Griva, Vasileios Rekkas, Kyriakos Koritsoglou, Sotirios Sotiroudis, Achilles Boursianis, Maria Papadopoulou and Sotirios Goudos

Ground-based cloud observation using wide-view optical and thermal representations
Dimitrios Tsourounis, Panagiotis Tzoumanikas, Alexios Kotronis, Andreas Kazantzidis, George Economou, Orestis Panagopoulos and Christos Theocharatos

An OTA-based power sensing integrated circuit for MPPT in photovoltaic energy harvesting applications
Zoi Agorastou, Vasileios Konstantakos, Konstantinos Siozios and Stylianos Siskos

Prediction of Analog Circuit Sizing Using an Artificial Neural Network
Shubham Agarwal, Benjamin Prautsch and Uwe Hatnik

Classification of Fetal Images using Deep Learning Methodologies: The Smart Embryo Project
Lazaros Alexios Iliadis, George Vergos, Paraskevi Kritopoulou, Achilleas Papatheodorou, Sotirios Sotiroudis, Achilles Boursianis, Kostas Kokkinidis, Maria Papadopoulou and Sotirios Goudos

Estimation of pedestrian trajectory using LSTM network architecture
Christos Theocharatos, Dimitris Kastaniotis and Vassilis Tsagaris

A deep learning approach for detecting defects in melt pool images in DED-AM process
Thanassis Polizogopoulos, Christos Theocharatos, Konstantinos Tzimanis, Nikolas Porevopoulos, Panagiotis Stavropoulos, Konstantinos Ntalas and Albano Memlika

Machine learning methods for the discrimination of refrigerant gases
Nikolaos Argirusis, Petros Karvelis, Georgia Sourkouni, John Konstantaras, Alice Baroncelli, Peter Segers and Christos Argirusis

HETiA
EXPANDING THE DIGITAL FRONTIERS

ARISTOTLE UNIVERSITY OF THESSALONIKI

EMERGING TECH CONFERENCE 2023
Edge Intelligence

Design of an Autonomous, Multi-functional Stress Assessment Sensor for Naval Applications: The AMSA project
Gregory Doumenis, Vasiliki Naskari, Evangelos Hristoforou, Polychronis Pattakos, Georgia Stamou, Christos Papakis, Ioannis Masklavanos

A unified framework for object and person 3D-pose estimation with application in ancient drama
Andreas Makedonas, Ioannis Papakonstantinopoulos, Panteleimon Alkinoos Peftikoglou and Christos Theocharatos

Deep Learning Architectures for Greek Orthodox Church Hymns Recognition: The Ymnodos Project
Lazaros Alexios Iliadis, Nikolaos Tsakatanis, Sotirios Sotiroudis, Achilles Boursianis, Kostas Kokkinidis, Georgios Patronas, Pavlos Serafeim, Maria Papadopoulou and Sotirios Goudos

An IoT System for Innovative Cultural Experience
Christos Sad, Aggeliki Ziaka and Kostas Siozios

An Improved Algorithm for Equalization and Energy Support for Lithium-ion Battery Storage Systems in Electric Motor Drive Applications
Nikolaos Jabbour, Evangelos Tsioumas, Dimitrios Papagiannis and Christos Mademlis

Design of an Autonomous Wireless Electric Field sensor for maritime applications: the EFOS project
Ioannis Masklavanos, Vasiliki Naskari, Christos Koutsos, Fotios Vartziotis, Gregory Doumenis, Stylianos Siskos, Achilleas Bardakas, Apostolos Segkos, Christos Tsamis, Christos Papakis and George Koukas

Area Allocation for Coverage Path Planning Using Affinity Propagation Clustering
Nikolaos Baras, Antonios Chatzisavvas, Irene Tabakis and Minas Dasygeni

Crosstalk exploration between PA and LNA inductors
Dimitrios Samaras, Alkis Hatzopoulos, Vasilis Pavlidis, Athanasios Stefanou, Georgios Chararas and Rafaela Themeli

Auditory Scene Profile Adaptation for ANC Headphones
Dennis Tsoukalas and Fotios Kontomichos

HETiA
EXPANDING THE DIGITAL FRONTIERS

ARISTOTLE UNIVERSITY OF THESSALONIKI

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 166 – 167

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

The Smart Fridge Project

Aikaterini I. Griva[1†††††] , Vasileios P. Rekkas[1], Kyriakos Koritsoglou[2], Sotirios P. Sotiroudis[1], Achilles D. Boursianis[1], Maria S. Papadopoulou[3], and Sotirios K. Goudos[1‡‡‡‡‡]

[1] ELEDIA@AUTh, School of Physics Aristotle University of Thessaloniki, Thessaloniki, Greece
[2] Department of Information and Telecommunications, University of Ioannina, Arta, Greece
[3] Department of Information and Electronic Engineering International Hellenic University Sindos, Greece

{aigriva, vrekkas, ssoti, bachi, sgoudo}@physics.auth.gr, kkoritsoglou@uoi.gr, mspapa@ihu.gr

Abstract

In this paper a technical report of the Smart Fridge project is presented. The Smart Fridge project (Project code: KMP6-0078831) combines Internet of Things technologies, wireless sensor networking, and artificial intelligence to reduce the energy footprint of companies involved in the retail food industry.

1 Introduction

Energy consumption is the key to both environmental and economic sustainability. Many academics and enterprises are working together to increase system performance while reducing energy usage. AUTH and SEEMS proposed the design and development of an innovative system that will fully comply with the legal framework and will be able to monitor, record, storage, manage, and automate regular the temperature values of a refrigerator chamber.

2 Results

Preliminary studies related to the Smart Fridge project were carried out in 2022. An IoT-based technology device was proposed to reduce the power consumption of refrigeration equipment. A DS18B20 digital temperature sensor was connected to a Raspberry Pi Zero W. The goal was to map the operation of the system and provide temperature control methods to minimize the power consumption (Koritsoglou et al, 2022).

In 2023 a simple refrigeration model was simulated on Simulink, Mathworks to evaluate the energy consumption of the model by adjusting the indoor temperature, the indoor moisture, and the interior temperature of the system (Griva et al, 2023). The results show that the indoor moisture has a small effect on the energy consumption of the fridge. However, the valid configuration of the temperature

†††††† Writing and original draft preparation
‡‡‡‡‡ Review and editing

(indoor and interior) holds great importance as shown in Fig. 1.

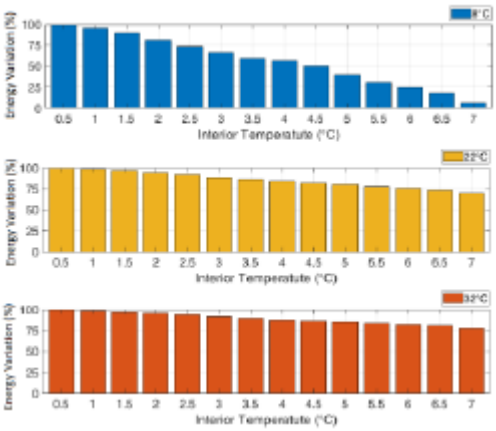


Figure 1: Quantitative change of the interior temperature in three indoor temperatures.

3 Conclusion

The main objective of the Smart Fridge project is to decrease the energy consumption of refrigeration equipment by controlling the operating parameters. An IoT-based preliminary model and a simulated approach have been examined thus far. Moreover, the overall system is currently being implemented. In the next few months, a machine learning model will be provided based on the dataset that will be developed from measurements in food retail stores in Thessaloniki, Greece. This model will map the desired period times in which the system can operate with the lower energy consumption. Finally, the results of the three approaches (simulation, experiment, and machine learning) will be evaluated and combined to suggest the optimal operation based on the energy consumption of refrigeration equipment, that will be important for environmental and economic sustainability.

4 Acknowledgment

This research was carried out as part of the project “Smart Fridge Using IoT Technology” (Project code: KMP6-0078831) under the framework of the Action “Investment Plans of Innovation” of the Operational Program “Central Macedonia 2014-2020”, that is co-funded by the European Regional Development Fund and Greece.

5 References

[1] Koritsoglou, K., Papadopoulou, M. S., Boursianis, A. D., Sarigiannidis, P., Nikolaidis, S., & Goudos, S. K. (2022, June). Smart Refrigeration Equipment based on IoT Technology for Reducing Power Consumption. In 2022 11th International Conference on Modern Circuits and Systems Technologies (MOCAST) (pp. 1-4). IEEE.

[2] Griva, A.I., Rekkas, V. P., Koritsoglou, K., Sotiroudis, S. P., Boursianis, A. D., Papadopoulou, M. S., & Goudos, S. K. (2023, June). Energy Consumption Assessment in Refrigeration Equipment: The SmartFridge Project. In 2023 12th International Conference on Modern Circuits and Systems Technologies (MOCAST) IEEE.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 168 – 174

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Ground-based cloud observation using wide-view optical and thermal representations

Dimitrios Tsourounis[1], Panagiotis Tzoumanikas[2], Alexios Kotronis[3], Orestis Panagopoulos[2], Andreas Kazantzidis[2], George Economou[1], and Christos Theocharatos[3§§§§§]

[1] *Electronics Laboratory, Physics Dept., University of Patras, Rio Patras 26504, Greece,*
[2] *Laboratory of Atmospheric Physics, Physics Dept.,University of Patras,Rio Patras 26504 Greece,*
[3] *Irida Labs, Patras InnoHub – Kastritsiou 4, Magoula Patras 26504, Greece*

dtsourounis@upatras.gr, tzumanik@ceid.upatras.gr, up1066667@upatras.gr, orestis.panagopou@upatras.gr, akaza@upatras.gr, economou@upatras.gr, htheohar@iridalabs.com

Abstract

A ground-based cloud observation system was developed in the context of the DeepSky project that utilizes optical and thermal cameras to capture wide-view images of the sky. A large dataset is created consisting of patch sky images in two modalities along with corresponding cloud category labels. Convolutional neural network models were trained using this dataset to classify the images into the cloud categories. Results highlight the complementary nature of the optical and thermal images in the task of cloud type classification.

1 Introduction

Clouds are collections of water droplets or ice crystals, and they form through the condensation of water vapor in the atmosphere, as it rises and cools, attaching to particles like dust, ice or sea salt and becoming visible [Liou1992]. Hence, there are various types of clouds according to multiple natural factors affecting their formation. The clouds significantly influence Earth's weather patterns, impacting the hydrological cycle, energy balance, and climate system. Consequently, accurate cloud observation is crucial for weather monitoring and relevant applications (e.g., air traffic control) as well as climate assessment or projections. Observation methods include space-based satellites, air-based radiosondes, and ground-based remote sensing. Satellite observations provide broad coverage, but they often lack the necessary temporal and spatial resolutions for localized and short-term cloud analysis. Air-based radiosonde observations excel in detecting cloud vertical structures however they are costly. In response, ground-based remote sensing technologies offer cost-effective solutions with high-resolution images for detailed analysis of local cloud characteristics [Kazantzidis2012].

2 System

Our ground-based cloud observation system is equipped with two camera sensors, an optical camera and a Thermal-InfraRed (TIR) camera. The optical camera, also known as an Red Green Blue (RGB)

§§§§§ Masterminded EasyChair and created the first stable version of this document

camera, captures images using visible light (390 – 770 nm), similar to how our eyes perceive the world. It measures the intensity of red, green, and blue light to create a color image. On the other hand, the infrared camera, also referred to as a thermal camera, detects infrared radiation (8 – 14 μm) emitted by the atmosphere and clouds. It measures the heat emitted by different objects and converts it into an image displaying variations in temperature. The spectral band of 8–14 μm is known as the Long Wave InfraRed (LWIR), which is an ideal spectral range for cloud observation. In this spectral window, the atmosphere has low emission and high transmittance while clouds emit strong infrared radiation. Clouds could be considered as black-bodies since the optically thick clouds' emission is similar to a black-body at or near the temperature of the clouds [Shaw2012].

Both cameras have their advantages and limitations. The optical camera relies on visible light, which means it is influenced by factors like sunlight and atmospheric haze, affecting the pixel values. Furthermore, it operates effectively only during daylight hours. Nonetheless, the optical camera can capture texture patterns of clouds that are derived from sunlight and cloud thickness. On the other hand, the saturation problem of circumsolar region is avoided when images captured from thermal cameras. Moreover, thermal representations of clouds may lack detailed texture information, especially in cases where large areas of the clouds have similar temperatures. However, these images are affected less by the atmospheric aerosols while the most important advantage is the capture of images under various air conditions during day and night.

The use of a TIR microbolometer in thermal cameras necessitates a germanium lens, increasing the camera's cost significantly. This cost escalation is due to the necessity of custom-made lenses to achieve a wide Field of View (FoV) and high resolution [Klebe2014]. In the work of Wang et al. [Wang2021], a self-made dual system, comprising TIR and optical all-sky view cameras, is a noteworthy example. While this system provides all sky imagery and high resolution, the number of collected images is relatively small covering a few months recordings and thus, its cloud recognition approach relies on meteorological models that are regulated by local environmental conditions. Some other approaches have employed multiple cameras and scanning operation in different zenith directions [Sun2011] or utilized spherical aluminum mirrors to reflect the sky [Aebi2018] however, these methods introduce many challenging installation issues, such as the need for motoring mechanism, combining multiple captured images into a unified image, cleaning the mirror, and determining the appropriate distance between camera and mirror.

The DeepSky dual camera system we employed consists of an optical and a thermal sensor, both offering a wide-angle FoV and capturing registered images with a small scaling factor between them. This system is based on the Mobotix M73 cameras configuration. More specifically, the optical camera is equipped with an IR cut filter, enabling its use in small luminance conditions. The thermal sensor operates in the infrared range of 7.5 to 13.5 μm and exhibits a Noise-Equivalent Temperature Difference (NETD) sensitivity of 50 mK. This thermal camera is enhanced with integrated Thermal Radiometry technology and a high-end thermal image sensor that is calibrated to measure thermal radiation across the entire image area and assign a temperature value per pixel. A summary of the key specifications of this dual camera system is presented in the following Table 1. Furthermore, our novel recognition system, designed for image processing, is firmly rooted in the principles of deep learning, leveraging a vast number of annotated images from both modalities for robust performance.

Specs of Dual camera system	Optical Camera	Thermal Camera
Resolution (H × V in pixels)	640 × 480	640 × 480
Field of View (FoV) (H × V in degrees)	95° × 50°	90° × 69°
Focal Length	5mm	5mm
Aperture	f/1.8"	-
Operational Temperature	-40° to 60° C	
Protection	IP66, IK07	

Table 1: Key specifications of DeepSky dual camera system

3 Dataset

The DeepSky system captures Sky Patch Images (SPI) using wide-angle lenses for both the optical and thermal cameras. As a result, the images cover the central portion of the sky rather than entire sky, as our dual camera system is positioned to look directly overhead (zenith). This is in contrast to Total Sky Images (TSI), which use a hemispherical chrome-plated mirror to reflect the sky onto a downward-pointing camera located above the mirror, and All Sky Images (ASI) which are typically obtained using a camera equipped with a fish-eye lens [Nie2022]. The images captured by the DeepSky system, located at the Physics Department of the University of Patras (38° 17' 29" N, 21° 47' 20" E), and covers the period between 2022 and 2023. The images are labeled with one of the five cloud categories: cumulus, altocumulus, cirrus, clear sky, and stratocumulus-cumulonimbus, based on the classification recommended by the World Meteorological Organization (WMO). The classification of the images was performed by professional human observers from the Laboratory of Atmospheric Physics. Since both optical and thermal images are captured simultaneously, the cloud labels are associated with both image types. Therefore, images for daytime data of thermal camera are used because only daytime images are available for the visible camera and human annotation. To ensure balanced training data, an equal number of images was selected from 2022 for training. This was achieved by randomly choosing images according to the number of the smallest class, ensuring that each class had the same number of training samples. For the test set, all available images from 2023 were used. However, since the clear sky class was overrepresented, a subset of

Cloud types - Classes	Description	Training (during 2022)	Test (during 2023)
I) Cumulus	Low clouds, Fluffy and puffy clouds with a distinct dome-shaped appearance	278	82
II) Altocumulus	Middle clouds, Patched clouds with mosaic-like appearance	259	87
III) Cirrus	High clouds, Thin, wispy, and fibrous clouds with a feathery or filamentous appearance	275	70
IV) Clear sky	Cloudless sky or a very few cloudiness	286	426
V) Stratocumulus-Cumulonimbus	Low clouds, Thick and lumpy clouds with almost to mostly overcast	223	93
Total		1321	758

Table 2: Information about the DeepSky dual representation dataset

randomly selected clear sky images was reduced in order to maintain a more balanced distribution across all classes. Table 2 provides a description of each cloud category along with the number of images available for each category. It's important to note that the total number of images in the dataset is doubled due to the inclusion of both optical and thermal images. This dataset consists of SPI with two modalities, optical and thermal, and includes the corresponding cloud category labels. Figure 1 shows some examples of the images in the dataset.

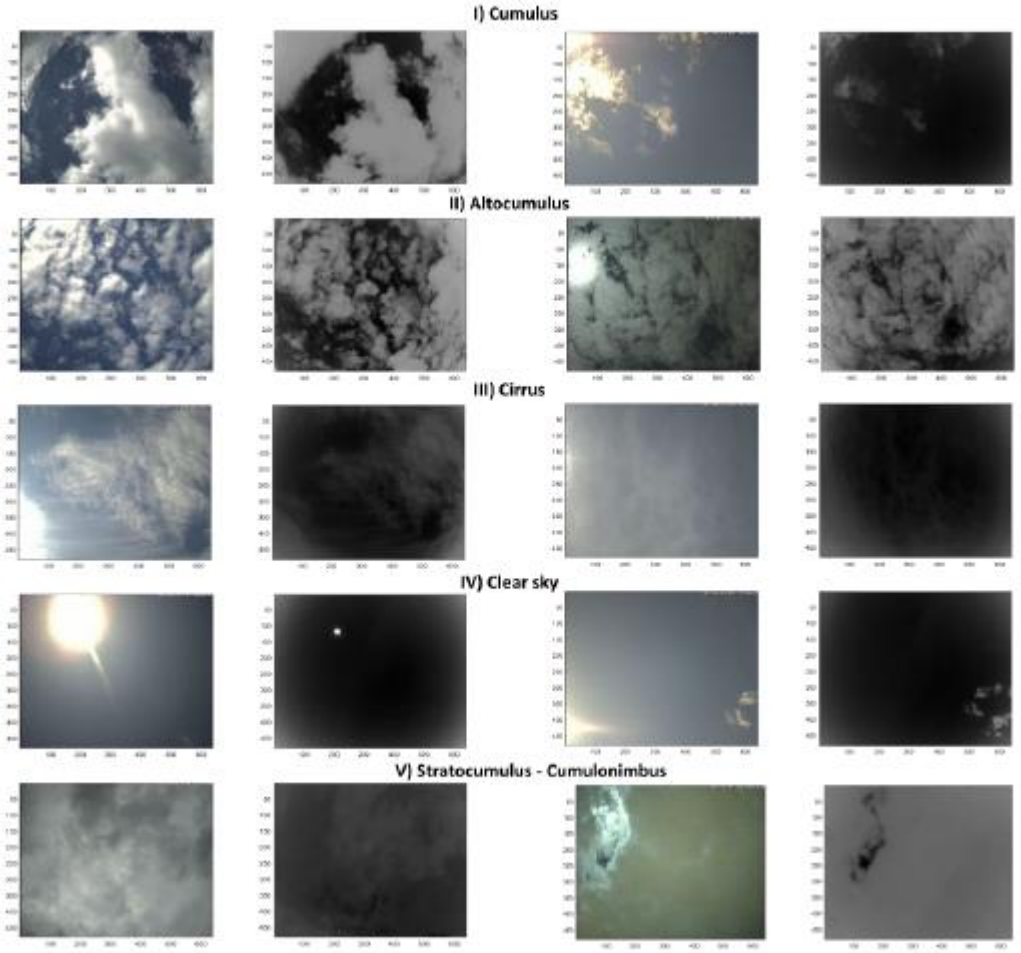


Figure 1: Examples of image pairs from the dual dataset with two modalities: optical (RGB images) and thermal (images with color scaling using a grayscale colormap). We display ten image pairs representing five cloud categories, with the RGB image on the left and the corresponding thermal image on the right.

4 Results

We employed a Convolutional Neural Network (CNN), specifically the ResNet-50 [He2016], to analyze the classification performance. Two separate CNN models were trained and evaluated using the optical and thermal images respectively. The images were resized to 224 x 224 pixels to reduce the computational cost. Also, the images are augmented by random horizontal and vertical flips during training. To ensure a coherent temperature range, the thermal images underwent pre-processing to limit temperature range between -80 and 25 degrees Celsius, where values exceeding 25 degrees

were set to 25. This pre-processing step was necessary to eliminate temperature values that could be influenced by external factors such as a portion of sunlight or objects, like airplanes, ensuring a more reliable and accurate analysis of the thermal data. The histograms of temperature values for the various cloud categories, generated using the training thermal images, provide a clear justification of this process, as shown in Figure 2. All images were normalized using mean subtraction and standard deviation division, calculated from the training data, before their input to CNN models.

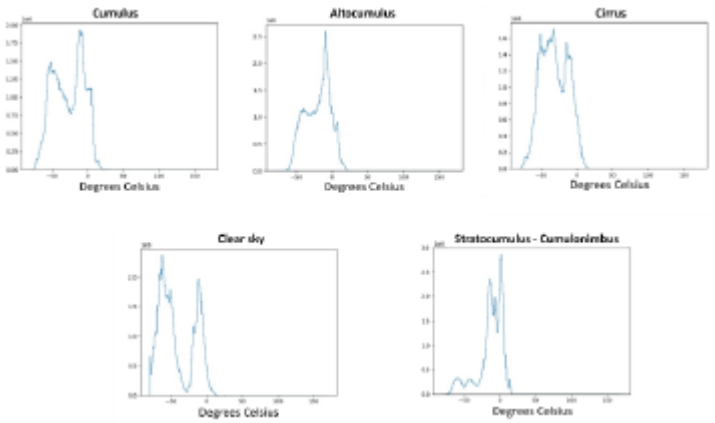


Figure 2: Histogram of temperature values in the thermal images for the five different cloud types presented in the training set of the dataset.

Optimization was performed by minimizing the cross-entropy classification loss with the RMSprop optimizer, employing a mini-batch size of 16 and an initial learning rate of 10-5 for 50 epochs. The classification results are presented using the confusion matrices in Figure 3 for the optical and thermal settings respectively. Although both CNN models achieved an average per-class classification accuracy of 76%, it is evident that optical and thermal representations have a different impact on the results. Specifically, the models performed similarly on cloudless sky (clear sky), middle clouds (altocumulus), and high clouds (cirrus). However, the classification accuracy for low clouds (cumulus and stratocumulus-cumulonimbus) differed between the two modalities. The model operating on optical images exhibited good performance in distinguishing cumulus clouds while the model operating on thermal images showed good performance in distinguishing stratocumulus-cumulonimbus clouds. This can be attributed to the information captured by each representation. The first model captured the texture and distinct features associated with cumulus clouds, such as clearly defined edges and white or light-gray color, whereas the second model detected the smaller temperature range (-30°C to 10°C) specific to stratocumulus-cumulonimbus clouds contrasting with the larger temperature range (-60°C to 10°C) observed in cumulus clouds.

5 Conclusions

In conclusion, a ground-based cloud observation system is developed incorporating optical and thermal cameras, enabling the capture of wide-view images of the sky. By utilizing this system, a substantial dataset comprising patch sky images in two modalities, along with associated cloud category labels, was created. Through the training of CNN models using this dataset, it was observed that the optical and thermal images complement each other in the classification of cloud types. This finding underscores the significance of utilizing multiple modalities for accurate and comprehensive cloud type classification. The integration of optical and thermal data enhances the overall

understanding of cloud formations and contributes to advancements in cloud observation and analysis. Future plans include the fusion of both modalities into the training of a CNN as well as the extension of the dataset using a motor mechanism that moves the cameras into some predefined positions to capture complete view of the entire sky.

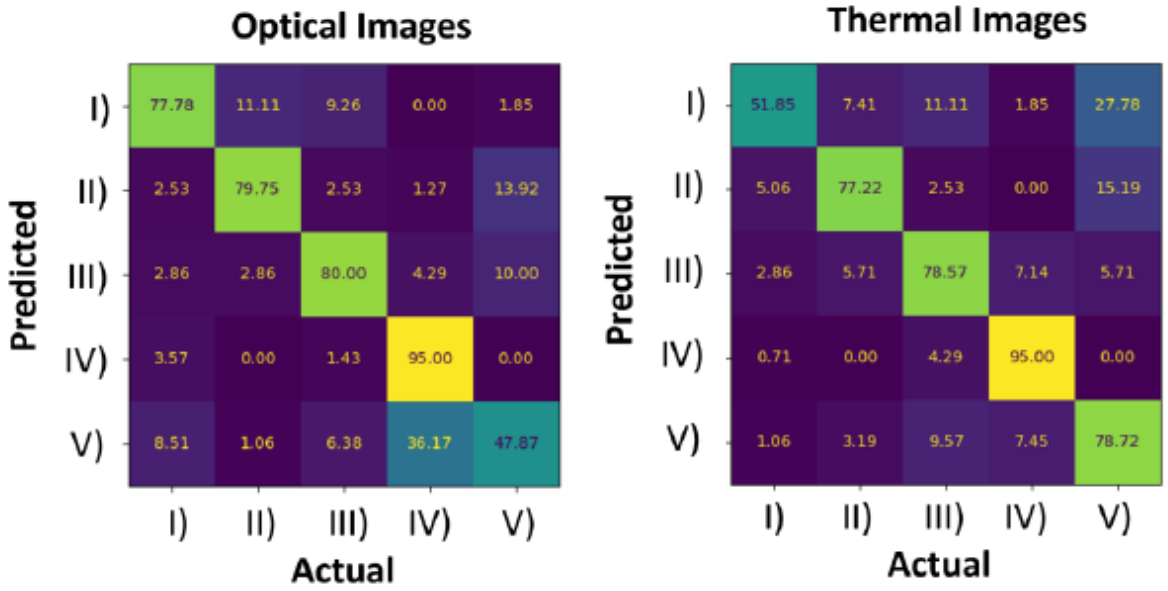


Figure 3: Confusion matrices using the Optical images (Left) and the Thermal images (Right).

6 Acknowledgments

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T1EDK – 00681, MIS 5067617).

7 References

[1] Aebi, C., Gröbner, J., & Kämpfer, N. (2018). Cloud fraction determined by thermal infrared and visible all-sky cameras. *Atmospheric Measurement Techniques*, 11(10), 5549-5563.

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[3] Kazantzidis, A., Tzoumanikas, P., Bais, A. F., Fotopoulos, S., & Economou, G. (2012). Cloud detection and classification with the use of whole-sky ground-based images. *Atmospheric Research*, 113, 80-88.

[4] Klebe, D. I., Blatherwick, R. D., & Morris, V. R. (2014). Ground-based all-sky mid-infrared and visible imagery for purposes of characterizing cloud properties. *Atmospheric Measurement Techniques*, 7(2), 637-645.

[5] Liou, K. N. (1992). Radiation and cloud processes in the atmosphere. Theory, observation, and modeling.

[6] Nie, Y., Li, X., Paletta, Q., Aragon, M., Scott, A., & Brandt, A. (2022). Open-Source Ground-based Sky Image Datasets for Very Short-term Solar Forecasting, Cloud Analysis and Modeling: A Comprehensive Survey. arXiv preprint arXiv:2211.14709.

[7] Shaw, J. A., & Nugent, P. W. (2013). Physics principles in radiometric infrared imaging of clouds in the atmosphere. *European Journal of Physics*, 34(6), S111.

[8] Sun, X., Liu, L., & Zhao, S. (2011). Whole sky infrared remote sensing of cloud. *Procedia Earth and Planetary Science*, 2, 278-283.

[9] Wang, Y., Liu, D., Xie, W., Yang, M., Gao, Z., Ling, X., ... & Xia, Y. (2021). Day and night clouds detection using a thermal-infrared all-sky-view camera. *Remote Sensing*, 13(9), 1852.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 175 – 181

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

An OTA-based power sensing integrated circuit for MPPT in photovoltaic energy harvesting applications

Zoi Agorastou[1], Vasileios Konstantakos[1], Konstantinos Siozios[1] and Stylianos Siskos[1]

[1] *Electronics Lab., Physics Dept., Aristotle University of Thessaloniki, Thessaloniki, Greece*

zagorast@physics.auth.gr, bkons@physics.auth.gr, ksiop@physics.auth.gr, siskos@physics.auth.gr

Abstract

In this paper, the design of a power sensing circuit based on the Operational Transconductance Amplifier (OTA) for photovoltaic energy harvesting applications is presented. The circuit extracts a series of pulses that have a frequency proportional to the measured power and can communicate with a digital processing unit that implements a Maximum Power Point Tracking (MPPT) algorithm. It can accommodate an input voltage range of 0.3-1 V and an adjustable input current range while consuming approximately 360 μW at a 3 V supply voltage. The circuit is suitable for integrated energy harvesting systems that include an MPPT unit, and it is compatible with photovoltaic energy harvesting systems.

1 Introduction

Maximum Power Point Tracking (MPPT) is a widely used technique in energy harvesting systems that harness the ambient available energy from the environment. By employing MPPT, an energy harvesting interface aims to extract the utmost possible energy from a source under varying conditions, and this is typically achieved by adjusting the load's electrical characteristics (current and voltage) to satisfy the Maximum Power Point conditions. More specifically, MPPT in photovoltaic systems, where a solar cell or panel converts the energy from solar irradiance to electrical energy, typically involves the measurement of the solar panel's output voltage and current to calculate the corresponding output power and the adjustment of the duty cycle of a DC-DC converter that is interposed between the transducer and the load, so that the maximum power is extracted from the former to be provided to the latter under varying temperature, solar irradiance, and shading conditions.

The prevalent MPPT algorithmic methods are Perturb and Observe (P&O) and Incremental Conductance [1-2] where the power or its derivative is calculated, and its behavior is studied for variations of the operating voltage or current, so that the MPP is tracked. By always operating at or around the MPP, the overall efficiency and performance of the system are enhanced, and the cost-effectiveness of the photovoltaic systems is improved. Both algorithms require knowledge of the behavior of the power in real-time to conduct the appropriate adjustments on the load characteristics. Typically, the voltage and the current are sensed separately, and the power is estimated by an analog or digital multiplier [3]. While voltage sensing is generally a non-complex

process, current sensing [4-6] involves the employment of a shunt resistor and the measurement of the voltage drop across the resistor, which is proportional to the current that flows through it. The resistor's value sets the sensitivity and accuracy of the current sensing circuit; therefore, a larger value would result in higher sensing efficiency but also in higher power losses. In addition, analog multiplication suffers from linearity and accuracy issues and limited dynamic range, whereas digital multiplication increases computational complexity as well as processing power demands.

In this paper an integrated analog circuit that provides an indication on the variations of the output power of a solar panel, which can be utilized by an MPPT unit, is presented. The circuit extracts a digital signal in the form of a series of pulses with a frequency proportional to the power of the solar panel, while consuming little power and occupying a small silicon area.

2 Description of the MPPT interface

A conventional MPPT interface for a photovoltaic panel consists of a DC-DC converter that is responsible for the MPPT process, the control unit of the DC-DC converter that typically conducts the power estimation and adjusts the duty cycle of the control signals of the converter, and the load.

In Figure 1 the proposed MPPT interface is depicted along with the positioning of the power sensing circuit on the DC-DC boost converter and its communication with the digital MPPT unit. The power estimation is conducted before the digital processing, which in this case is responsible only for the creation of the control signals of the converter.

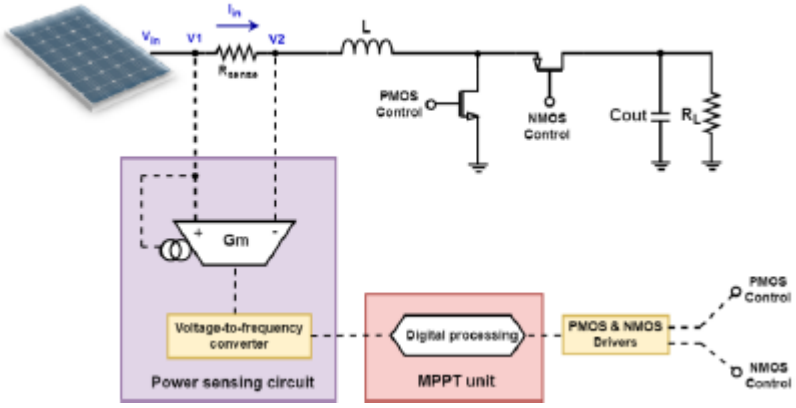


Figure 1. Proposed MPPT interface block diagram

The inputs of the power sensing circuit are the terminals of a shunt resistance that is inserted between the PV panel and the input inductor of the DC-DC converter. The voltage difference of the two terminals *ΔV=V1-V2* is proportional to the input current of the DC-DC converter, I_{in} according to:

$$\Delta V = I_{in} R_{sense} \quad (1)$$

and it is converted into a current by an Operational Transconductance Amplifier (OTA).

The input voltage is converted into a proportional current which is used as a bias current of the OTA. Since the output current of the OTA is proportional to both the voltage difference at its inputs and its transconductance, which in certain operating conditions is proportional to its bias current, the output current of the OTA is proportional to the output power of the PV panel. A current sensing and a power sensing circuit implementation using the OTA circuit were presented in [7] and [8], respectively,

however the proposed circuit provides a less complex as well as more area-efficient solution. In Section 3 a more thorough mathematical analysis is presented.

3 Proposed power sensing circuit

Figure 2 depicts the proposed power sensing circuit which consists of a transconductor (Figure 2a) that converts the input voltage of the DC-DC converter (output voltage of the PV panel) into a proportional current, according to:

$$I_b = \frac{V1}{R_c} \tag{2}$$

The core of the power sensing circuit is the OTA circuit (Figure 2b), which converts the voltage difference at its inputs into a current, according to:

$$I_{OTA} = G_m(V1 - V2) \tag{3}$$

where G_m is the amplifier's transconductance. According to [9] the transconductance of the OTA has a linear dependence on the bias current when the amplifier operates at the weak inversion or with low currents. Thus, ideally:

$$I_{OTA} = kI_b(V1 - V2) \tag{4}$$

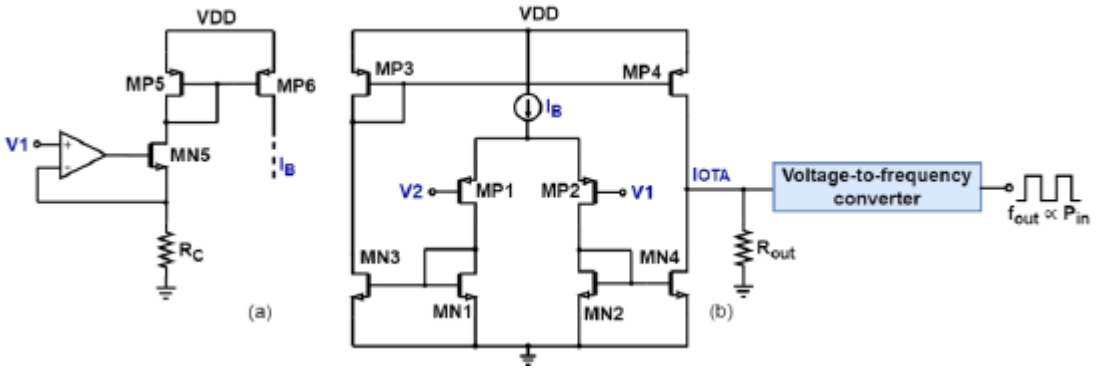


Figure 2. Proposed power sensing circuit (a) Transconductor (b) OTA and voltage-to frequency converter

And from (2) and (4):

$$I_{OTA} = k\frac{V1}{R_C}(V1 - V2) \tag{5}$$

Since *V1* is the input voltage ($V1 \equiv Vin$) and the voltage difference ΔV is proportional to the input current, Iin:

$$I_{OTA} \propto V_{in}\, x\, I_{in} \propto P_{in} \tag{6}$$

The OTA's output voltage, V_{out} which is inserted to the voltage-to-frequency converter's input, is the product of its output current and the load resistance, R_{out}:

$$V_{out} = I_{OTA}\, x\, R_{out} \tag{7}$$

Finally, the frequency of the pulse series at the output of the voltage-to-frequency converter (Figure

3) – which is based on [10] - is proportional to the voltage at its input, which is V_{out}, therefore:

$$f_{out} = mV_{out} \propto I_{OTA} \propto P_{in} \quad (8)$$

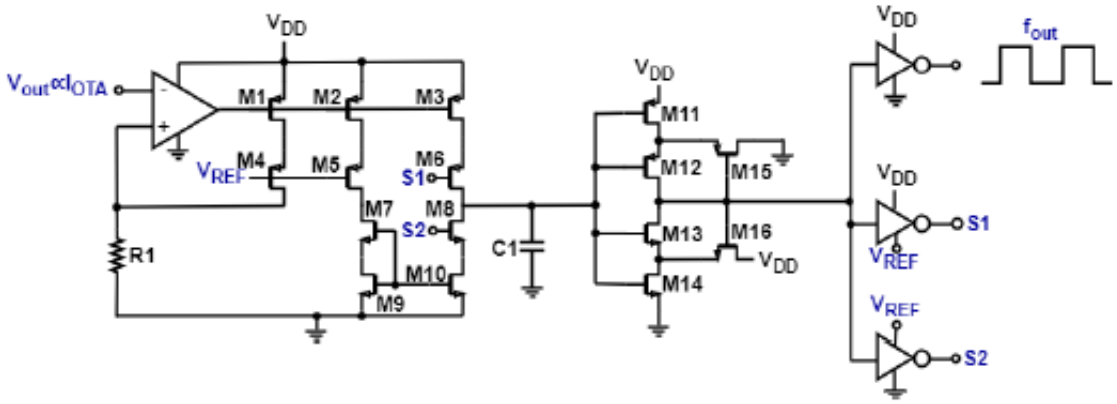


Figure 3. Voltage-to-frequency converter

Figure 1: Examples of image pairs from the dual dataset with two modalities: optical (RGB images) and thermal (images with color scaling using a grayscale colormap). We display ten image pairs representing five cloud categories, with the RGB image on the left and the corresponding thermal image on the right.

Table 1 includes the component values of the power sensing circuit and the voltage to frequency converter and Figure 4 depicts the layout of the total power sensing circuit.

Components	W/L (μm/μm)	Components	W/L (μm/μm)
MP1-MP2	10/8	M1-M10	40/2
MN1-MN2	20/2	M11-M12	10/1
MN3-MN4	20/2	M12-M13	10/3
MN5	10/0.35	M15	15/1
MP3-MP4	20/2	M16	2/1
MP5-MP6	10/4	C1	73.8pF
R_C	1MΩ	R1	2.4MΩ
R_{out}	2MΩ		

Table 1. Component values of the power sensing circuit and the voltage-to-frequency converter

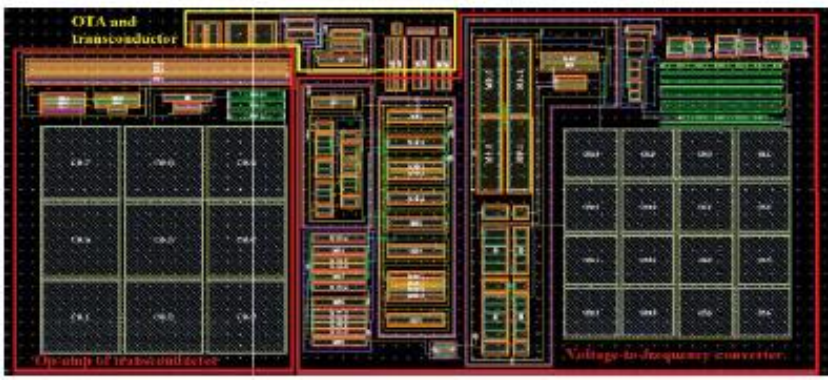


Figure 4. Physical design (layout) of the power sensing circuit (303 μm × 144 μm)

4 Post-layout simulation results

The specifications of the energy harvesting module appear in Table 2. The input voltage range can be adjusted by varying the value of RC to maintain proper bias conditions of the OTA circuit (~0.2-1 μA) and the input current range can be adjusted by varying the value of the shunt resistor, R_{sense}, so that the voltage difference remains inside the measurable range of 0-100 mV. However, R_{sense} contributes to the power losses, so its value should be carefully selected in relation to the input current range.

Technology	180 nm CMOS (XH018)
Input voltage	0.3-1 V (adjustable)
Input voltage difference	0-100 mV (fixed)
Input current @ Rsense=100 mΩ	0-1 A (adjustable)
fout (kHz)	2.55-14.11 kHz
Power consumption @ 3V	0.36 mW

Table 2. Specifications of the power sensing circuit

The diagrams in Figure 5 depict the plots of the output frequency versus the voltage V1 for various voltage differences ΔV (Figure 5a) and the output frequency versus the voltage difference ΔV for various V1 values. In Figure 5c the post-layout results of the output frequency plotted against the product of the voltage difference $\Delta V \propto I_{in}$ and the voltage $V1 \equiv V_{in}$ are presented. The frequency is linearly dependent on this product and therefore on the input power of the DC-DC converter.

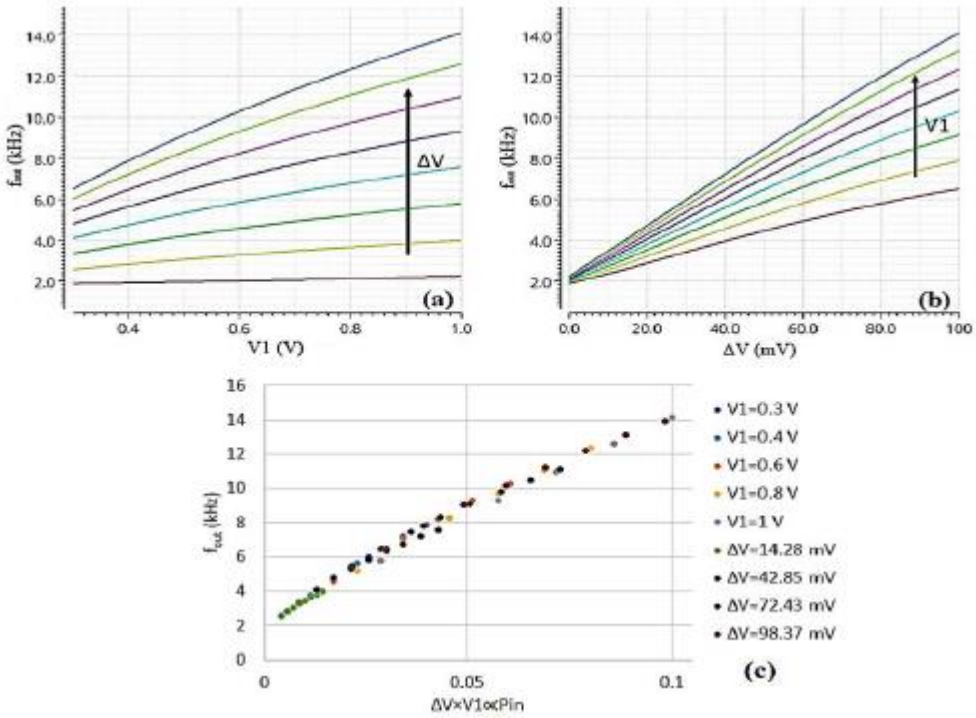


Figure 5. Output frequency vs. (a) V1, (b) ΔV and (c) the product of ΔV×V1

The derived straight-line equation is:

$$f_{out} = 123.08\ x\ n\ x\ \ P_{in}\ +\ 2.5312\ kHz,\quad R^2 =\ 0.989$$

which indicates that the output frequency is linearly dependent on the input power.

5 Conclusions

An integrated power sensing circuit was designed and simulated in XH018 CMOS technology using Cadence software. The power sensing circuit utilizes the linear relationship between an OTA's output current and the voltage difference on its inputs, as well as the linear relationship between an OTA's transconductance and its bias current when the amplifier operates in the weak inversion or low current region. The extracted voltage at the load of the OTA circuit is converted into a series of pulses with a frequency proportional to the input power of the DC-DC converter, allowing a digital MPPT unit to detect its behavior and adjust the converter's duty cycle accordingly. The final circuit occupies 0.0436 mm2 and consumes 360 μW, which renders it feasible to be integrated in an energy harvesting system.

6 Acknowledgments

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH—CREATE—INNOVATE (project code: T2EΔK-01681).

7 References

[1] A. Ali et al., "Investigation of MPPT Techniques Under Uniform and Non-Uniform Solar Irradiation Condition–A Retrospection," IEEE Access, vol. 8, pp. 127368–127392, Jan. 2020, https://doi.org/10.1109/access.2020.3007710

[2] S. H. Hanzaei, S. A. Gorji, and M. Ektesabi, "A Scheme-Based Review of MPPT Techniques with Respect to Input Variables Including Solar Irradiance and PV Arrays' Temperature," IEEE Access, vol. 8, pp. 182229–182239, Oct. 2020, https://doi.org/10.1109/access.2020.3028580

[3] S. Singh, D. Mandal, B. Bakkaloglu and S. Kiaei, "Low-Power/Low-Voltage Integrated CMOS Sense Resistor-Free Analog Power/Current Sensor Compatible With High-Voltage Switching DC–DC Converter," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 66, no. 6, pp. 2208-2218, June 2019, https://doi.org/10.1109/TCSI.2019.2897049

[4] V. Kalenteridis, Z. Agorastou and S. Siskos, "A novel current sensing technique for photovoltaic MPPT applications," 2022 11th International Conference on Modern Circuits and Systems Technologies (MOCAST), Bremen, Germany, 2022, pp. 1-4, https://doi.org/10.1109/MOCAST54814.2022.9837565

[5] C. -W. Hsu, K. -P. Chiu and J. -F. Chen, "The Vds signal of power switch for boost PFC current detection," IEEE 4th International Future Energy Electronics Conference (IFEEC), pp. 1-6, 2019.

[6] B. Li, K. S. Low, and B. S. Kang, An accurate lossless current sensing approach for a DC-DC converter with online calibration. 2014. https://doi.org/10.1109/appeec.2014.7066050

[7] J. -J. Chen, Y. -S. Hwang, J. -H. Wu, C. -H. Lai and Y. -T. Ku, "A New Improved V-Square-Controlled Buck Converter With Rail-to-Rail OTA-Based Current-Sensing Circuits," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 29, no. 7, pp. 1428-1436, July 2021, https://doi.org/10.1109/TVLSI.2021.3071652

[8] T. Papaioannou, I. Kosmadakis and S. Siskos, "A power sensing circuit for solar cells MPP tracking," Design of Circuits and Integrated Systems, Madrid, Spain, 2014, pp. 1-6, https://doi.org/10.1109/DCIS.2014.7035570

[9] K. R. Laker and W. M. C. Sansen, Design of Analog Integrated Circuits and Systems. McGraw-Hill Science, Engineering & Mathematics, 1994.

[10] C. Azcona, B. Calvo, N. Medrano, S. Celma, and M. R. Valero, A CMOS micropower voltage-to-frequency converter for portable applications. 2011. https://doi.org/10.1109/prime.2011.5966237

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 182 – 190

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Prediction of Analog Circuit Sizing Using an Artificial Neural Network******

Shubham Agarwal, Benjamin Prautsch, Uwe Hatnik

Department of Efficient Electronics, Fraunhofer IIS/EAS Dresden, Germany

shubham.agarwal@eas.iis.fraunhofer.de, benjamin.prautsch@eas.iis.fraunhofer.de, uwe.hatnik@eas.iis.fraunhofer.de

Abstract

This paper presents a method for predicting the sizing data of an analog circuit meeting a certain performance with the help of a neural network. The performance data for training is given in two ways. First, an executable function represents the target circuit and second, a lookup table (LUT) is generated from an actual design in the design environment. In order to avoid repeatability and to ensure that the model is tested on a wide range of input datasets, three different datasets are generated through both pre-defined and randomized methods. The model is trained targeting high accuracy. The results are compared and show a good prediction accuracy which verifies the efficiency of the method. We believe that when using this approach, initial sizing of circuits will be eased once the performance was sampled at the beginning. This approach does not aim to completely replace the electrical simulation involved in analog design. Reuse-oriented design flows will take advantage of the method by identifying "go" vs. "no-go" scenarios early in the design process.

Keywords—Artificial neural network, Deep neural network, analog circuits, sizing data, Python

1 Introduction

Analog circuit design is a crucial step in bringing an electronic design project from concept to a fully functional reality as the analog circuits form an integral part of the integrated circuits. However, implementation of the analog circuits cannot be deemed successful until and unless the sizing values are not in line with the required performance values. Sizing values can be described as the input parameter values of a circuit that are required to obtain specific performance output values. Before the advancement of simulation, engineers would manually calculate the design parameters and there was no way to verify the inefficiencies beforehand. The increasing complexity and rising demand for accurate and high performance circuits led to the need of automated circuit designing techniques. Simulating the circuits is helpful, but it is also a very time consuming process. In analog circuit design, sizing is a very important step as it aligns the parameters of the circuits to work in favor of the desired performance. It takes very experienced analog design engineers to implement accurate sizing of components. Using simulation techniques also requires multiple iterations to choose the correct sizing data and get to the desired performance [1]. In order to save this method from being a very costly and time consuming process, artificial intelligence (AI) methods are increasingly used [2] [3] [4].

****** This research was funded by the General Federal Ministry of Education and Research (BMBF) within the frame of the HoLoDEC project, grant number 16ME0700.

AI methods are gaining more and more recognition in providing aid to efficient design of analog circuits and other aspects of realizing analog ICs. Some of the proposed works [5] [4] include automatic synthesis of high-performance analog ICs. In [5], the authors implemented an evolutionary approach using neural-fuzzy performance models that were trained from SPICE simulations. Another work introduces analog layout placement for FinFET technology using reinforcement learning [4].

This paper focuses on the implementation of an AI model that predicts the sizing data as per the required performance data from a lookup table (in order to ease library-oriented reuse). For this purpose, the performance data is taken as input and the sizing data is taken as the output in the AI model. This method will save time by eliminating multiple iterations carried out by the simulation to achieve the correct sizing data values. However, this approach in essence, does not replace the entire process of electrical simulation by the AI model. The AI model is first trained by means of supervised learning. To verify the proper functioning of the AI model before moving on to a complicated circuit, the AI model is first implemented on a voltage divider circuit. The second case taken is an actual LDO circuit. The dataset consisting of sizing data and performance data is created using permutation and combination, hereafter referred to as PnC within a specific range of values in an excel sheet. This AI model is trained using 80% of this dataset and the remaining 20% of dataset is kept aside for testing the AI model [6]. In order to test the robustness and accuracy of the AI model on a wide range of inputs, two more datasets are generated using a random generator and Latin Hyper Cube Sampling (LHS). There is a similar work that introduces prediction of the sizing using machine learning techniques, but the authors generated the optimized dataset using simulation based on the gm/Id technique [2].

The contributions of this paper are as follows:

- The presented work utilizes both sizing values and performance measures directly and builds an AI model that allows mapping from target performance to an initial sizing estimate.
- We propose an AI model, trained using the supervised learning technique that adapts to different circuits and changed inputs and outputs.
- We investigated the robustness and accuracy of proposed AI model by utilizing three data sampling techniques.

This paper further continues with Section 2 which provides an insight into the proposed AI model and its features. Section 3 deals with the generation of the different datasets leading to Section 4 that shows the comparison and results obtained pertaining to the performance of the AI model. Section 5 summarizes with a conclusion and gives an outlook to future work.

2 Proposed AI model

A. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANNs) are [7] networks that consist of artificial neurons shown as circles in Figure 1. The first and the last layers are input and output, respectively, while the layers in-between are called hidden layers. The output of each neuron is connected to every neuron of the next layer. The inner layers remain hidden. Each layer passes the information to the next layer until the data reaches the output layer. Unlike the simple networks that only consist of one hidden layer, there are also many complex networks comprising of up to several hundred hidden layers. Such network is termed as a deep neural network (DNN) [8].

In our work, the AI model was trained by using the supervised learning technique [9]. It is a form of learning that requires classified datasets. The dataset provided to the AI model is specified in terms

of input data and output data. During the training process, the AI model is iteratively adapted such that it creates an estimation pattern between input and output. This is an effective way to ensure that the AI model functions with commendable accuracy across usage scenarios.

B. Model Framework

The proposed AI model is a fully connected deep neural network that predicts the sizing data as per the required performance data. The AI model was first designed to work on a voltage divider circuit to check if our approach yields positive results. Hence three input and three output nodes were used. In order to be adaptive, the number of input and output nodes varies according to the number of performance data values and sizing data values. To further facilitate the network to suggest sizing given a target performance, for training, the performance data was taken as the input and the sizing data was the output. Our AI model has three hidden layers with 128, 64 and 32 nodes respectively. The ReLu activation function was thereby used in the neurons. It adds non-linearity to the neural network, as it decides which node should be activated, which further ensures that the neural network uses only the necessary information for the transformation of the input and ignore the irrelevant information. Out of the available activation functions like Linear, Sigmoid, Tanh to name only a few, the ReLU (Rectified Linear Unit) activation function [10] was chosen. This function does not activate all the nodes at once, hence it increases the computational efficiency of the network.

The proposed AI model can be seen in Figure 1. Figure 2 shows the structure of the model.

In addition to the activation function, optimizers are an essential aid to the AI model. Optimization is an interactive process that utilizes mathematical functions to help in minimizing a given error function. This is because optimizers are methods that can change some properties of the neural network like weights and biases. Adam (Adaptive Moment Estimation) [7] is the optimizer that was used in the AI model.

It is a well-known gradient descent optimization method that calculates the rate of adaptive learning for each parameter.

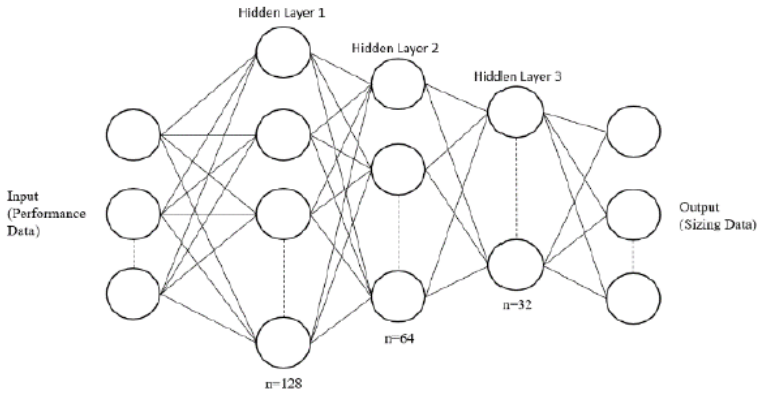


Figure 1. Proposed AI model, where the number of input and output nodes vary as per the requirements of the circuit

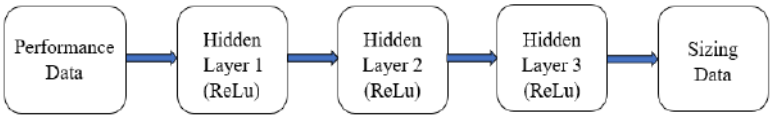


Figure 2: Structure of the model

3 Generation of Datasets

In order to test the AI model, it is necessary to have a dataset which is used to train the AI model. Some of the data samples would be kept aside to test the AI model after it has been trained. But, to check if the AI model functions well on a wide variety of data samples, it is beneficial to have different data sets. Three datasets are generated using different ways to ensure versatile data samples. In this case, the dataset is first generated using permutation and combination and the AI model is trained on it [11]. Then two more datasets are generated using random sampling and Latin Hypercube Sampling (LHS), which are tested on the already trained AI model to obtain the mean squared error (MSE) values, mean absolute percentage error (MAPE) and overall accuracy for each dataset. MSE and MAPE are calculated using inbuilt python library. MAPE (Mean Absolute Percentage Error) is used to measure the magnitude of error produced by the AI model while predicting the output. MSE and MAPE are mathematically represented by equations (1) and (2) respectively.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2 \quad (1)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \quad (2)$$

$$\text{Relative Error} = \left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \quad (3)$$

Where N is number of samples, Y_i is actual data sample, and $\hat{Y}_i$ is predicted data sample.

A. Permutation

The first dataset was created manually for a voltage divider circuit by using permutation and combination for a specified range of sizing values. This same method was used to generate the dataset for an actual LDO circuit.

B. Random Value Generation

Another dataset was created that provides a uniform set of sizing values. This dataset was generated using the random generator from NumPy [12]. NumPy is a Python [13] library that provides quick operations on arrays like mathematical, logical, random simulation, etc. The lower and upper range of values were retained from the dataset generated by permutation and combination. The size of this dataset was also 125 data samples. This dataset was used to test the AI model and the performance characteristics such as MSE, MAPE, and accuracy of the model were obtained.

C. Latin Hypercube Sampling (LHS)

Latin Hypercube Sampling (LHS) [11] is used to sample random values in a way that they are evenly distributed in the sample space. For n variables (in this case I, VR1, and VR2), the sample space of every value is divided into n evenly sampled sections. A random sample is then picked from that

specific section. For the purpose of this paper, LHS proved to be very helpful in providing a uniformly distributed dataset.

D. Testing on an actual LDO circuit

The AI model is also tested on an actual LDO (low dropout generator) circuit to verify if the AI model works effectively on actual circuits. The dataset is generated through the design environment Cadence® Virtuoso®. The purpose is to check if the AI model could handle complexities and show similar accuracies across circuits by just varying the number of inputs and outputs. The sizing parameters to be predicted as output values are the VDD voltage dc(V), mirror transistor width of Error amplifier W_m, pass transistor length l_p and load current Corner nom loadcurrent. The performance parameters taken as input are feedback voltage V_fb, reference voltage Bgr_op and output voltage of regulator ldo_out. The diagram of a LDO and the transistor level LDO can be seen in Figure 3and Figure 4 respectively.

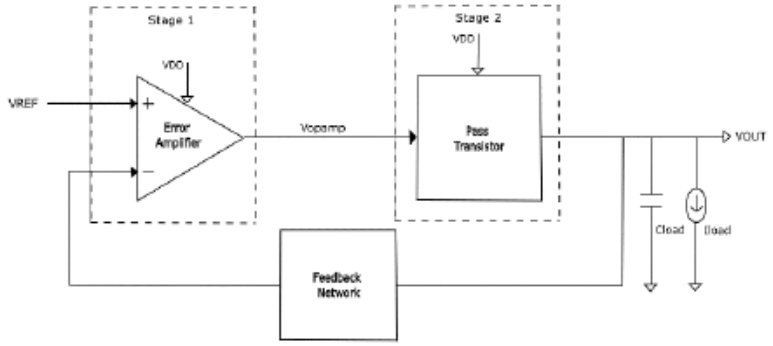


Figure 3. LDO circuit

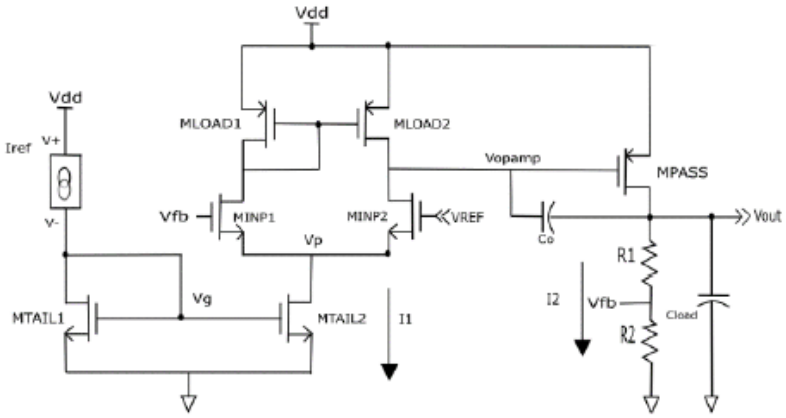


Figure 4. Transistor level LDO circuit

4 RESULTS

After testing the AI model using all three datasets mentioned above, it was observed that the proposed AI model was able to train efficiently from the permutation lookup table and perform with great accuracy on all the sampled datasets. Some important factors to examine the efficiency of the system

such as the MSE (Mean Squared Error), MAPE (Mean Absolute Percentage Error), and accuracy were calculated for each parameter of the sizing data (Vs, V1 and V2) for the voltage divider circuit. Upon testing of the AI model, there was consistency observed with respect to the accuracy in predicting the sizing data for all the datasets, as seen in Table 1.

The values in Table 1 show that the MSE values are very low regardless of how the dataset was generated. However, the lowest value can be observed for the dataset that was generated using the Latin Hypercube Sampling method. As seen in the Table 1, the MAPE is 5.6 for the test data generated using permutation and combination method. This value comes down to 2.63 and 2.58 in case of the datasets generated by random value generation and Latin Hypercube Sampling respectively. Their accuracy was also maintained at 94.4% for the lookup table and 97.36 % for the dataset generated by random generator and 97.41 % for the dataset generated by LHS. The AI model also showed an accuracy of 92.5 % for the LDO circuit, which is expectedly sufficient for initial sizing.

For visualization of the AI model's performance on all the three datasets, graphs were also plotted as seen in Figure 5, Figure 6, and Figure 7. The color bar indicates the relative error of the AI model's resulting sizing values. The relative error is calculated as shown in equation(3). The similar color is also assigned to the related performance data to ease visual mapping of both. This way, a designer can more easily spot the group of good-enough performance data values and find its corresponding sizing data values including the error the performance data of the permutation lookup table on which the AI model was tested and the predicted data. Figure 6 shows the performance of the AI model on randomly generated dataset, and Figure 7 shows the performance on the dataset generated using Latin Hypercube Sampling.

In another scenario, three AI models were created using Permutation, random sampling and LHS. Then a new dataset was generated using LHS (named LHS2) which was tested on these three AI models. The resulting parameters can be seen in Table 2. Here again it is observed that the LHS AI model attains the highest accuracy of 98.07 %.

Dataset	MSE	MAPE	Accuracy (%)
PnC	0.03	5.6	94.4
Random	0.012	2.63	97.36
LHS	0.0102	2.58	97.41
LDO circuit	0.0121	8	92.5

Table 1: Accuracy analysis of the AI model

AI model	MSE	MAPE	Accuracy (%)
PnC	0.013	3.1	96.89
Random	0.008	2.48	97.5
LHS	0.007	1.93	98.07

Table 2: Performance of LHS2 dataset on 3 different AI models

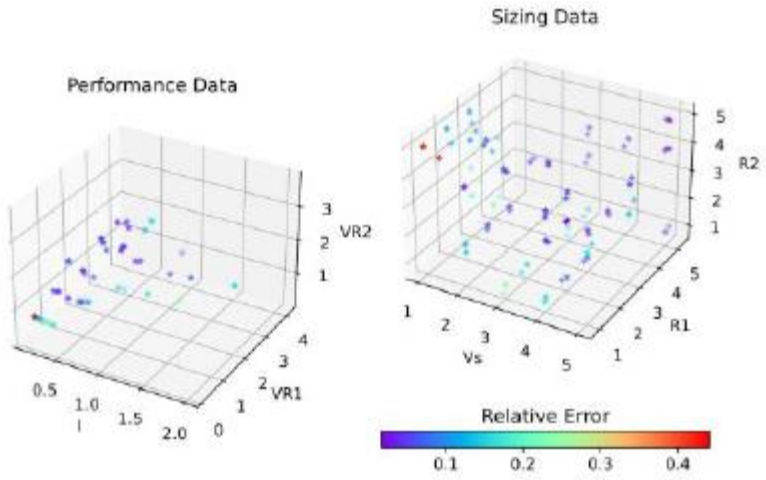


Figure 5: Performance and sizing data for Permutation & Combination (PnC) dataset (125 samples)

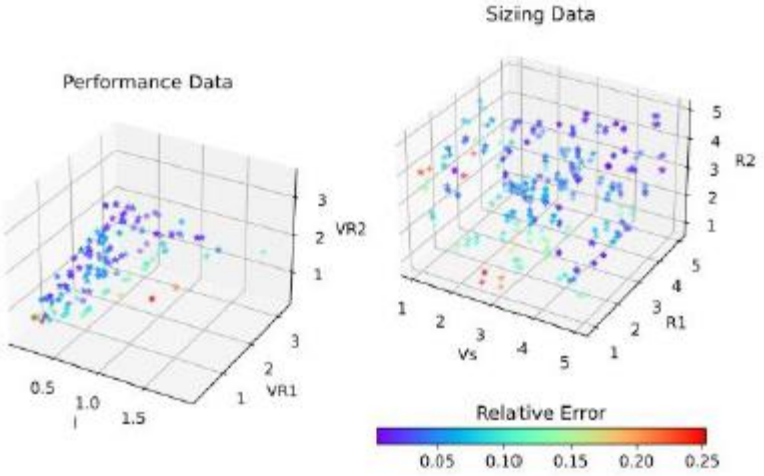


Figure 6: Performance and sizing data for Random dataset (125 samples)

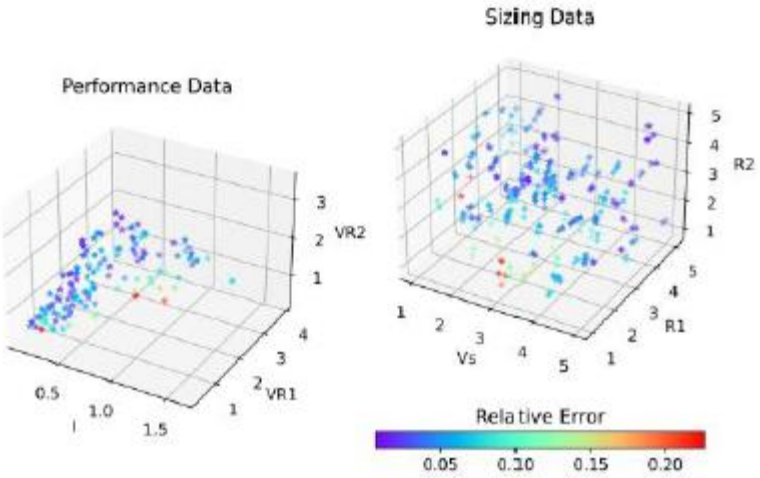


Figure 7: Performance and sizing data for LHS dataset (125 samples)

5 Conclusion and future scope

This paper implemented an AI algorithm to predict the sizing data for the required performance data of circuits. This method is very useful for initial sizing for reuse without completely replacing the electrical simulation for circuit design, as with a given data set, it consumes less time due to no necessity of iterative simulations. The proposed AI model consists of five layers of which three are hidden. It accurately predicts the sizing data. In order to test the trained AI model, a number of different datasets were used to check for the robustness and accuracy of the model. The proposed AI model was able to maintain good accuracy for the three datasets generated. An important thing to note was that the AI model was recorded to be the most robust with an accuracy of 97.41 % and the least value of MSE 0.0102 for the dataset generated using Latin Hypercube Sampling. It was also able to generate an accuracy of 92.5 % on the LDO circuit. In another scenario, when the AI model was trained separately on Permutation dataset, Random sampling dataset and LHS dataset, it was again observed that the LHS trained model attains the highest accuracy of 98.07 % among all the datasets. An important point to be noted is that the models show an increase in accuracy by 1-2% when they are trained on more number of samples. So the number of sample data also plays a role in increasing the accuracy of the AI model. This model proves to be very fast, cost effective, and works well on different circuits with good accuracy.

While this model functions well in terms of accuracy, there is still some scope to improve the model. The datasets generated in Cadence® Virtuoso® design environment are taken in permutation format, which shows comparatively less accuracy. So there is a scope of Cadence® Virtuoso® datasets being generated using LHS method so that the model works with increased accuracy. The model can also be modified in a way that it is able to handle more complex circuits and predict the sizing data regardless of the circuit taken into consideration.

6 References

[1] M. Fayazi, Z. Colter, E. Afshari and R. Dreslinski, "Applications of artificial intelligence on the modeling and optimization for analog and mixed-signal circuits: A review," IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I, 2021.

[2] D. S, G. Tilwankar and R. Zele, "Automated design of analog circuits using Machine Learning techniques," 25th International Symposium on VLSI Design and Test, 2021.

[3] R. Mina, C. Jabbour and G. Sakr, "A review of Machine Learning techniques in analog integrated circuit design automation," Electronics, 2022.

[4] M. Ahmadi and L. Zhang, "Analog layout placement for FinFET technology using Reinforcement Learning," IEEE International Symposium on Circuits and Systems, 2021.

[5] G. Alpaydın, S. Balkır and G. Dündar, "An evolutionary approach to automatic synthesis of high-performance analog integrated circuits," IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 7, NO. 3, 2003.

[6] F. Chollet, Deep learning with python, Manning Publications, 2017.

[7] D. Kingma and J. L. Ba, "ADAM: A method for Srochastic Optimization," ICLR, 2017.

[8] U. Hatnik and B. Prautsch, "Bringing AI to sensors – simulation of hardware-aware AI models," EASS 2022; 11th GMM-Symposium, 2022.

[9] A. Singh, N. Thakur and A. Sharma, "A review of supervised machine learning algorithms," IEEE, 2016.

[10] C. Nwankpa, W. Ijomah, A. Gachagan and S. Marshall, "Activation functions: comparison of trends in practice and research for deep learning," 2nd International Conference on Computational Sciences and Technologies, 2020.

[11] R. Iman, "Latin Hypercube Sampling," in Encyclopedia of Quantitative Risk Analysis and Assessment, John Wiley & Sons, Ltd, 2008.

[12] C. Harris., J. Millman, S. v. d. Walt, R. Gommers, P. Virtanen and D. Cournapeau, "Array programming with NumPy," Nature, 2020

[13] V. Rossum, Guido and F. Drake, "Python 3 Reference Manual," CreateSpace, 2009.

[14] M. Harsha and B. Harish, "Artificial Neural Network Model for Design Optimization of 2-stage Op-amp," 24th International Symposium on VLSI Design and Test (VDAT), 2020.

[15] R. Khandelwal, "Design of a fully on-chip Linear Regulator," New Delhi, 2017.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 191 – 192

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Classification of Fetal Images using Deep Learning Methodologies: The Smart Embryo Project††††††

Lazaros A. Iliadis[1], George Vergos[1], Paraskevi Kritopoulou[2], Achilleas Papatheodorou[3], Sotirios P. Sotiroudis[1], Achilles D. Boursianis[1], Konstantinos-Iraklis D. Kokkinidis[2], Maria S. Papadopoulou[4], and Sotirios K. Goudos[1]

[1] *ELEDIA@AUTH, School of Physics, Aristotle University of Thessaloniki, Thessaloniki, Greece*
[2] *Dept. of Applied Informatics, University of Macedonia, Thessaloniki, Greece*
[3] *Embryolab Fertility Clinic, Thessaloniki, Greece*
[4] *Dept. of Information and Electronic Eng., International Hellenic University, Sindos, Greece*

liliadis,gvergos,ssoti,bachi, sgoudo@physics.auth.gr,
pkritopoulou, kostas.kokkinidis@uom.edu.gr
a.papatheodorou@embryolab.eu, mspapa@ihu.gr

Abstract

This paper provides a technical report on the Smart Embryo project. The Smart Embryo project (Project code: KP6-0079459) combines computer vision, image processing, and artificial intelligence methodologies to classify and characterize fetal images for assisted reproduction applications.

1 Introduction

Millions of children owe their existence to the advancements in in-vitro fertilization (IVF) technology. However, despite the lengthy and costly procedures involved, only one-third of couples undergoing IVF treatment successfully conceive a child. To address challenges such as age, embryo quality, and technological limitations, embryologists and researchers are actively seeking new tools and methods to improve outcomes. One such method introduced over a decade ago is time-lapse imaging incubators (TLI) as part of the IVF process. TLI captures photographs at regular intervals, creating a video showcasing the embryo's development. This technique allows for precise control over culture conditions while capturing and annotating key developmental events, known as morphokinetic parameters, such as cell divisions, blastocyst formation, and expansion.

ADeep learning (DL), a subset of machine learning techniques, has gained significant traction in the field of computer vision (CV) and is recognized as the most successful approach. Convolutional neural networks (CNNs), renowned for their accomplishments in image classification tasks, have emerged as the dominant method in CV and have found widespread applications, including medical

†††††† This research was carried out as part of the project "Classification and characterization of fetal images for assisted reproduction using artificial intelligence and computer vision" (Project code: KP6-0079459) under the framework of the Action "Investment Plans of Innovation" of the Operational Program "Central Macedonia 2014 2020", that is co-funded by the European Regional Development Fund and Greece.

imaging. Over the past five years, there has been a growing research trend in employing DL for IVF-related applications [2]. In this paper, we present a DL model based on convolution, which is trained on a proprietary dataset to classify images of blastocysts, a crucial stage in IVF.

2 Results

Preliminary results related to the Smart Embryo project were carried out during the period 2022-2023. Several CNN architectures were designed and applied to our dataset. Furthermore, transfer learning approaches have been also utilized. In 2023 a custom CNN that is based on the VGG architecture, was combined with two state-of-the-art DL models, to construct an ensemble learner that achieved satisfactory results [1].

In this paper, a custom VGG model with eight convolution layers, and five max-pooling layers, using the rectified linear unit function (ReLU) as an activation function, and three fully connected layers, is trained to classify blastocysts' images in terms of their quality. To study the generalization capabilities of our approach, 5-fold cross-validation is used. Four super-classes are formed to create a balanced dataset. The results are presented in Table 1 and Table 2.

Training Accuracy	Validation Accuracy	Test Accuracy
47.53	45.91	45.74

Table 1: Performance of proposed DL model: Accuracy

Precision	Recall	F1-score
0.45	0.46	0.45

Table 2: Performance of proposed DL model: Statistical measures

3 Conclusions

The main objectives of the Smart Embryo Project are the classification and the characterization of fetal images for IVF applications using SOTA DL and CV methodologies. Preliminary results show that DL models can achieve satisfactory performance. Future research includes the development of novel DL and ensemble learning approaches, utilizing transfer learning methods, and creating an application that will be tested in real-world case scenarios.

4 References

[1] George Vergos and et al. Ensemble learning technique for artificial intelligence assisted ivf applications. In 2023 12th International Conference on Modern Circuits and Systems Technologies (MOCAST), pages 1–4, 2023.

[2] Nikica Zaninovic and Zev Rosenwaks. Artificial intelligence in human in vitro fertilization and embryology. Fertility and Sterility, 114(5):914–920, 2020.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 193 – 199

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Estimation of pedestrian trajectory using LSTM network architecture

Christos Theocharatos[1ǂǂǂǂǂǂ], Dimitris Kastaniotis[1] and Vassilis Tsagaris[1]

[1] *Irida Labs, Patras InnoHub – Kastritsiou 4, Magoula Patras 26504, Greece*

htheohar@iridalabs.com

Abstract

A method is presented for estimating the trajectory of moving objects using deep learning techniques, with application to people / pedestrian detection and tracking. The method utilizes long short-term memory (LSTM) networks that try to learn how to successfully predict the trajectory of people in a plane, given the current position and the last n-positions. Results on real and synthetic data seem promising, compared to conventional computer vision approaches. Moreover, the method seems to illustrate robust behavior in outdoor, unconstrained environments where human trajectories are influenced by external factors and different spatial domains.

1 Introduction

Trajectory tracking of moving objects is a challenging task used in a plethora of areas including robotic technology, automotive industry, logistics (e.g. AGV coordination), traffic management and others. Similarly to how humans identify moving objects using the sequence of trajectory positions, computer vision techniques try to model the time sequence observations to predict the future position of moving objects. The most well-known method for estimating the trajectory of objects is optical flow [Baron1994], which calculates the motion of each pixel between two consecutive image frames, providing robust results to get the trajectory of objects. Optical flow estimates the relative motion of objects, providing important insight on the spatial arrangement of objects. Typical algorithms have been utilized broadly in different application domains such as Lucas – Kanade algorithm [Lucas1981] and Horn – Schunck algorithm [Horn1981]. Kalman Filter, which is an optimal recursive filter for predicting the next state of a discrete data-controlled process from measurements that are typically noisy, is also utilized in some cases with the optical flow for smoothing the motion history and estimating the motion of the object in consecutive frames [Simon2006]. Combined with the use of spatial or temporal relationship of neighbor pixels and the same pixel in several adjacent frames, Kalman Filter can provide robust trajectory estimates to noisy data as well [Dupond2019].

With the advancement of machine and deep learning techniques, specific neural network architectures have been proposed for estimating and predicting the next state of a system operation. For example, a variety of recurrent neural networks (RNNs) are capable of learning long-term dependencies, especially in sequence prediction problems [Dupond2019]. Within the RNN family,

ǂǂǂǂǂǂ Masterminded EasyChair and created the first stable version of this document

long short-term memory (LSTM) networks [Hochreiter1997] are a special field of deep learning architecture that uses feedback connections for processing entire sequences of data (e.g., video data), making them ideal for processing and predicting data. Recently, an attempt has been made to predict human trajectories using LSTM models [Rossi2021], proposing new methods and metrics to help understand trajectories.

In this work, the use of an LSTM network architecture is proposed for estimating the current state of moving objects and predicting the next state within a trajectory sequence. The method is applied to people / pedestrian detection and tracking in a plane, given the current position and the last n-positions. Training is performed on a variety of input data, including real and synthetic ones. Results on real-world data seem promising, compared to conventional computer vision approaches.

2 Methodology

a. Overview of LSTM Network

Long short-term memory (LSTM) is an artificial neural network that is commonly used in the fields of artificial intelligence and deep learning technology. These types of networks, that constitute a type of recurrent neural network (RNN), have been successfully applied in various fields, including natural language processing (NLP) and time series analysis. While LSTM networks are primarily known for their applications in sequential data, they can also be employed in computer vision tasks.

In computer vision, LSTMs are typically used to model and analyze sequential dependencies within visual data such as video analysis (e.g., by capturing temporal dependencies for recognizing actions and generate video captions), object tracking (e.g. by incorporating LSTM layers into tracking algorithms), image captioning (e.g., in combination with CNNs that extract visual features from an image, which are then fed into the LSTM network to generate a relevant caption), image generation (e.g., by using them in generative models such as Variational Autoencoders (VAEs) [Pinheiro2021] and Generative Adversarial Networks (GANs) [Goodfellow2014] to generate realistic images) and image segmentation (e.g., by capturing long-range contextual information and improving the accuracy of image segmentation).

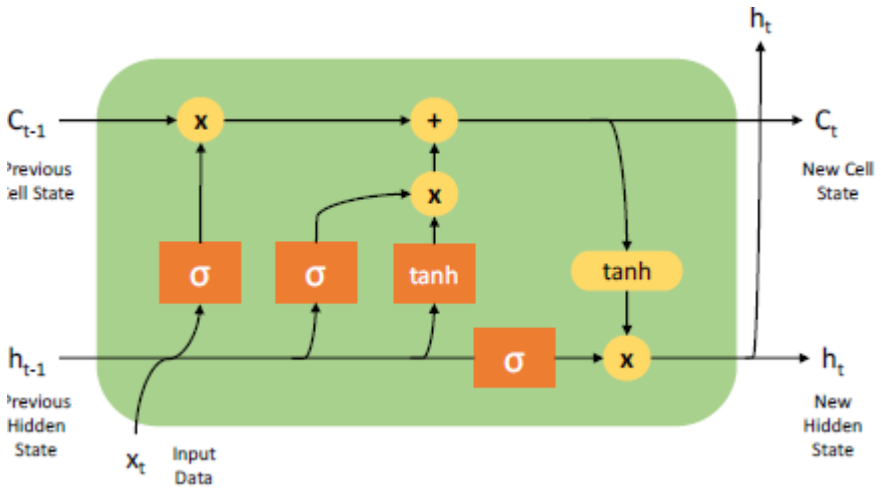


Figure 1: The typical Long Short-Term Memory (LSTM) cell for processing sequential data.

In general, an LSTM cell is a module that can be used to create a bigger neural network. Compared to

a fully connected layer that is purely matrix multiplication of the weight tensor and the input to generate an output tensor, an LSTM building block is more complicated. A typical LSTM cell is presented in Figure 1. The main task is held by a memory cell known as a "cell state" that maintains its state over time, shown as the horizontal line running at the top of the graph. Information can be added or removed from the cell state, regulated by structures called "gates" that optionally control how the information in a sequence of data flows in and out of the cell. Three gates are typically included in an LSTM: the forget gate, the input gate and the output gate. The cell also contains a pointwise multiplication operation and a sigmoid neural net layer that assist the mechanism.

b. LSTM Model Architecture

In the context of our analysis, a modified neural network architecture incorporating an LSTM layer was designed and implemented, utilized in a way to estimate the current state of moving objects and predicting the next state within a trajectory sequence. The input layer takes a portion of an entire trajectory, with each data point containing two variables x and y. This sequence is fed to the LSTM with a hidden state number of 50, followed by two linear layers that lead to the output returning two values, the predicted X and Y coordinates. The utilized activation function between all layers is ReLU (Rectified Linear Unit), which is the most common non-linear activation function in neural networks and returns the maximum of 0 and the input value X, offering simplicity, computational efficiency, and effective training in our LSTM model architecture. Figure 2 illustrates the designed and implemented LSTM model architecture for predicting trajectory sequences and estimating the (X, Y) coordinate positions of moving people, given their former states.

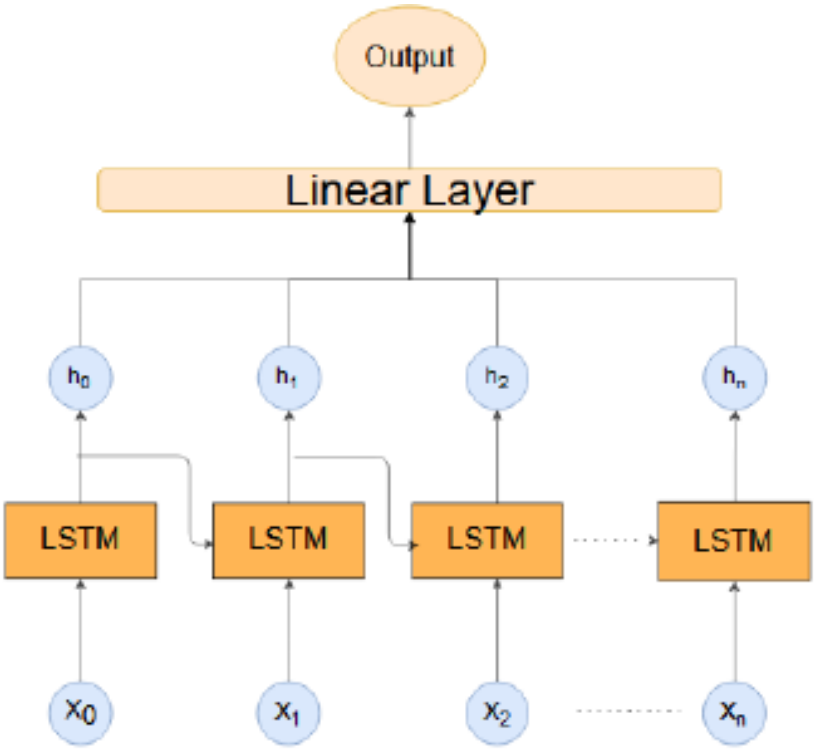


Figure 2: The LSTM model architecture developed for predicting a trajectory sequence.

3 Dataset and Model Training

a. Generation of Image Dataset

Two sources of input image data were utilized in our analysis (a) real world data captured from surveillance cameras in indoor and outdoor environments and (b) synthetic data trajectories that were created using data synthesis tools. In the first case, 220 human data trajectories were captured and sampled to include 30 data-points each, and then anonymized by blurring all people present in the scenes (to avoid GDPR issues). Next, data were manually annotated by drawing a rectangular box around moving people and extracting the central point in an (X, Y) coordinate nomination. To avoid the time-consuming annotation, a people detection CNN model (i.e., Yolo-v5) was utilized for extracting the initial bounding box, based on which the center of mass was afterwards estimated.

In the second case, synthetic data generation was performed using Blender and Unreal engine environments. These tools provide the ability to generate a plethora of annotated data, in different viewpoints and environmental conditions, which can, not only effectively augment the utilized dataset, but also generate data that might not be able to capture in real-world conditions. In our case, 500 tracklets were created comprised of 30 data-points each, that can be active during different frames. Synthetic data trajectories were created by using a mouse pointer to draw trajectories on a specific plane. Then, Blender and Unreal engine were used in a script mode for producing more realistic trajectories, by setting variable backgrounds on the trajectories' drawing window and generating more realistic views. This option allowed us to create different sequences of people walking in real-life scenarios. Using a Python script, the (X, Y) coordinates were extracted from tracklets and stored as sequences. Figure 3 illustrates an input scene and generated trajectories using the synthetic data synthesis tool.

Figure 3: Generation of synthetic data trajectories.

b. LSTM Model Training

The LSTM model was initially trained using part of the Official Dataset (OD) and more specifically the ETH scenes [Pellegrini2009] consisting of 780 people trajectories in outdoor environment with high linearity and low diversity, having a sampling period of 400ms and 30 data-points length of trajectories. All utilized data were proportionally normalized to fall in the range [-1,1]. The OD dataset was augmented using the real-world data sources described in the previous section using the same data characteristics (i.e. sampling period, data points and normalization), and the entire dataset was created in order to proceed with training of the LSTM model. This process is divided into two main stages: (1) data preparation and (2) data training and validation.

In the process of data preparation, generated data from both sources (real and synthetic ones) are saved in csv files with three distinct columns: *Trackid, X, Y*. Each trajectory related to the original data dimensions is normalized between [0-1]. Additionally, zero mean unit variance normalization is performed to avoid exploding or vanishing gradients of the neural network. Next, data augmentation is performed. Points are deleted from a trajectory to introduce sparsity into the data. In addition, trajectory reversal is performed to make more data available and to avoid overfitting the network from trajectories in the same direction. Additionally, the trajectories are split into smaller sub-trajectories using the nmin and nmax parameters. A typical range for input subpaths is within the range of [4-10].

Following data preparation, we proceed with LSTM model training. The OD dataset was entirely utilized in the training process, augmented by 60% of the real-world and synthetic data presented in Section 3.1, while the rest 40% scenes were used for testing and evaluation. PyTorch is utilized to orchestrate the entire training process, incorporating the ADAM optimizer with mean square error loss to tune the gradients and parameters of the neural network. Training was performed on an Nvidia 3070 GPU requiring a total of 18 hours to acquire robust stability. During validation, two types of visualizations are drawn, that is (1) design of the target and the predicted point and (2) design of the target point, the predicted point and future points calculated by feeding a combination of initial network input and output(s). For validation, the MSE and Euclidean distance are calculated as the LSTM model metrics.

4 Experimental Results

Experimental results are provided and evaluated for different real-world videos captured in variable outdoor conditions. The LSTM based trajectory estimation technique is compared against the classical Kalman filter method. During the testing and evaluation phase, the trajectory predictions made by the LSTM network are visualized on real videos that are manually annotated. For each person whose track is annotated, the test script feeds the last 5 positions to the LSTM network and returns the prediction of the next position and the following four future positions.

Based on our analysis, it is shown that the LSTM network can provide solid results and can work effectively as trajectory predictor, especially at fixed levels. Moreover, based on our analysis we have noticed that a completely different image perspective (top-down, high-angle, etc.) has a huge impact on model performance, which needs to be taken into consideration during the data generation and model training phases.

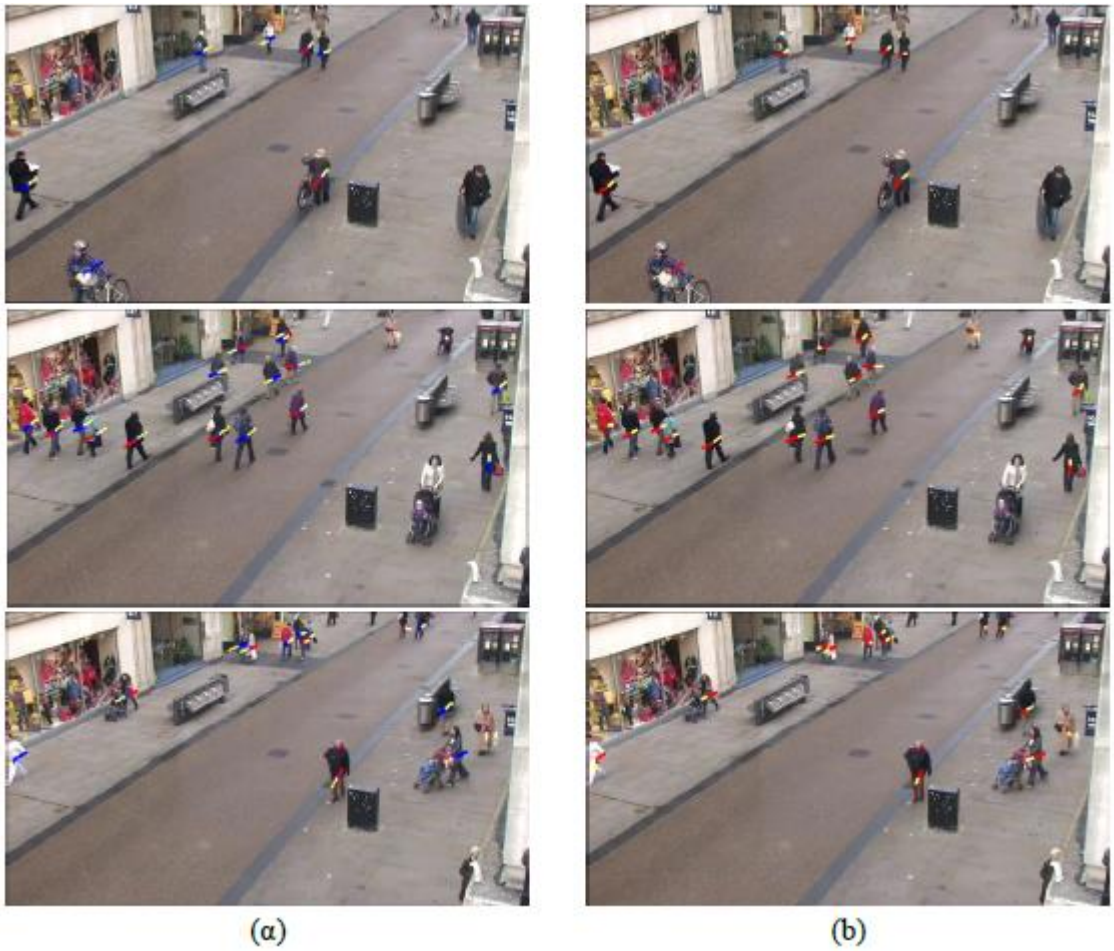


Figure 4: Prediction made by (a) the LSTM network and (b) Kalman filter, for trajectory estimation.

5 Conclusions

In conclusion, an LSTM based architecture is developed for estimating trajectory predictions of people moving in variable scenarios, trained using real and synthetic data. Experimental results show that LSTM networks can work as trajectory predictors, especially at viewpoints that has been trained to predicts. For future improvements, a combination of object detection and LSTM can be used for improving the evaluation results. In addition, trajectory positions can be combined with CNN-based extracted features on each individual person in the image, utilizing reidentification methodologies, in order to resolve the dependence of perspective level on the LSTM performance. Finally, challenges that needs to be addressed, as future improvements of our work, is the analysis of long trajectories and the activity prediction in human analysis.

6 Acknowledgments

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T2EDK-02263).

7 References

[1] Barron, J.L., Fleet, D.J., & Beauchemin, S.S. (1994). Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1), 43–77.

[2] Dupond, S. (2019). A thorough review on the current advance of neural network structures. *Annual Reviews in Control*. 14, 200–230.

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014)*. pp. 2672–2680.

[4] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*. 9(8), 1735–1780.

[5] Horn, B.K.P., & Schunck, B.G. (1981). Determining optical flow. *Artificial Intelligence*, 17, 185-204.

[6] Lucas, B.D., & Kanade, T. (1981). An image Registration Technique with an application to stereo vision. *Proceedings of Image Understanding Workshop*, (pp.121-130).

[7] Pellegrini, S., Ess, A., Schindler, K., & Van Gool, L. (2009). You'll never walk alone: modeling social behavior for multi-target tracking, *Proceedings of IEEE 12th International Conference on Computer Vision*, IEEE, pp. 261–268.

[8] Pinheiro C.L., et al. (2021). Variational Autoencoder. *Variational Methods for Machine Learning with Applications to Deep Networks*. Springer. pp. 111–149.

[9] Rossi, L., Paolanti, M., Pierdicca, R., & Frontoni, E. (2021). Human trajectory prediction and generation using LSTM models and GANs. *Pattern Recognition*, 120, 108136.

[10] Simon, D. (2006). Optimal State Estimation: Kalman, H∞, and Nonlinear Approaches. John Wiley & Sons.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 200 – 206

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

A deep learning approach for detecting defects in melt pool images in DED-AM process

Thanassis Polizogopoulos[1], Christos Theocharatos[1§§§§§§], Vassilis Tsagaris[1], Konstantinos Tzimanis[2], Nikolas Porevopoulos[2], Panagiotis Stavropoulos[2], Konstantinos Ntalas[3] and Albano Memlika[3]

[1] *Irida Labs S.A., Patras InnoHub – Kastritsiou 4, Magoula Patras 26504, Greece*
[2]*Laboratory for Manufacturing Systems and Automation (LMS), Department of Mechanical Engineering and Aeronautics, University of Patras, 26504 Rio Patras, Greece*
[3]*Gizelis Robotics S.A., Kormatzini Area, PC 32009, Schimatari Viotias*

polizogopoulos@iridalabs.com, htheohar@iridalabs.com, tsagaris@iridalabs.com, tzimanis@lms.mech.upatras.gr, porevopoulos@lms.mech.upatras.gr, pstavr@lms.mech.upatras.gr, k.ntalas@grobotics.eu, a.memlika@grobotics.eu

Abstract

A method is presented for monitoring the melting pool of a laser based Directed Energy Deposition (DED-LB/w) Additive Manufacturing (AM) processes and detecting defects during the process using a comprehensive vision system. The approach utilizes a robust deep learning network for extracting prominent features from the melt poll images based on a Variational Autoencoder (VAE) model that is able to distinguish the inherent variations of the melt pool data. The feature vectors are then classified using the SVM algorithm, providing robust classification results on the utilized test set.

1 Introduction

During the past decade, Laser based Directed Energy Deposition (DED-LB/w) Additive Manufacturing (AM) processes with wire as raw material feedstock has become an emerging technology in the industrial manufacturing sector due to its ability to produce big metal parts with high deposition rate, low cost and high material utilization [Gong et. al., 2021]. Despite of the evident advantages, current DED-LB technologies face important limitations that challenge the wide adoption in various manufacturing industries. For example, the final surface quality of 3D metal parts requires machining as further processing, and the quality certification is an issue due to process robustness, stability and repeatability [Rey et. al., 2022]. Due to the increased quality requirement, knowledge of correlations between the main process parameters (laser power, laser head velocity, wire feed rate) and the melt pool behavior, is needed [Theocharatos et. al., 2018]. Lately, a necessity has emerged for being able to monitor the part quality, detect defect formations and make corrections or repairs in situ, as a part is being built. Therefore, real-time process monitoring and closed loop control is needed to better understand and control the thermal behavior of the process, as well as to detect any unpredicted fault and continuously control the interior quality of the machine [Stavropoulos et. al., 2014].

§§§§§§ Corresponding Author

In this work, a vision-based solution for melt pool monitoring of a DED-LB/w process and detecting defects in real-time is presented. Our work is conducted as part of the HybridR National project [HybridR] and is based on the real-time monitoring of the DED-LB/w process with a comprehensive vision sensing system that is able to interact with the machine process algorithms in order to detect defects in real-time and correct deposition errors, leading in an optimal shape of the manufactured part and leading towards zero-defect AM. The interoperable vision system is comprised by an optical camera with appropriate lens and optical filters that is integrated on the laser head in an off-axial configuration. The vision system monitors the geometric and intensity properties of the melting pool and identifies defects during the AM procedure using a variational autoencoder deep learning model that extracts prominent features of the melt pool distribution that are classified using an SVM algorithm. Experimental results on real-time data in different processes seem promising.

2 Hardware Configuration and Data

2.1 Experimental Set-up

As illustrated in Figure 1, the experimental set-up consists of two subsystems: the DED-LB/w system using an articulated industrial robot as the motion mechanism and the monitoring system. A Meltio head is used, being equipped with 6-off axis laser beams with maximum power of 200W each one and totally 1200W as well as an on-axis wire feeding system. For process monitoring, a Basler high-speed optical camera of 227fps at 1440x1080 resolution was selected, equipped with an appropriate Fujinon lens that is able to provide sharp centralized melting pool images at 200mm distance and combined with a Ø25.0 mm premium short-pass filter with a cut-off wavelength at 900 nm for filtering out the frequency of the laser profile. The camera system was applied in an off-axial angled configuration to observe the melt pool along the path where laser is rapidly irradiated during the deposition process.

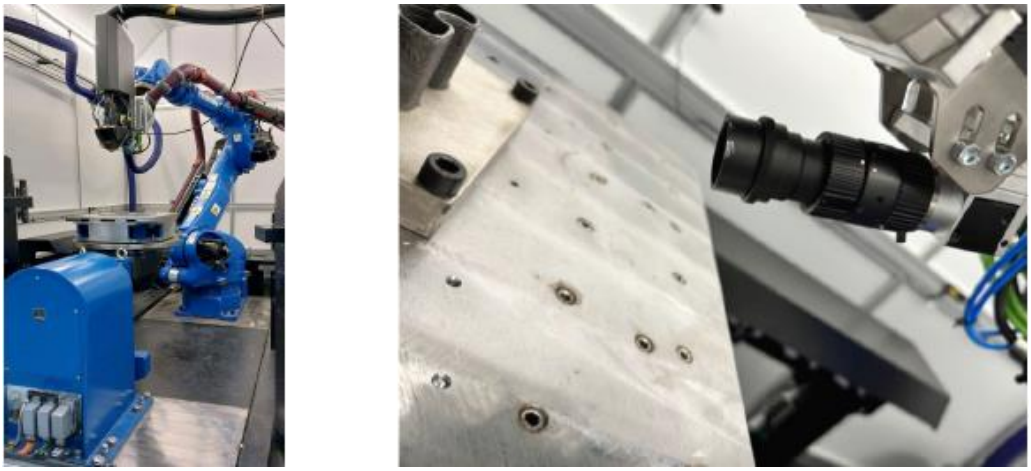


Figure 1: Left: The laser deposition head mounted on the robotic arm. Right: Set-up of the off-axial camera system.

2.2 Experimental Process and Data Description

Experiments were designed to obtain various deposition states, using different parameters of the camera system (e.g., exposure time, fps and angle between the camera lens and the deposition direction) and different parameters of the laser parameters (e.g., laser power varying from 650W –

950W). In particular, besides of the good process, four types of defects were obtained: (Type-1) Wire bending, (Type-2) Wire cutting and depositing, (Type-3) Material drop and (Type-4) Footprint.

Figure 2 illustrates the surface morphology of the melt pool in different deposition states, each of which presents different melt pool morphology. For example, the morphology of the good melting state is relatively regular, while wire cutting (Type-2 defect) present a slightly irregular morphology in the form of a ball of material on the melt pool. In addition, wire bending (Type-1 defect) causes a distortion in the melting morphology due to the contact with the melt pool and is mainly observed at the beginning and ending of a track but also at points of direction change. Material drop (Type-3 defect) is depicted at the starting point (downward flow) due to high heat build-up and presents a morphology pretty similar to a footprint (Type-4 defect) that indicates heat build-up.

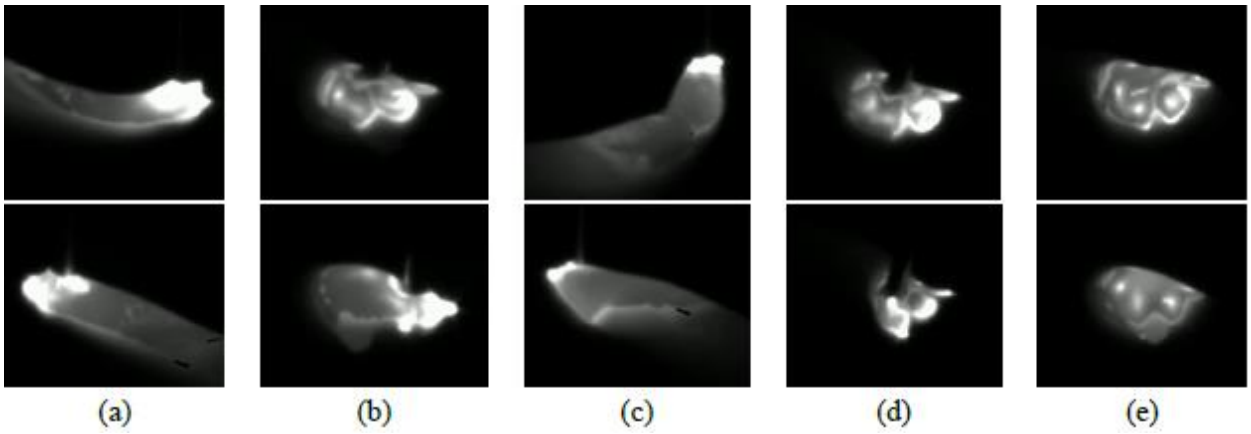


Figure 2: Dataset samples: (a) good process, (b)-(e) Type-1, Type-2, Type-3 and Type-4 defects respectively.

In this study, the target is to identify defects on melt pool images and classify them accordingly using a deep learning model and respective machine learning methodology. For this scope, a melt pool dataset was initially created and is being continuously extended in the current phase of the project. Currently it is comprised of 1200 defected images from all defect types, most of which belong to Type-1 defect class (i.e., 800 out of the total number of defected images). Since the number of defected images in each defect type of the dataset is unbalanced, all different defect types are merged into a defective dataset, leading in a classification problem that is able to identify good vs. defected melt pool images. Therefore, the dataset in the current study is comprised of 1200 good (class-0) and another 1200 defected (class-1) melt pool images, which is randomly split into training and validation sets in a rate of 80-20% respectively, preserving the analogy of normal and defected data on every set.

3 Methodology and Results

In a typical machine learning classification scheme, robust features that are able to capture the geometrical and intensity features of the data distribution can be extracted from the melt pool images and classified properly using a SoA algorithm. In the deep learning era, suitable CNN architectures have been applied to classify molten pool images in different AM or laser welding cases [Wu et. al., 2018; Xia et. al., 2020]. In our study, a hybrid methodology is proposed. Variational Autoencoders (VAEs) are initially used for extracting sophisticated features from the melt pool images, which are then classified using a Support Vector Machine (SVM) classifier as shown in Figure 3.

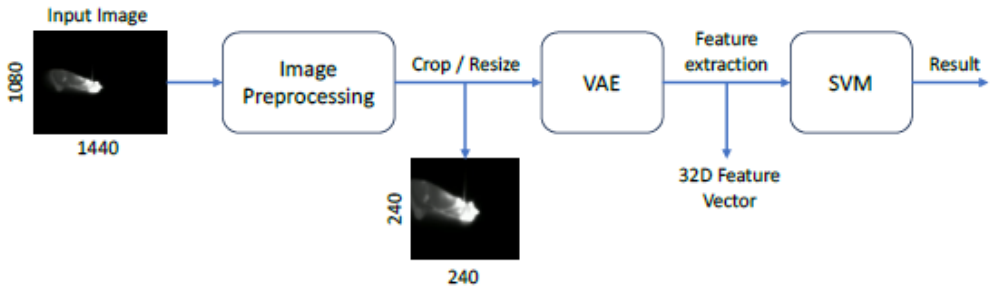


Figure 3: Proposed methodology scheme for detecting defects in melt pool images.

a. Variational Autoencoders

Variational Autoencoders (VAEs) are a type of deep learning generative model commonly used in the field of computer vision, combining elements of both autoencoders (consisting of an encoder and a decoder) and variational inference to learn a compact representation of the input data distribution [Pu et. al., 2016]. More specifically, VAEs are particularly useful because they can learn meaningful latent representations of images, capturing important visual features and variations in the data. In cases like the melt pool data, where the visual distribution does not have a clear dominant pattern, VAEs have the ability to encode the salient features of the input images in the latent space by optimizing a specific objective function, compared to traditional computer vision techniques that rely on handcrafted features and require careful engineering and tuning of the algorithms for specific datasets thus being less capable to automatically learn high-level abstract representations from the input data. In the VAE approach, the encoder maps the input images into a lower-dimensional latent space and the decoder reconstructs the input data from the latent space representation.

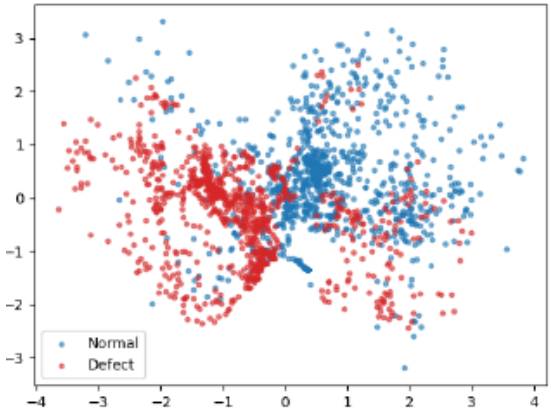


Figure 4: 2-D scatter plot of the extracted feature vectors on the training set, based on the proposed defect detection classification methodology of the melt pool images.

To perform feature extraction using VAEs, initially the image dataset is preprocessed and normalized appropriately. The size of the original image is 1440 × 1080 pixels. Since the melting pool area is located at center of the image and many black pixels exist on the image boundary, the images are cropped to 1080 × 800 pixels to reduce the amount of calculation, and then resized to 240 × 240 pixels. Then, the VAE model was trained using the prepared dataset. The training process involved optimizing the model's parameters to reconstruct the input images accurately while regularizing the latent space distribution. This was done by minimizing a loss function that consists of a reconstruction loss (e.g.,

pixel-wise mean squared error) and a regularization term (e.g., Kullback-Leibler divergence between the learned latent distribution and the target distribution). Once the VAE was trained, images were encoded into the latent space by passing each image through the encoder and sampling multiple latent vectors from the learned latent distribution. The result was a 32-D latent vector for each input image that can be considered as feature representations capturing the most salient and important features of the images, as learned by the VAE model. In order to visualize the extracted features from the training dataset, PCA was performed to reduce the feature dimensionality into a 2-D plane. Figure 4 illustrates the 2-D scatter plot of the extracted feature vectors. The blue dots correspond to the good melt pool samples (class-0), while the red dots to the defected ones (class-1).

b. Classification Results

An SVM was trained on the training dataset of the VAE-based extracted features for performing image classification. The SVM utilized an RBF kernel, while instant C and kernel parameter gamma were selected to be 10 and 1 respectively. In order to validate the results, Accuracy (ratio of correct prediction to the total observations) and Precision (percentage of correct prediction in positive predictions) were used, calculated as:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Accuracy = \frac{TP+TN}{Total\ Samples} \quad (2)$$

where TP are the true positive samples, TN are the true negative samples and FP are the false positive samples. Based on our analysis, experimental results provided $Precision = 92.0\%$and $A\,Accuracy = 91.5\%$.

Figure 5(a) present the confusion matrix on the test set that visualizes and summarizes the performance of the classification algorithm. Finally, by applying the PCA to the test set feature vectors and plotting the first 2 dimensions, the visual illustration depicted in Figure 5(b) is obtained, which shows that the 2-D feature vectors are separable enough and thus the proposed feature extractor works good enough also on the test set.

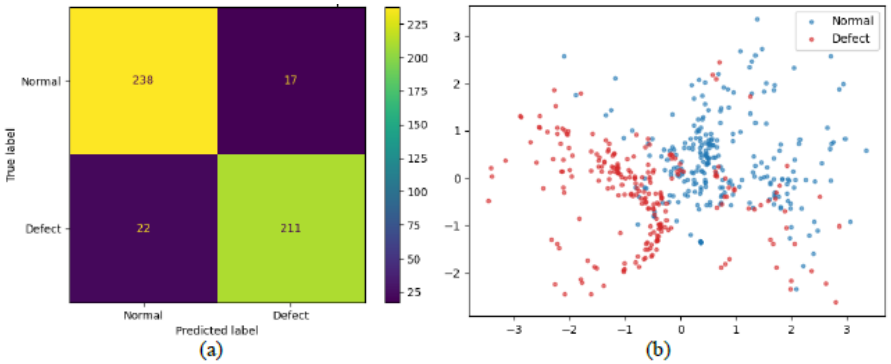


Figure 5: (a) Confusion matrix of the proposed VAE and SVM scheme and (b) 2-D scatter plot of the extracted feature vectors on the test set.

4 Conclusions

A machine learning methodology for detecting defects in melt pool images of a DED-LB/w process is

presented. The images are captured using a comprehensive vision system comprised by a high-speed optical camera and appropriate optics, targeting to monitor the geometrical and intensity characteristics of the melting pool and identify defective images. The machine learning methodology utilizes a VAE deep learning model for extracting prominent visual features from the melt pool data capturing the inherent variations that are able to distinguish a good melting procedure compared to a defective one. The extracted feature vectors are then classified using an SVM classifier, providing promising classification results on the utilized test set.

As a future extension of this study, our intension is to extend the HybridR dataset in a way to include more defective data from the low representative types (i.e. Types-1, -2 and -3), in order to generate a balanced dataset that will contain around 1000 images from each defect class and provide the respective classification per class type. Also, the inclusion of more melt pool states should be considered, and the dataset will be expanded to improve the robustness of the deep learning model and respective machine learning classification methodology. Furthermore, the VAE algorithm will be optimized to improve accuracy and real-time performance.

5 Acknowledgments

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T2EDK-03896, MIS: 5069999).

6 References

[1] Gong, G., Ye, J., Chi, Z., Wang, Z., Xia, G., Du, X., Tian, H., Yu, H., & Chen, C. (2021). Research status of laser additive manufacturing for metal: a review. *Journal of Materials Research and Technology*, 15, 855–884.

[2] Rey, P., Prieto, C., González, C., Tzimanis, K., Souflas, T., Stavropoulos, P., Rathore, J.S., Bergeaud, V., Vienne, C., & Bredif, P. (2022). Data Analysis to assess part quality in DED-LB/M based on in-situ process monitoring, in *Proceedings of the 12th CIRP Conference on Photonic Technologies [LANE 2022]*, 111, 345-350, 4-8 September, Fürth, Germany.

[3] Theocharatos, C., Vassalos, V., Besyris D., & Tsagaris, V. (2018). Closing the loop in Additive Manufacturing - An embedded solution for real-time melt pool monitoring, in *Proceedings of Embedded World Conference 2018*.

[4] Stavropoulos, P., Chantzis, D., Doukas, C., Papacharalampopoulos, A., & Chryssolouris, G. (2014). Monitoring and control of manufacturing processes: a review. In *Proceedings of the 14th CIRP Conference on Modelling of Machining Operations*, Turin.

[5] HybridR website: http://www.hybridr-project.eu/.

[6] Wu, B., Pan, Z., Ding, D., Cuiuri, D., Li, H., Xu, J., & Norrish, J. (2018). A review of the wire arc additive manufacturing of metals: properties, defects and quality improvement. *Journal of Manufacturing Processes*, 35, 127-139.

[7] Xia, C., Pan, Z., Fei, Z., Zhang, S., & Li, H. (2020). Vision based defects detection for Keyhole TIG welding using deep learning with visual explanation, *Journal of Manufacturing Processes*, 56, 845-855.

[8] Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., & Stevens, A. (2016). Variational Autoencoder for Deep Learning of Images, Labels and Captions, In *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS2016)*, Barcelona, Spain, 2360-2368.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 207 – 213

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Machine learning methods for the discrimination of refrigerant gases

Nikolaos Argirusis[1], Petros Karvelis[2], Georgia Sourkouni[3], John Konstantaras[4], Alice Baroncelli[5], Peter Segers[5], Alexandros Athanasiadis[2] and Christos Argirusis[6]

[1] *mat4nrg GmbH Clausthal-Zellerfeld, Germany*
[2] *Department of Informatics and Telecommunications University of Ioannina, Greece*
[3] *TU Clausthal, Centre of Materials Technology 38678 Clausthal-Zellerfeld, Germany*
[4] *Energy and Environmental Research Laboratory, National and Kapodistrian University of Athens Psachna, Evia, Greece*
[5] *Daikin Europe N.V., Ltd.Brussels, Belgium*
[6] *School of Chemical Engineering National Technical University of Athens 15773 Zografou/Athens, Greece*

nikos.argirusis@mat4nrg.de, pkarvelis@uoi.gr, cogsa@tu-clausthal.de, johnkonst@uoa.gr, baronchelli.a@bxl.daikineurope.com, segers.p@bxl.daikineurope.com,csst9604@yahoo.com, amca@chemeng.ntua.gr

Abstract

An IoT-based approach, combined with machine learning algorithms, presents a promising solution for the discrimination of pure refrigerant gases (R32, R134a) in various applications. By utilizing a network of interconnected gas sensors, continuous monitoring and real-time analysis of the environment can be achieved, generating valuable data for machine learning models. It is possible to use supervised learning techniques, like random forests, to create classification models that correctly distinguish between pure gases based on their distinctive characteristics, such as spectral signatures or sensor responses. This innovative coupling of IoT and machine learning enables the development of robust, automated gas detection systems that ensure high accuracy and low latency in the discrimination of pure gases, paving the way for improved safety, efficiency, and environmental sustainability in numerous industrial and commercial settings.

1 Introduction

The rapid advancements in the Internet of Things currently increases the number of interconnected devices, which generate an enormous amount of data. This data holds significant potential for improving the performance and efficiency of various systems and processes. Mobile gas detection and identification is one such application, where the combination of IoT and machine learning techniques can offer a more accurate and reliable means to distinguish between different gaseous substances, such as refrigerants, which are a major environmental problem as they contribute to the greenhouse effect. HVAC-R like air conditioning, and refrigeration systems frequently employ these gases and are subject to strict regulations due to their environmental impact (ASHRAE, 2019). Accurate detection and classification of these gases are of paramount importance for safety, environmental, and economic reasons.

In the context of IoT, gas sensors can be deployed in a networked configuration, allowing for continuous monitoring and real-time analysis of the surrounding environment. The data collected by these sensors can be processed and analyzed by machine learning algorithms, either remotely or locally, depending on the requirements of the application (Shi, 2016). By leveraging the combination of IoT and machine learning, it is possible to develop robust gas detection systems that can automatically identify and distinguish between different gases, such as R32 and R134a, with high accuracy and low latency. These systems can provide critical information for decision-making, ensuring the safe and efficient operation of HVAC-R systems while minimizing their environmental impact.

Contrarily, machine learning, focuses on creating algorithms that can recognize patterns in data without needing to be explicitly programmed. This ability to learn from data allows machine learning algorithms to adapt to changing conditions, making them well-suited for gas classification tasks in IoT environments (Al-Fuqaha, 2015). For instance random forests (Cutler, 2007, Pal, 2005, Khaidem, 2016, Karvelis, 2017) and neural networks, can be employed to build models that can accurately classify gases (Zhang 2018) based on their unique features, such as spectral signatures or sensor responses [3].

This study presents a promising solution for automated gas detection and classification in IoT environments, leveraging the synergies between IoT and machine learning to ensure the safe and efficient operation of HVAC-R systems while minimizing their environmental impact. This approach has the potential to be extended to other gas detection applications, further illustrating the versatility and efficacy of integrating machine learning with IoT for enhanced decision-making and system performance. There are no portable devices in the market, which are able to upload the data to a business-to-business (B2B) database. The device is using infrared irradiation in order to obtain data on the nature and purity of the refrigerant gases. This is a portable device which can be used in the field allowing for an accurate estimation of the purity of recovered refrigerant gases. The device also identifies the corresponding recovery bottle through a QR-code and uploads the data automatically to a B2B database helping direct business by reusing the gases and avoiding fraud.

2 IoT Approach

The approach followed in the present work was to establish a hardware environment consisting of a reactor with an IR source and a 4 channel infrared sensor working at different wavelengths (3.91 serving as reference and 4.26 μm for CO2 detection) or wavelength regions (5.5-18 μm and 8-14 μm). Based on the obtained signals from each channel with time one can perform data analysis which results in a percentage of the input gas.

Traditionally the data analysis can be performed based on Fourier transformation for the identification of the gases and peak mathematics for the gas quantification or simply by the use of the Beer-Lambert law, which demands a calibration of the procedure based on different gas concentration. The decision is then taken by comparing the value resulting from the Beer-Lambert law with the calibration curve (ideally linear dependence).

Purity specifications for refrigerants are needed for the verification of the composition, and to specify the associated methods of testing for acceptability of the refrigerants in the recycling market. These specifications can be found in the AHRI Standard 700 and the AHRI Standard 740 (AHRI, 2015a). The compositions given therein build the basis for the decision making if a gas is useful for recirculation in the market or if it must be destroyed.

The hardware periphery (pumps, valves, voltage sources, etc) is controlled by an STM32 μ-controller, which also performs the data acquisition with a scan rate of 1 point/ms. The IR source is pulsed at 10 Hz with a duty cycle of 62 %. Within a pulse 100 data points per channel are collected of which only 60 are usable (the ones taken during the on time of the IR-source). The collection time is 1000 pulses. The collected data are then transferred to a raspberry pi μ-computer, where the data analysis is performed.

3 Feature Extraction and Machine Learning

Feature extraction (Christ 2018) is a critical step in the process of signal classification, as it involves transforming raw signals into a set of meaningful features (Quinlan 1986). The objective of feature extraction is to keep the most critical information while lowering the number of dimension. thus enabling more accurate and computationally efficient classification models (Wang 2017). Various techniques for feature extraction have been proposed in the literature. One of these type of feature extraction techniques like time-domain methods focus on computing features, such as mean, variance, and kurtosis, directly from the raw signals (Guyon 2003, Ince 2009). Frequency-domain methods, on the other hand, analyze the spectral content of the signals, often using Fourier or wavelet transformations to obtain features related to the signal's frequency components.

One of the significant advantages of using tsfresh (Christ 2018) is its built-in feature filtering mechanism, which helps in identifying and retaining only the relevant features while discarding the others. The tsfresh module computes a comprehensive set of features, including statistical, spectral, and temporal characteristics, by analyzing the properties of time series data. Some of these features include mean, standard deviation, skewness, kurtosis, autocorrelation, Fast Fourier Transform (FFT) coefficients, and linear regression coefficients, among many others. A detailed description of the list of features can be found in (Christ 2018).

Random forests (RF) was introduced by (Breiman 2001) because to their durability, interpretability, and high predictive accuracy, and have experienced tremendous success for machine learning applications. They build decision trees (DT) and try to estimate the correct class of the data. The algorithm is based on the principles of bagging and feature randomness, which lowers the chance of overfitting and enhance the generalization (Genuer 2010). Additionally, random forests provide a natural way to estimate feature importance.

The underlying concept of the random forest algorithm can be briefly summarized in the following steps:

1. Bootstrap sampling: Given a training dataset DD of size NN, we generate MM bootstrap samples (datasets) by sampling NN instances with replacement. Each of these datasets, $D_i, i = 1,2,\dots,M$, is used to train a decision tree.
2. Tree construction: For each bootstrap sample D_i, a decision tree T_i is constructed. For each node, a smaller and randomly selected set of number features (usually $\sqrt{p}$ or $p/3$, where p is the total number of features) is considered for splitting. There are a number of statistical measures to measure of data impurity like Gini impurity or information gain and this is used to estimate the splitting point. Finally, the tree stops growing using a stopping criterion like minimum number of samples per leaf or maximum depth.
3. Aggregation: After training all M trees, the random forest classifier combines their predictions.

Mathematically, let(x) be a sample, and $h_1(x)$ the prediction of the i-th tree for that sample. For classification, the final output yy of the random forest is given by:

$$y = model\ (h_1(x), h_2(x), \ldots, h_{1M}(x))$$

4 DataSet

In this study, we employed an IoT-based approach for collecting data from a variety of refrigerant gases, including R32, R134a and others. A custom-designed sensor array consisting of multiple gas sensors was deployed to detect and measure the concentration levels of the target gases. These sensors were selected based on their sensitivity, selectivity, and compatibility with the refrigerants under investigation. The sensor array was connected to an IoT platform, which facilitated real-time data collection, transmission, and storage.

The data acquisition process involved exposing the sensor array to controlled gas environments containing different refrigerant gases at various concentrations. These environments were created in a laboratory setting using gas chambers and precisely controlled gas mixtures. The sensors were calibrated before each experiment to ensure accurate and consistent measurements. During the experiments, the sensor array continuously monitored the gas concentrations and transmitted the data to the IoT platform.

In this study, a total of 221 data samples were collected, where each sample corresponds to a specific gas concentration and sensor response. The collected data have been summarized in Table 1 which provides an overview of the number of samples per refrigerant gas type. The table illustrates the distribution of the dataset across the different gases, highlighting the balanced nature of the data collection process. This comprehensive dataset enabled the effective training and evaluation of the machine learning algorithms, ensuring that the discrimination of refrigerant gases was conducted in a robust and reliable manner.

	Number of signals					
Dilution with Air	*0%*	*5%*	*10%*	*25%*	*50%*	*70%*
R134a	20	20	20	12	12	12
R32	25	20	20	20	20	20

Table 1: Distribution of data samples across different refrigerant gases in the dataset.

5 Results

Using a stratified holdout method, the data was split into two sets namely the train and test subsets in order to assess the effectiveness of the machine learning algorithms. This approach ensured that the proportions of each gas class were maintained in both the training and testing sets, reducing the likelihood of biased performance estimates while several performance metrics can be employed. These metrics will provide insights into the accuracy, reliability, and overall effectiveness of the classifier. Here, we describe some common performance metrics for a scientific manuscript:

Accuracy: This number is calculated by the ratio of correct classified instances divided by the total instances. A common performance metric for categorization tasks is this one. However, it may not be the most appropriate measure in cases with imbalanced class distributions, as it can be misleading.

Precision: Precision calculates the ratio of true positive data among the positive instances. It

represents the classifier's ability to correctly identify positive instances while minimizing false positive errors.

Recall (Sensitivity): Recall calculates the ratio of true positive instances divided by the actual positive instances. It evaluates the classifier's capacity to find every positive case.

F1 Score: Is calculated as the harmonic mean of recall and precision. When dealing with imbalanced datasets or when both false positives and false negatives must be considered, it provides a balance between precision and recall, making it a valuable metric. The accuracy was measured and found to be 93.02% for the two classes. Below we also provide the performance metrics for our dataset:

	Precision	Recall	F1-Score
R32	0.9200	0.9583	0.9388
R134a	0.9444	0.8947	0.9189

Table 2: Distribution of data samples across different refrigerant gases in the dataset.

6 Conclusions

In conclusion, this study has successfully demonstrated the effectiveness of an IoT-based approach combined with machine learning algorithms for the discrimination of refrigerant gases. By leveraging the power of IoT sensors and data collection techniques, we were able to gather high-quality and representative samples of gas measurements. The application of the the RF classifier, enabled us to analyze this complex data and accurately distinguish between different refrigerant gases, such as R32 and R134a.

The proposed approach holds significant potential for improving the detection of refrigerant gases in various industrial and environmental contexts. By incorporating advanced data processing techniques, including dimensionality reduction and feature extraction, we have shown that it is possible to enhance the interpretability and accuracy of gas classification models. In the context of evolving regulations, it will be of high importance and the need for effective management of refrigerant gases to mitigate their environmental impact and ensure compliance with safety standards. Moreover, the integration of IoT technology and machine learning algorithms in this study opens new avenues for further research and development in the field of gas discrimination. Future work might explore deep learning or ensemble techniques, to further improve classification performance. Additionally, the scalability and flexibility of the IoT platform enable the incorporation of more advanced sensing technologies and the extension of the system to cover a wider range of gases and applications.

7 Acknowledgment

This project has received funding from the LIFE Programme of the European Union under grant agreement LIFE19 CCM/AT 001226 - LIFE Retradeables.

8 References

[1] Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of things: A survey on enabling technologies, protocols, and applications. IEEE Communications Surveys & Tutorials, 17(4), 2347-2376.

[2] ASHRAE. (2019). ASHRAE Standard 34-2019, Designation and Safety Classification of Refrigerants. ASHRAE, Atlanta, GA.

[3] AHRI STANDARD 700-2015 Specifications for Refrigerants, https://www.ahrinet.org/search-standards/ahri-700-700c-and-700d-specifications-refrigerants

[4] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[5] Christ, M., Braun, N., Neuffer, J. and Kempa-Liehr A.W. (2018). Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). Neurocomputing 307 (2018) 72-77, doi: 10.1016/j.neucom.2018.03.067.

[6] Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792.

[7] Daubechies, I. (1992). Ten lectures on wavelets. Society for Industrial and Applied Mathematics.

[8] Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. Pattern Recognition Letters, 31(14), 2225-2236.

[9] Graps, A. (1995). An introduction to wavelets. IEEE computational science & engineering, 2(2), 50-61.

[10] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

[11] Ince, T., Kiranyaz, S., & Gabbouj, M. (2009). A generic and robust system for automated patient-specific classification of ECG signals. IEEE Transactions on Biomedical Engineering, 56(5), 1415-1426.

[12] Khaidem, L., Sivakumar, R., & Raman, S. (2016). Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:1605.00003.

[13] Karvelis, P., S. Kolios, G. Georgoulas and C. Stylios, "Ensemble learning for forecasting main meteorological parameters," 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 2017, pp. 3711-3714, doi: 10.1109/SMC.2017.8123210.

[14] Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. R news, 2(3), 18-22.

[15] Lucas, J., Gaël, V., & Damoulas, T. (2019). Time series classification with random forests and application to zero-shot learning. arXiv preprint arXiv:1906.10885.

[16] Pal, M. (2005). Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1), 217-222.

[17] Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301.

[18] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106. doi:10.1007/bf00116251

[19] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637-646.

[20] Zhang, J., Yan, J., Feng, S., Lei, Z., Hu, X., & Meng, Q. H. (2018). Gas identification using deep convolutional neural networks. Sensors, 18(1), 191

[21] Learning Research, 9, 2579-2605.

[22] Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: A comparative review. Journal of Machine Learning Research, 10, 66-71.

[23] Wang, H., & Raj, B. (2017). On the origin of deep learning. arXiv preprint arXiv:1702.07800.

[24] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. Chemometrics and Intelligent Laboratory Systems, 2(1-3), 37-52.

[25] Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. Maaten, Laurens van der and Geoffrey E. Hinton. “Visualizing Data using t-SNE.” Journal of Machine Learning Research 9 2579-2605.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 214 – 219

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Design of an Autonomous, Multi-functional Stress Assessment Sensor for Naval Applications: The AMSA project

Gregory Doumenis[a], Vasiliki Naskari[a], Evangelos V. Hristoforou[b], Polychronis Pattakos[b], Georgia Stamou[b], Christos Papakis[c] and Ioannis Masklavanos[c]

[a]*Autonomous Systems Laboratory, Dept. Informatics and Telecommunications, University of Ioannina, Arta, Greece*
[b]*Laboratory of Electronic Sensors, National Technical University of Athens, Athens, Greece*
[c]*METIS Cyberspace Technology, Athens, Greece*

naskvasil@uoi.gr

Abstract

The AMSA research project addresses the need for a reliable, cost-effective system to measure stress in metal structures of maritime transportation systems. Utilizing the innovative Steel Health Monitoring (STEHEMON) non-destructive test methodology, the project assesses structural integrity through stress analysis in magnetically excited steels. The STEHEMON device is based on two Hall sensors and a Yoke and operates autonomously, drawing energy from ship movements, enabling real-time stress monitoring. Applicable to various transportation systems, such as maritime, fuel transportation, and land transport, the project integrates subsystems into an autonomous collection and processing system (WADSA node), paving the way for commercial exploitation. The validation phase includes laboratory evaluations and real-world demonstrations in the engine room of seagoing ships. AMSA showcases scientific innovation across Materials, Communications, IoT, and Machine Learning domains, enhancing safety and efficiency in cargo transport systems.

1 Introduction

Steel structures' maintenance is a costly process, hindered -among other things- by the lack of in-the-field cost effective monitoring methodologies. The shipping industry faces significant challenges due to the lack of a reliable and cost-effective system for measuring stress in metal structures. Traditional inspection methods are often intrusive, time-consuming, and expensive, hindering real-time monitoring of structural integrity. Consequently, undetected stress concentrations and weaknesses in metal components pose potential safety hazards and can result in substantial economic losses.

The alarming frequency of structural failures in cargo transport systems has raised serious concerns within the industry. Catastrophic incidents, such as structural collapses and material fatigue, have not only endangered the lives of personnel but also led to devastating environmental impacts and financial burdens for companies involved. Moreover, the inability to detect and address stress-related issues in a timely manner has hindered the industry's progress towards ensuring safer and more efficient transportation of goods.

In response to these critical issues, we proposed the AMSA research project. The goal is to create a system that could continuously and accurately assess stress levels in metal structures during the operational phase of cargo transport systems. For this purpose, sensors based on Hall effect can be used to measure magnetic flux density of a ferromagnetic material, such as steel and thus monitor changes in its magnetic properties, due to stresses or other effects.

2 Project description

The primary objective of the AMSA research project is to develop and evaluate a reliable, cost-effective, and energy-autonomous measurement system for monitoring the fatigue and load of metallic structures. This system is intended for installation and operation in challenging and remote locations on vessels or equivalent cargo transport systems, such as containers and trains. It will derive its energy from the ship's motions and transmit wireless measurements pertaining to fatigue and stress to the control center.

To achieve this, the project will investigate and assess the novel STEHEMON method for non-contact analysis of stress and estimation of structural integrity in metallic elements. Concurrently, hardware and software subsystems will be developed to realize a prototype system, reaching a Technology Readiness Level (TRL) of 6, for field testing and real-world evaluation.

The installed system will operate perpetually, continuously processing measurements in near real-time, providing instantaneous analysis of momentary fatigue and overall deterioration of structural components. It will proactively alert relevant personnel about potential structural failures, enabling timely preventive actions.

The developed sensing element arrangement consists of two Hall sensors, which measure the magnetic flux density (B) of a steel structure and a yoke, formed of a permanent magnet and two high permeability soft ferromagnetic poles. The yoke is placed above the Hall sensors and forms a magnetic circuit, when placed closely to the steel surface under measurement.

The use of permanent magnets instead of excitation coils has the advantage of not requiring an additional power source for generating the magnetic field. The measured magnetic flux density by this device is relative to the material’s microstructural configuration, due to defects, inhomogeneities, stress or permeability tensors [1].

The total device includes also the required electronics, connected to the above-mentioned arrangement to power the sensors, receive and process their signal. The output of the two linear bipolar SS49E Hall sensors is directed to the ADS1115 Analog-to-Digital Converter (ADC) with a resolution of 16-bits. The digital signal produced in this stage is received by a microcontroller and converted to magnetic flux density values in mT, which can indicate the stress level of the material with the use of appropriate algorithms [2].

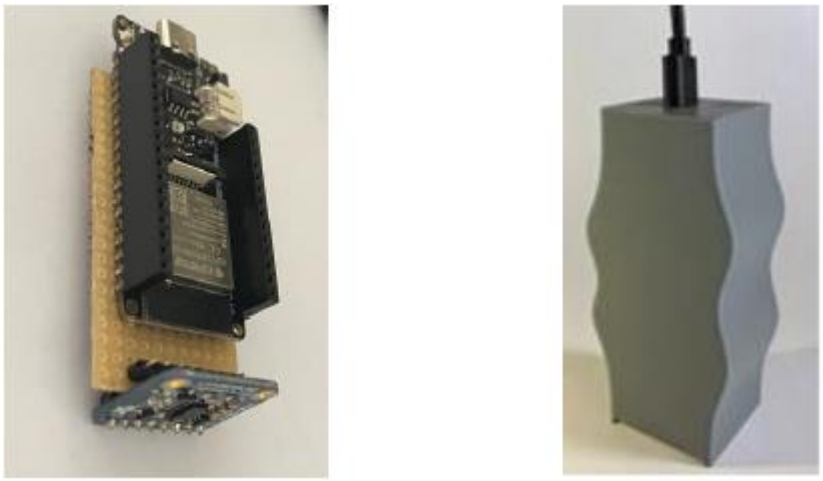


Figure 1. Experimental sensor and 3-d printed enclosure

Additionally, the project explores innovative means of utilizing ambient energy at the installation site, to power the sensor without relying on external power connections, thereby enhancing system autonomy, and minimizing maintenance demands. Towards this direction, an energy harvesting device is been developed and integrated, utilizing an electromagnetic microgenerator to transform the stochastic vibrations occurring on a ship to electric energy [3].

Figure 2. Rendering of an experimental electromagnetic microgenerator

A key element for the success of the project is the transformation of the individual sensor's readings into meaningful data, capable of supporting costly decisions (such as the recall of a commercially operating ship for maintenance/repair). Following the calibration of the sensor in a laboratory environment, a dataset will be generated from the installation of a set of sensors into ships. Advanced AI and machine learning techniques will be employed to derive meaningful insights from the dataset, related to element load distribution, cumulative fatigue, and the generation of comprehensive reports and notifications pertaining to operational deviations and proactive maintenance needs.

The results of the analysis will be readily available to ship operators and engineers through the METIS cloud-based ship data analysis platform.

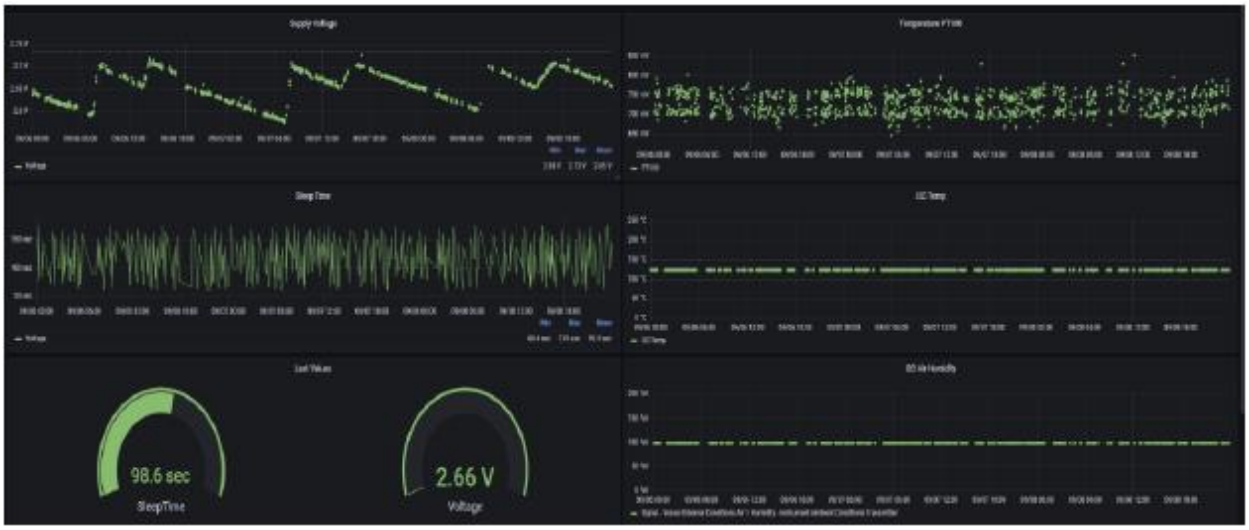


Figure 3. METIS Dashboard

3 Project validation

The validation phase of the AMSA research project constitutes a pivotal step in establishing the commercial viability and real-world applicability of the developed subsystems. This section presents the methodology and scenarios utilized to assess the performance of the system components, ensuring their seamless integration and functionality under demanding conditions.

In the laboratory setting, each subsystem will undergo evaluation to gauge its individual performance and functionality. During the testing phase in the laboratory, the sensors' measurements and performance are checked on several steel samples with or without defects, caused by stresses, corrosion, or after heating an area of the sample and rapidly cooling it.

To demonstrate the practical feasibility and real-world readiness of the AMSA project, the WADSA nodes are installed and put to operation in a rigorous scenario. The selected scenario involves the deployment of WADSA nodes within the engine room of a seagoing ship. This scenario presents a challenging environment for data collection and transmission, emulating the demanding conditions typically encountered in maritime transportation systems.

The objective of this real-world testing is to evaluate the effectiveness and accuracy of the stress measurement and assessment functionalities of the project under operational conditions. By subjecting the system to a realistic environment, any potential limitations or optimizations required for optimal performance can be identified and addressed.

4 Expected results

The AMSA project aims to create the necessary conditions that will lead to progress in applied research in the fields of safety and predictive maintenance of steel structures, while also fostering the development of strong (and protected) technological expertise:

- Conducting cutting-edge research in the domain of predictive maintenance for metallic elements in transportation systems.

- Exploring advanced magnetic sensor technology and energy autonomy methods for IoT devices.
- Designing and developing a prototype fatigue measurement unit for steel structures, which monitors the condition of a ferromagnetic material, such as steel, by detecting changes in its magnetic properties, through the measurement of the magnetic flux density.
- Creating an integrated system for failure warning and predictive maintenance of steel constructions in vessels.

The results obtained from the project's validation hold broad applicability across various transportation systems that incorporate steel (magnetic) carriers or plates as their structural elements. Notable sectors that stand to benefit from this cutting-edge technology include:

Maritime Transport: The stress and load assessments conducted in ship hulls, holds, engine rooms, and engine shafts are critical for ensuring the structural integrity and security of seagoing vessels. The AMSA project's insights and real-time warnings contribute to enhancing the safety and reliability of maritime transportation.

Transportation of Liquid and Gaseous Fuels: Tanks used for the storage and transport of liquid and gaseous fuels must maintain structural integrity under varying stress conditions. STEHEMON's capabilities in assessing stresses and structural integrity offer valuable insights for optimizing fuel transportation systems.

Land Transport: The AMSA project's applications extend to land transportation systems, including trains and containers. Structural assessments in these contexts ensure the safe and efficient operation of the transportation infrastructure.

The impact on the national economy is direct, both through the exploitation of results by a Greek exporting company targeting the global market and the direct effects on the maritime industry, where Greece holds a leading position.

Societal and environmental benefits: The project aims to improve maritime safety, with a direct impact on human safety and environmental protection from maritime accidents. Furthermore, the implementation of predictive maintenance measures will optimize the lifespan of metallic structures, leading to resource efficiency and sustainability.

5 Conclusions

By adopting an interdisciplinary approach, integrating expertise from the domains of Materials Science, Communications, Internet of Things (IoT), and Machine Learning, the AMSA project aims to deliver a commercially exploitable method and TRL6-7 apparatus for the in-situ, non-destructive monitoring of steel structures. Specific research objectives include the adaptation of the STEHEMON method to suit the unique requirements of ship structural elements, encompassing the development of robust measurement techniques and processing algorithms. The project’s added value refers to:

- The preventative maintenance in naval environments
- The autonomous wireless sensor devices

The seamless integration of the developed subsystems into an economic, transportable, and autonomous collection and processing system (WADSA node) further enhances the system's

practicality and ease of implementation. By connecting the WADSA nodes to a decision support platform (METIS platform), the project creates avenues for immediate commercial exploitation in key transportation sectors.

6 References

[1] P. Vourna, A. Ktena, P. Tsarabaris, and E. Hristoforou, "Magnetic Residual Stress Monitoring Technique for Ferromagnetic Steels," Metals, vol. 8, no. 8, Art. no. 8, Aug. 2018, doi: 10.3390/met8080592.

[2] S. Angelopoulos et al., "Steel health monitoring device based on Hall sensors," J. Magn. Magn. Mater., vol. 515, p. 167304, Dec. 2020, doi: 10.1016/j.jmmm.2020.167304.

[3] P. Pattakos, S. Angelopoulos, A. Katsoulas, A. Ktena, and E. Hristoforou, "Magnetic Harvester for an Autonomous Steel Health Monitoring System Based on Hall Effect Measurements," Micromachines, vol. 14, no. 1, Art. no. 1, Jan. 2023, doi: 10.3390/mi14010028.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 220 – 226

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

A unified framework for object and person 3D-pose estimation with application in ancient drama

Andreas Makedonas, Ioannis Papakonstantinopoulos, Panteleimon Alkinoos Peftikoglou, Christos Theocharatos and Vassilis Tsagaris*******

Irida Labs, Patras InnoHub – Kastritsiou 4, Magoula Patras 26504, Greece

anmack@iridalabs.com, papakonstantinopoulos@iridalabs.com, peftikoglou@iridalabs.com, htheohar@iridalabs.com, tsagaris@iridalabs.com

Abstract

A framework for an interactive approach to ancient drama using state-of-the-art immersion technology in augmented reality environments is presented. The framework aims to provide tools and spaces for experiential cultural experiences of ancient theater, allowing actors, directors, designers, and spectators to engage with the genre and understand its role in a modern context. A unified network of a vision and a 3D projection subsystem is introduced, where with the use of a dedicated application all participants can interact with the play.

1 Introduction

Ancient drama is considered a national cultural treasure and a source of inspiration for modern artistic creation. Efforts have been made to promote and explore ancient drama at both national and international levels [Chessa2013, Shilkrot2014]. This includes training theater professionals, familiarizing the audience with the genre, and fostering collaborations between theater practitioners. However, there is a need for further study and understanding of the performance aspects of ancient drama, such as set design, costumes, props, and audience interaction. Reviving historical performances of ancient drama is seen as essential to highlight the functional and interactive nature of these performances. Modern technologies can be utilized to visually reconstruct parts of iconic performances and explore the interconnectedness of different elements in the performance process [Clay2014]. This approach aims to deepen the understanding of the role of these elements and their educational value for theater professionals and the audience.

The proposed framework aims to achieve several goals in relation to ancient drama, such as

- the rescue of structural elements of ancient drama and their utilization in the revival and rescue of historical performances.
- the highlighting and promotion of an experiential understanding of the function of individual elements in an ancient drama performance.

******* Corresponding Author

Additionally, it creates a dynamic technological tool that enables the reconstruction and approach to the experience of ancient drama performances in an outdoor space, without relying on material elements of the performance. Furthermore, it creates conditions for activating the viewer and offers a new way of viewing and participating in the performance process.

Overall, the proposed framework seeks to provide an interactive approach to ancient drama for actors, directors, designers, and spectators, particularly young people, allowing them to engage with the genre and understand its role and function in a modern context.

2 System

The development of the system is based on the creation of tools and spaces of experiential cultural experience of the ancient theater through state-of-the-art immersion technology in augmented reality environments. The indicative architecture of the proposed system is shown in Figure 1, and its main building blocks are: (a) the computer vision subsystem, (b) the 3D Projection subsystem, (c) the app building framework for viewers, and (d) the graphical Interface for artistic creators.

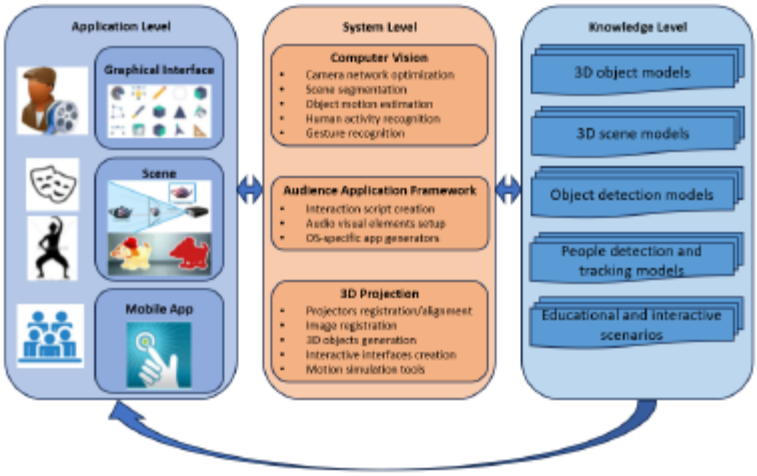


Figure 1: Prometheus system

The computer vision subsystem includes: (a) geometric dynamic representation of the theater stage for the purpose of automatic scene understanding by distributed camera networks, (b) foreground/background segmentation of the theater stage for the purpose of initial separation of moving and immovable areas, (c) recognition, separation and localization, in real-time, of the actors from the theater stage in changing lighting conditions and occlusions, (d) recognition, separation and localization, in real-time, of specific objects located within of the stage set under changing lighting and shading conditions, and (e) recognition of specific movements or poses of the actors, in real-time, in order to cause specific effects to be produced.

The selection of appropriate cameras and lenses is made based on the dimensions and possible placement positions of these systems in the field (Forest Theater), of which a snapshot with indicative dimensions is presented in Figure 2. The 3D rendering subsystem includes: (a) creation of static and dynamic virtual objects with 3D rendering techniques, (b) development of interactive interfaces of virtual objects, and (c) methods of simulating the interaction and movement of dynamic objects in a 3D model of space of the show. The rest of the paper deals with the implementation of the computer vision subsystem.

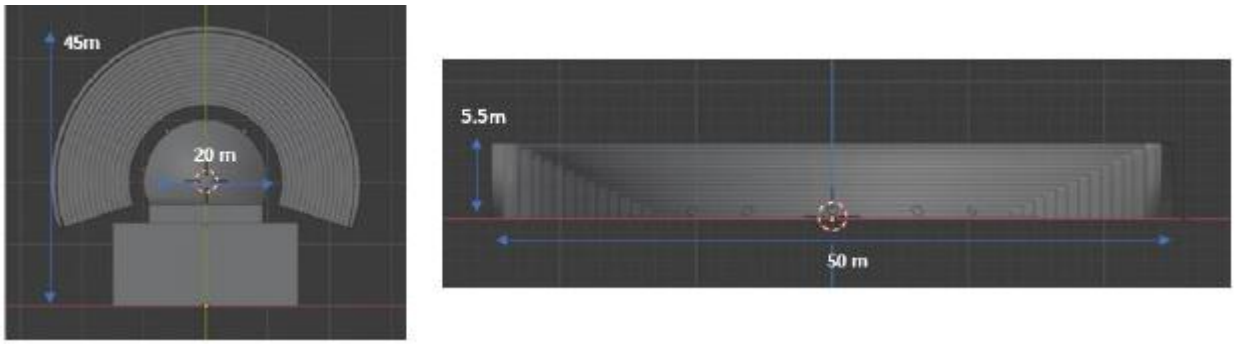


Figure 2. Forest Theater dimensions.

3 Special Objects Dataset Curation

Two different objects in terms of their shape and texture were chosen to be used in the detection learning procedures. These are the same two objects that will be used afterwards for the needs of completing the theatrical performance. More specifically, the network is trained to detect a lekythos and a shield, the characteristics of which are presented in Figure 3. The plans for the two objects have been created by the theater staff. The detection neural network was trained to detect both objects. For data collection, shots were taken according to some predefined scenarios related to the movement in the space and the lighting.

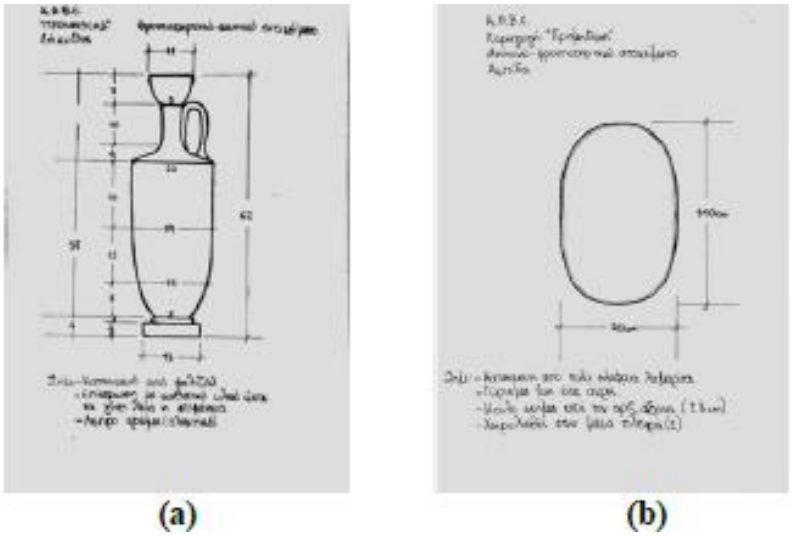


Figure 3: Drawings of (a) lekythos and (b) shield.

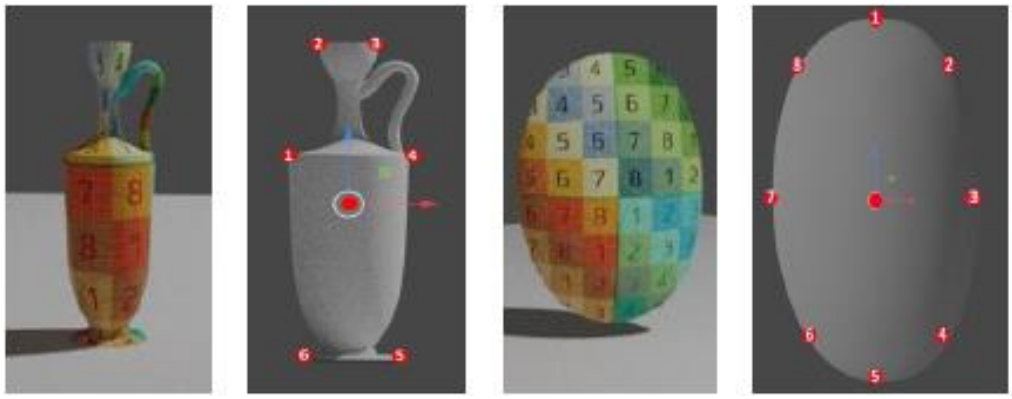


Figure 4: 3D models of the objects along with their key points.

For the synthetic data, the 3d models of the objects, as they were created by NTNG, were used (Figure 4). In both real and synthetic data captured or created, an annotation stage took place, where for every object present in the captured scene a predefined number of key points was manually set (in case of real data) or computed (for the synthetic data).

4 Pose Estimation

3D-pose estimation of an object or human is a process in which one can estimate the relative position and orientation of the object and the camera using only an RGB image. However, it can only be achieved when the 3D dimensions of the projected object are known by matching sets of corresponding 3D and 2D points. In this work, the objects required to estimate their pose, which have a specific shape and dimensions, are the shield and the lekythos. Then, the information to be extracted from the RGB images is the landmark points on the detected objects. To achieve this, a 2-stage approach using convolutional neural networks was developed.

Detection and pose estimation of the actors was performed based on the OpenPose framework [Cao2019], which is considered state-of-the-art in the domain. Regarding pose estimation of the given objects, a 2-step approach was applied. The first stage is used to locate the objects of interest in an RGB image. Using an object detector, the region of interest (ROI) is first located within the image plane. In the second stage, the ROI is used to extract specific 2D points. Then, using the 3D object's known dimensions, a set of 2D points is mapped to a set of 3D points and the palette pose is derived. Object detection is performed using the Yolo convolutional neural network (CNN) [Redmon2016], and more specifically the Yolov5s architecture, which was released only a month after its predecessor (i.e., Yolov4) by Ultralytics [Jocher2021] and is proven to have several significant improvements over existing YOLO detectors [Thuan2021]. Yolov5s consists of three parts (a) Backbone: CSPDarknet, (b) Neck: PANet and (c) Head: Yolo Layer.

The data are first fed into CSP Darknet for feature extraction and then fed into PANet for feature fusion. Finally, the Yolo Layer outputs the detection results (i.e., class, score, spatial information (detection position) and size). Additionally, the Pytorch framework was used in the training phase. The CNN is trained using annotated data and the label is described by a 2D bounding box. During inference, the CNN takes a single RGB image as input and returns a 2D bounding box.

The training process was performed using 300 epochs by using the SGD optimizer with a decaying learning rate of initial value 0.01. Images were resized to 576x576 and fed to network in batches of

four (4). Several augmentation techniques were implemented like image scaling, rotating, shearing, translating and color space transformations.

Figure 5 illustrates the training and evaluation loss curves, combined with the mean Average Precision curves.

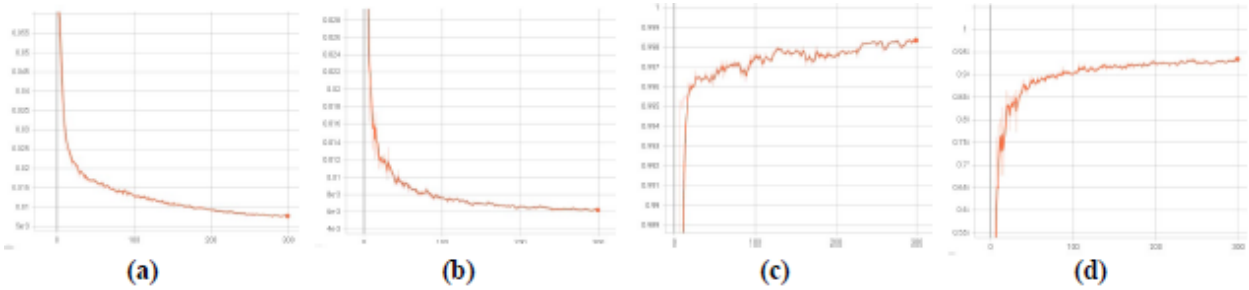


Figure 5: (a) Yolov5 Training Loss curve, (b) Yolov5 Validation Loss curve., (c) Mean Average Precision – mAP_0.5, (d) Mean Average Precision – mAP_0.5:0.95.

Landmark or Key-point detection is also developed using HRNet [Wang2020], which is a universal architecture for visual recognition and has become a standard technology for object pose estimation. Additionally, the Pytorch framework was used for the training phase. HRNet is trained using RGB images and labels from 2D landmarks in the image. In inference mode, CNN takes an image as input and returns a number of 2D landmarks points. These 2D landmarks are used in the next step to estimate the objects' pose.

In this stage, the training process was performed for 500 epochs by using Adam Optimizer with a learning rate equal to 0.0001. The image size was 256x256 and the batch size 16. Rotation and scaling were also applied. Figure 6 illustrates the training and evaluation loss curves.

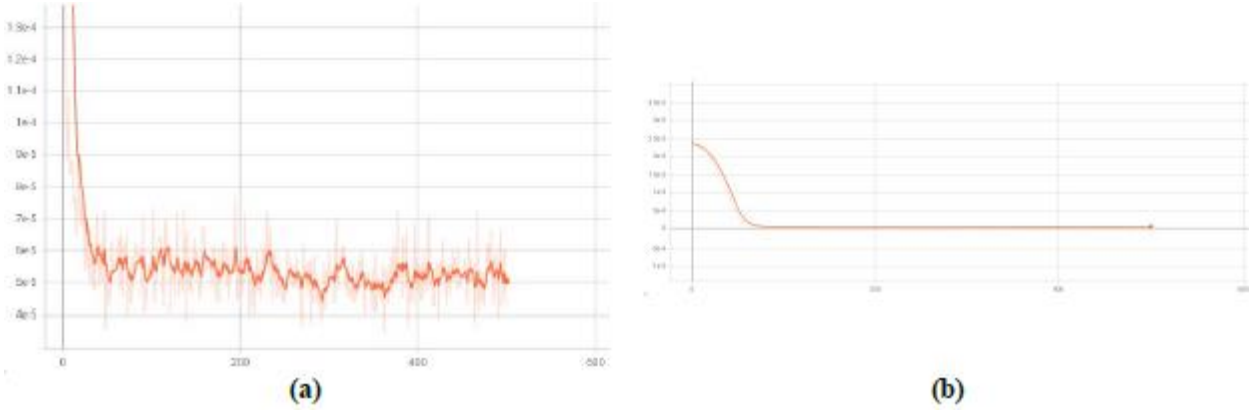


Figure 6: (a) HRNet Training Loss curve, **(b**) HRNet Validation Loss curve.

5 Results

We followed a two-stage framework for the object/person detection and their pose estimation. In the first stage, Yolov5s acted as the object detector while OpenPose initiated for person detector, identifying the object/actor of interest within the given image. The initial stage helped to narrow down the focus to the relevant regions for further analysis.

In the second stage HRNet is utilized to detect and localize key points on the identified object, allowing for the creation of the object's pose. By locating these key points, such as joints or key landmarks, the network is able to infer the position, orientation, and scale of the object, resulting in an accurate estimation of its pose. This two-stage approach facilitates a more robust and comprehensive solution to pose estimation tasks, enhancing the understanding and analysis of object configurations in computer vision applications.

Detection results for both actor and object detection are shown in Figure 7.

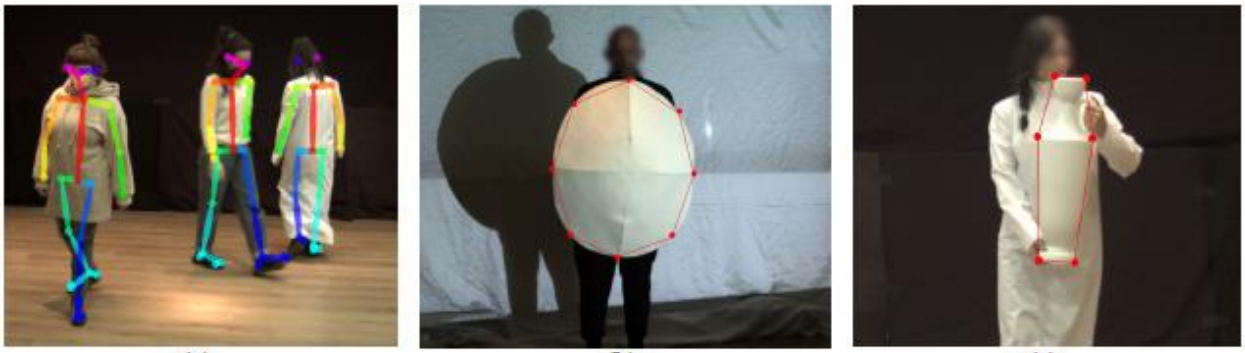


Figure 7: Detection and pose estimation results.

The seamless integration of the developed subsystems into an economic, transportable, and autonomous collection and processing system (WADSA node) further enhances the system's practicality and ease of implementation. By connecting the WADSA nodes to a decision support platform (METIS platform), the project creates avenues for immediate commercial exploitation in key transportation sectors.

6 Conclusions

A unified framework for object and person 3D-pose estimation in the context of ancient drama is presented. It explores the use of modern technologies, such as augmented reality and computer vision, to reconstruct and connect different elements of the performance process in ancient drama. The framework aims to deepen the understanding and educational value of ancient drama for theater professionals and the audience. It highlights the experiential understanding of individual elements in ancient drama performances and offers a dynamic technological tool for the reconstruction and approach to ancient drama performances in outdoor spaces. The proposed system includes components such as computer vision subsystem, 3D projection subsystem, app building framework for viewers, and graphical interface for artistic creators. The computer vision subsystem is further analyzed, in terms of system architecture, data creation and curation.

7 Acknowledgments

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the Call Special Actions "Aquaculture" - "Industrial materials" - "Open innovation in culture" (project code: Τ6ΥΒΠ-00490, MIS code: 5052072).

8 References

[1] Cao, Z., Martinez, G. H, Simon, T., Wei, S., and Sheikh, Y. A. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[2] Chessa M., Garibotti M., Canessa A., Gibaldi A., Sabatini S.P., Solari F., "Veridical Perception of 3D Objects in a Dynamic Stereoscopic Augmented Reality System". In: Csurka G., Kraus M., Laramee R.S., Richard P., Braz J. (eds) Computer Vision, Imaging and Computer Graphics. Theory and Application. Communications in Computer and Information Science, vol 359. Springer, Berlin, Heidelberg, 2013.

[3] Clay, A., Domenger, G., Conan, J., Domenger, A., and Couture, N. (2014) Integrating Augmented Reality to Enhance Expression, Interaction & Collaboration in Live Performances: A Ballet Dance Case Study. *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*.

[4] Jocher, G., et. al., ultralytics/yolov5: v6.0 -YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support, https://doi.org/10.5281/zenodo.5563715, Oct. 2021.

[5] Redmon, J., Divvala, S. K., Girshick, R. B, and Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *Computer Vision and Pattern Recognition*.

[6] Shilkrot R., Montfort N., Maes P., "nARratives of Augmented Worlds", In proceedings of 2014 IEEE International Symposium of Mixed and Augmented Reality-Media, Art, Social Science, Humanities, 2014.

[7] Thuan, D., Evolution of YOLO algorithm and YOLOv5: the state-of-the-art object detection algorithm, 2021.

[8] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., and Xiao B. (2020). Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 227 – 228

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Deep Learning Architectures for Greek Orthodox Church Hymns Recognition: The Ymnodos Project[†††††††]

Lazaros A. Iliadis[1], Nikos Tsakatanis[1], Sotirios P. Sotiroudis[1], Achilles D. Boursianis[1], Konstantinos-Iraklis D. Kokkinidis[2], Georgios Patronas[3], Pavlos Serafeim[3], Maria S. Papadopoulou[4], and Sotirios K. Goudos[1]

[1] *ELEDIA@AUTH, School of Physics, Aristotle University of Thessaloniki, Thessaloniki, Greece*
[2] *Dept. of Applied Informatics, University of Macedonia, Thessaloniki, Greece*
[3] *Dept. of Music Science and Art, University of Macedonia, Thessaloniki, Greece*
[4] *Dept. of Information and Electronic Eng., International Hellenic University, Sindos, Greece*

liliadis, ntsakata, ssoti, bachi, sgoudo@physics.auth.gr, kostas.kokkinidis@uom.edu.gr, gpatronas, serapavl@uom.edu.gr, mspapa@ihu.gr

Abstract

In this work a technical report of the Ymnodos project is presented. The Ymnodos project (Project code: KMP6-0078938) combines artificial intelligence methodologies, audio signal processing, and computer vision techniques to provide automatic identification and characterization of cultural items especially Greek Orthodox Church hymns.

1 Introduction

The study of cultural and religious heritage sectors involves dealing with a vast and diverse amount of data. Traditional techniques have been supplemented with machine learning (ML) approaches in recent times. Deep learning (DL) algorithms, particularly those based on computer vision (CV), have been employed to extract valuable information from available data. Specifically, convolutional neural networks (CNNs) have been proven successful in image classification problems due to their ability to apply filters for feature extraction, leveraging the grid topology of images [1].

Greek Orthodox Church hymns are an integral part of the Orthodox Christian tradition, characterized by vocal performances without musical accompaniment. Recognizing and classifying these hymns using ML methods poses a significant challenge. The distinctive nature of these vocal compositions, reliant on chanting voices, and their climaxes differing from Western hymns, further complicates the task [2]. This study addresses this challenge by assembling a corpus of 23 Greek Orthodox Church hymns, each comprising over 200 samples. The audio data is transformed into Mel-spectrograms,

[†††††††] This research was carried out as part of the project "Recognition and direct characterization of cultural items for the education and promotion of Byzantine Music using artificial intelligence" (Project code: KMP6-0078938) under the framework of the Action "Investment Plans of Innovation" of the Operational Program "Central Macedonia 2014 2020", that is co-funded by the European Regional Development Fund and Greece.

which are then used as input to a novel shallow CNN for accurate classification. Notably, this research project represents the first application of DL techniques to the recognition of Greek Orthodox Church hymns, to the best of the authors' knowledge.

2 Results

Preliminary results related to the Ymnodos project were carried out during the period 2022-2023. Several DL models were designed and trained on our dataset. Furthermore, transfer learning approaches have been also utilized. In 2023 a custom CNN with two convolution layers was compared to several state-of-the-art DL models, achieving better performance in terms of accuracy. Furthermore, the proposed model was lightweight requiring few computational resources [3].

In this paper, a custom CNN with three convolution layers, three max-pooling layers, using the leaky rectified linear unit function (Leaky ReLU) as an activation function, and two fully-connected layers, are trained to identify Greek Orthodox Church hymns. To study the generalization capabilities of our approach, 5-fold cross-validation is used. The results are presented in Table 1 and Table 2.

Training Accuracy	Validation Accuracy	Test Accuracy
95.67	95.92	94.04

Table 1: Performance of proposed DL model: Accuracy

Precision	Recall	F1-score
0.95	0.93	0.94

Table 2: Performance of proposed DL model: Statistical measures

3 Conclusions

The main objective of the Ymnodos project is the automatic recognition and characterization of Greek Orthodox Church hymns, using AI methodologies. Preliminary results show that DL models can achieve satisfactory results. Future research includes the development of novel DL approaches, the utilization of transfer learning techniques, and the creation of a mobile application that will be tested in real-world scenarios.

4 References

[1] Lazaros Alexios Iliadis, Sotirios P. Sotiroudis, Kostas Kokkinidis, Panagiotis Sarigiannidis, Spiridon Nikolaidis, and Sotirios K. Goudos. Music deep learning: A survey on deep learning methods for music processing. In 2022 11th International Conference on Modern Circuits and Systems Technologies (MOCAST), pages 1–4, 2022.

[2] Lazaros Moysis, Lazaros Alexios Iliadis, Sotirios P. Sotiroudis, Achilles D. Boursianis, Maria S. Papadopoulou, Konstantinos-Iraklis D. Kokkinidis, Christos Volos, Panagiotis Sarigiannidis, Spiridon Nikolaidis, and Sotirios K. Goudos. Music deep learning: Deep learning methods for music signal processing—a review of the state-of-the-art. IEEE Access, 11:17031–17052, 2023.

[3] Nikolaos Tsakatanis and et al. Greek orthodox church hymns recognition using deep learning techniques. In 2023 12th International Conference on Modern Circuits and Systems Technologies (MOCAST), pages 1–4, 2023.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 229 – 230

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

An IoT System for Innovative Cultural Experience

Christos Sad, Aggelizi Ziaka and Kostas Siozios

Aristotle University of Thessaloniki, Thesasaloniki, GR.

ksiop@auth.gr

Abstract

In an increasingly interconnected world, cultural exchange and understanding have gained paramount importance. This study presents a novel and innovative product designed to enhance cultural experiences through a fusion of advanced technology and immersive engagement. Leveraging cutting-edge augmented reality (AR) and Internetof- Things (IoT) technologies, the proposed product aims to offers an unparalleled opportunity for individuals to engage with and understand cultures in an immersive and interactive manner. The paper elucidates the development, features, and applications of this pioneering product, emphasizing its ability to revolution traditional education, tourism, and cultural preservation efforts. By fusing technological innovation with cultural exploration, the product promises to redefine how people perceive, appreciate, and connect with the world's diverse cultural tapestry.

1 Introduction

Intangible Cultural Heritage consists of intangible features of a culture, which are often preserved through cultural and religious customs, and are reflected in practices, representations, knowledge, and techniques. Some of the intangible traditions are preserved thanks to the oral transmission and collective memory, but also to their material cultural references. Due to the peculiarity of the intangible cultural heritage, its preservation and recording contribute to the beneficial updating and promotion of modern culture internationally. This proposal strengthens the effort made by organizations such as UNESCO, the UN, the Mediterranean Forum, to save and highlight the innovative interaction between monuments and the intangible narratives about them. The region of Eastern Macedonia and Thrace is of particular interest due to the long and rich coexistence of three cultural and religious traditions, Christianity, Islam, and Judaism.

Due to the importance of cultural heritage, more and more products are appearing on the market in order to support actions related to the Museums of the future. Despite the intense interest, the tour applications usually have prefabricated and static content, the same for all categories of visitors. Contrary to common practice, ICE seeks to design and implement an innovative system for the promotion of cultural heritage. More specifically, in the framework of the ICE project, it is proposed: (i) the creation of an innovative product called ICE (Innovative Cultural Experience) for the active and experiential tour, as well as for the promotion of the cultural or commercial product, and (ii) the

development of a content aggregator mechanism for the enrichment of the Augmented Reality material for the highlighted cultural or commercial exhibit. The final ICE product will be addressed to exhibition, educational and other spaces, with the possibility of providing a comprehensive exhibit presentation service, with dynamic content that will be adapted to the preferences, needs and profile of users, through a knowledge management system.

2 Architecture of the Proposed System

The approach of the project is guided by the technology and all the partners that will contribute to the development of the innovative system that is proposed. The ICE system is an advanced implementation (beyond the state of the art) of the Transparent multi-touch Window, in which each exhibit that will be entered, will be able to display to the user / visitor additional information through Augmented Reality in the form of video (simple or 360) and audio (admin enabled).

ICE's innovation includes functional extensions that are not available in the originally mentioned Transparent multi-touch Window technology. More specifically, its innovation is based on the fact of the cognitive background of technology (can be used in a variety of knowledge areas, education, culture, tourism, commercial applications), 360 video viewing (pre-stored or dynamically generated by the user), providing API for connection with external applications-platforms, the use of sensors (e.g., motion), to include meta-information in the descriptions of the exhibits utilizing Internet of Things technologies.

Figure 1: The ICE platform for the presentation of exhibits with Transparent multi-touch Window technology.

3 Acknowledgment

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T2EDK-02564).

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 231 – 239

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

An Improved Algorithm for Lithium-ion Battery Equalization and Energy Support in Electric Motor Drive Applications

Nikolaos Jabbour[1], Evangelos Tsioumas[1], Dimitrios Papagiannis[1] and Christos Mademlis[1]

[1] *Aristotle University of Thessaloniki, Thessaloniki, GR*

njabbour@auth.gr, etsioumas@auth.gr, dipapagi@auth.gr, mademlis@auth.gr

Abstract

An improved battery management system (BMS) is proposed in this research paper that provides a combined approach involving a non-dissipative direct cell-to-cell equalization algorithm and an energy support algorithm. The primary objective is to safeguard the longevity of a lithium-ion battery pack. To achieve this, an auxiliary energy storage system, such as supercapacitors (SCs) or batteries, is employed to temporarily store or recover electric energy from any cell within the primary battery pack. The equalization algorithm specifically controls the magnitude of each battery balancing current by considering factors such as State of Health (SOH), the State of Charge (SOC), the residual available energy of each battery cell, the energy losses of the matrix switch converter during steady-state operation, and the residual available capacity. Additionally, the energy support algorithm provides assistance to weak or problematic battery cells during dynamic conditions, with the aim to prolong the battery bank's safe operating duration. The practicality and efficacy of this combined equalization and energy support method have been validated through comprehensive simulation analysis and selective simulation results are presented in the paper.

1 Introduction

Over the years, batteries have emerged as the predominant choice for energy storage across various applications, including solar power systems, wind power systems, and electric vehicles (EVs) [1]. In these applications, there is high demand for energy storage systems that can exhibit rapid charge and discharge rates, extended lifespan, and superior energy density [2]. Recent research findings confirm that lithium-ion batteries fulfill the above requirements, surpassing other battery types such as nickel and lead acid cells, thanks to their inherent advantages [3].

The control strategies employed to achieve equalization in a Battery Storage System (BSS) can be classified into two categories: dissipative and non-dissipative, based on the form of energy transfer [4]. Dissipative schemes involve topologies that use resistors to dissipate excess energy from cells with high SOC. These schemes offer several advantages such as low complexity, low cost, and high reliability.

However, the energy that is equalized is wasted, leading to the reduction of the overall BSS efficiency [5]. On the other hand, non-dissipative control schemes achieve battery equalization by exchanging

energy between cells [6]. Although these schemes are more complex, they provide faster equalization and increased efficiency for the BSS [7]. As a result, non-dissipative equalization methods and topologies have gained significant research interest in recent years [8].

The objective of this research paper is to present an improved non-dissipative direct cell-to-cell equalization method and energy support algorithm specifically designed for lithium-ion battery storage systems used in battery-powered electric motor drives. The proposed method comprises two subalgorithms: the equalization algorithm and the energy support algorithm. The equalization algorithm utilizes the particle swarm optimization (PSO) technique to determine the optimal participation of cells in the equalization process. Factors such as the operation of the motor drive, distribution of residual energy among the cells, and energy losses are taken into account. Additionally, the algorithm regulates the balancing current for each cell, ensuring effective equalization. The energy support algorithm focuses on evaluating the SOH of the battery cells. It initiates the energy assistance process for cells with the lowest SOH using the auxiliary energy storage system, particularly during high dynamic conditions. The energy support algorithm aims to both safeguard the battery lifespan and extend the time interval during which the battery bank remains within the safe operating zone. Selective simulation results are presented and discussed to validate the effectiveness and operational enhancements achieved by the proposed integrated method. The outcomes of the study demonstrate the improved performance of the BMS and substantiate the benefits of the proposed approach.

2 Overview of the proposed system

The equalization and energy support circuit for each battery stack is illustrated in Figure 1. The circuitry consists of a bidirectional dc-dc converter and a matrix power switch system, both of which are controlled by an integrated control system. The dc-dc converter has the capability to change the energy flow direction. It can either store energy from a battery cell into the auxiliary energy storage system or retrieve energy from the auxiliary energy storage and deliver it to weak batteries. A system is implemented to detect weak cells, which involves online estimation of the SOH and SOC for each battery module.

Figure 1 showcases the matrix power switch system, which operates at a lower switching frequency compared to the *dc-dc* converter [9]. This system is composed of two sets of MOSFET power switches: the battery cell switches (*BCSW*) and the polarity switches (*PSW*) [10]. These switches facilitate the connection of each battery cell to the bidirectional *dc-dc* converter, enabling the energy transfer between the cells and the converter.

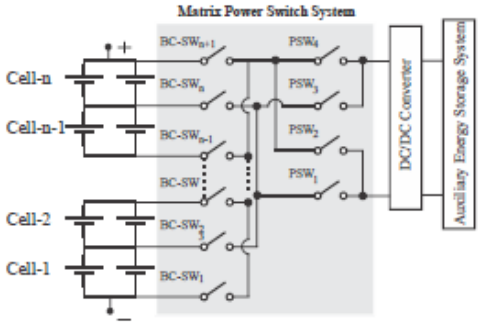


TABLE I
PARAMETERS OF THE BATTERY AND THE SC-BASED BMS

SC-bank equivalent capacitance	C_{sc} = 200 F	Lithium-Ion battery parameters:	
SC-bank rated voltage	8.1 V	Nominal Capacity	7Ah
SC-bank max. continuous current (ΔT=15°C)	50.0A (RMS)	Nominal Voltage	3.7V
		Internal Resistance	5.2mΩ

Figure 1: The circuit of the matrix switch-based converter.

3 Algorithm of the proposed system

A *BSS* can experience performance degradation when a significant residual capacity discrepancy occurs among its cells. However, it is important to note that the advantage of eliminating cell inconsistency may not always outweigh the energy losses incurred in the equalization circuit. In the literature, a commonly employed equalization strategy involves setting a predefined value of *SOC* and regulating the cells to be equalized based on this threshold. Typically, this threshold value is arbitrarily set at 10%, with the upper and lower limits defined as the highest and lowest *SOC* cells, respectively [11].

a. Theoretical Analysis

The *SOC* parameter needs to be defined in order to accurately describe the battery's energy level. Consequently, the initial value of the SOC for the *i*-th cell can be expressed as follows

$$SOC_{init,i} = f\left(V_{ocv,i}\right) \tag{1}$$

where *f* is the mapping function of the *SOC* and the open circuit voltage. Then, the value of each cell's *SOC* at the time *t* is defined by

$$SOC_{i,t} = SOC_{init,i} + \frac{\int_0^t I_i dt}{Q_{nom,i}} \cdot 100\% \tag{2}$$

with $Q_{nom,i}$, the battery module's nominal capacity is denoted. The current of the *i*-th cell is given by

$$I_i = I_{pack} - I_{dal,i} \tag{3}$$

where I_{pack} is the current of the battery pack and the $I_{dal,i}$ is the balancing current of the i-th cell.

The residual capacity of the i-th cell is given by

$$Q_{res,i} = SOC_i \cdot Q_{nom,i} \tag{4}$$

and the residual available capacity of the battery pack is

$$Q_{RAC} = \min_{l<i<N} \quad Q_{res,i} \tag{5}$$

During discharge mode of the battery pack, the residual available capacity of each cell is constrained to Q_{RAC}. As a result, the *SOC* that represents the minimum level of charge for a cell within the pack can be determined as follows

$$SOC_{0,i} = SOC_{init,i} - \frac{Q_{RAC}}{Q_{nom,i}} \tag{6}$$

and the residual available energy is calculated by

$$E_{RAE,i} = \left(\int V_{OCV,i}\,(t)\, I_i dt - \int I_i^2\, R_i dt\right|_{[SOC_{0,i}, SOC_{init,i}]} \tag{7}$$

Another important parameter that an equalization system should consider is the energy losses in the equalization circuit. The energy losses at the boost mode (where energy is transferred from the *i*-th battery to the auxiliary energy storage system) is given by

$$E_{loss,boost,i} = \left(1 - n_{boost}(I_{bal,i})\right) \cdot E_{cell,i} \tag{8}$$

and the energy losses at the buck mode (where energy is transferred from SC-bank to the *i*-th weak cell) is

$$E_{loss,buck,i} = \left(1 - n_{buck}(I_{bal,i})\right) \cdot E_{AUX,i} \quad (9)$$

where $I_{bal,i}$ is the *i*-th cell balancing current and the n_{boost} and n_{buck} are the efficiency of the equalization circuit at the boost and buck modes, respectively, considering both the efficiency of the bidirectional *dc-dc* converter and the matrix power switch system. In order to be classified the *i*-th as equalized, the $E_{cell,i}$ energy should be obtained by that cell and the $E_{AUX,i}$ the energy should be provided from the auxiliary energy storage system to that cell.

From (8) and (9), it is observed that the n_{boost}t and n_{buck} depend directly on the magnitude of the balancing currents, and therefore, the algorithm should regulate them to minimize the energy losses. The total energy losses at the end of the equalization can be defined by

$$E_{l,total} = \sum_{i=1}^{N_1} E_{loss,boost,i} + \sum_{i=1}^{N_2} E_{loss,buck,i} \quad (10)$$

where N_1 denotes the number of the battery cells that give energy to the auxiliary energy storage system and N_2 denotes the number of battery cells that absorb energy from it. Thus, the total number of battery cells can be expressed as, $N=N_1+N_2+N_3$, where N_3 denotes the number of battery cells that does not participate in the balancing process.

b. Proposed integrated strategy

The objective is to achieve an optimal balance between reducing equalization energy losses and maximizing the residual available energy of the battery pack by setting the appropriate values for the balancing current, denoted as $I_{bal,i}$. Consequently, the cost function for the proposed Particle Swarm Optimization (PSO)-based strategy can be expressed as follows

$$\begin{cases} J = w_1 \cdot \frac{E_{RAE,max}}{E_{RAE,total}} + w_2 \cdot \frac{E_{l,total}}{E_{l,max}} \\ E_{RAE,total} = \sum_{i=0}^{N} E_{RAE,i} \\ w_1 + w_2 = 1 \end{cases} \quad (11)$$

where w_1 and w_2 are the weighting factors for the residual capacity and the energy losses, respectively. The $E_{RAE,max}$x and the $E_{l,max}$ are the maximum residual energy and energy losses in the equalization circuit, respectively.

The proposed integrated strategy is illustrated in Figure 2. The *BMS* system initiates the operation of the proposed algorithm by sending a signal and determining whether the application is in dynamic or non-dynamic operation. This determination is made by evaluating the absolute value of the *dc*-link current, denoted as I_{dc}. If the I_{dc} is found to be lower than a predefined threshold value, $I_{dc,th}$, it indicates that the system is in a non-dynamic mode. In this case, the equalization PSO-based algorithm is activated. On the other hand, if the I_{dc} exceeds the threshold value, it indicates that the system is in dynamic operation.

In the non-dynamic mode, the algorithm periodically monitors the time-dependent variable *SOH*flag. Once this variable reaches a certain value, the *PSO*-based algorithm is triggered to search for the optimal values of the balancing currents. *PSO* is a metaheuristic search method known for its simplicity and effectiveness in finding near-optimal solutions quickly [13]. The *PSO* algorithm aims to

minimize the cost function *J*, and once the cost function is minimized, the equalization process is activated using the optimal values of the balancing currents ($I_{bal,i}$) for each cell (*i*-th cell).

In the case of dynamic operation, the energy support algorithm is activated. The energy support control scheme is depicted in Figure 3. The first *PI* controller receives inputs of the reference voltage and the measured voltage of the weak *i*-th battery, and it adjusts the reference value for the weak battery balancing current based on the error between these two voltages. The real balancing current is then measured, and the second *PI* controller determines the duty cycle of the switches in the *dc-dc* converter. It is important to note that, in the energy support process, the voltage value of the weakest cell is chosen as the variable instead of the *SOC* or the residual available capacity. This choice is made because the cut-off voltage of a cell has a significant impact on the power performance of the entire battery pack.

The decision for initializing the energy support process is obtained with respect to the amplitude of the difference between the lowest *SOH* cell ($SOH_{c,1}$) and the highest *SOH* cell ($SOH_{c,n}$). If the error is greater than a predefined value (SOH_{th}), the reference voltage value of the voltage *PI* controller is set to the nearest cell voltage value *V*cell,i that satisfies the condition $V_{cell,l} \leq V_{cell.i} \leq V_{cell,j}, \forall\, j \neq i$.

4 Simulation results

To validate the effectiveness and operational improvements of the proposed integrated system for electric motor drive applications, a simulation model was used, consisting of a 6-cell Lithium-ion battery pack. The parameters of the battery cells and the SC-based auxiliary energy storage system are provided in Table I.

Figure 4 illustrates the performance of the proposed *BMS* during both steady-state and dynamic operation of the motor drive. The capacity of the six battery cells is specified as follows: 7Ah, 6.95Ah, 6.9Ah, 6.85Ah, 6.8Ah, and 6.2Ah. The voltage of the SC-bank is set at 7V. The proposed algorithm is initiated when the *SOC* reaches 12% and it terminates when one of the cells reaches the cut-off voltage of 2.775V.

In the proposed algorithm, the battery pack is considered balanced when the difference between the highest and lowest residual available capacity is less than 0.002Ah. This predefined value is selected by the user of the application, while the charging and discharging balancing currents of the battery cells are determined by the proposed algorithm. As shown in Figure 4, during the 80-second equalization process, the residual available capacity of the proposed BMS reaches 0.459Ah prior to dynamic operation. The *SOC* of the backup *SC*-bank fluctuates within a small range. The proposed *BMS* effectively assists the operation of the battery cell with the lowest *SOH* and ensures satisfactory performance during the dynamic operation, which concludes at 99.5 seconds when the voltage of the cell reaches the cut-off voltage (2.775V). It is noteworthy that the *SOC* of the *SC*-bank remains nearly constant at 60.125% throughout the equalization process, making it ready to provide energy support to a weak battery cell during dynamic conditions.

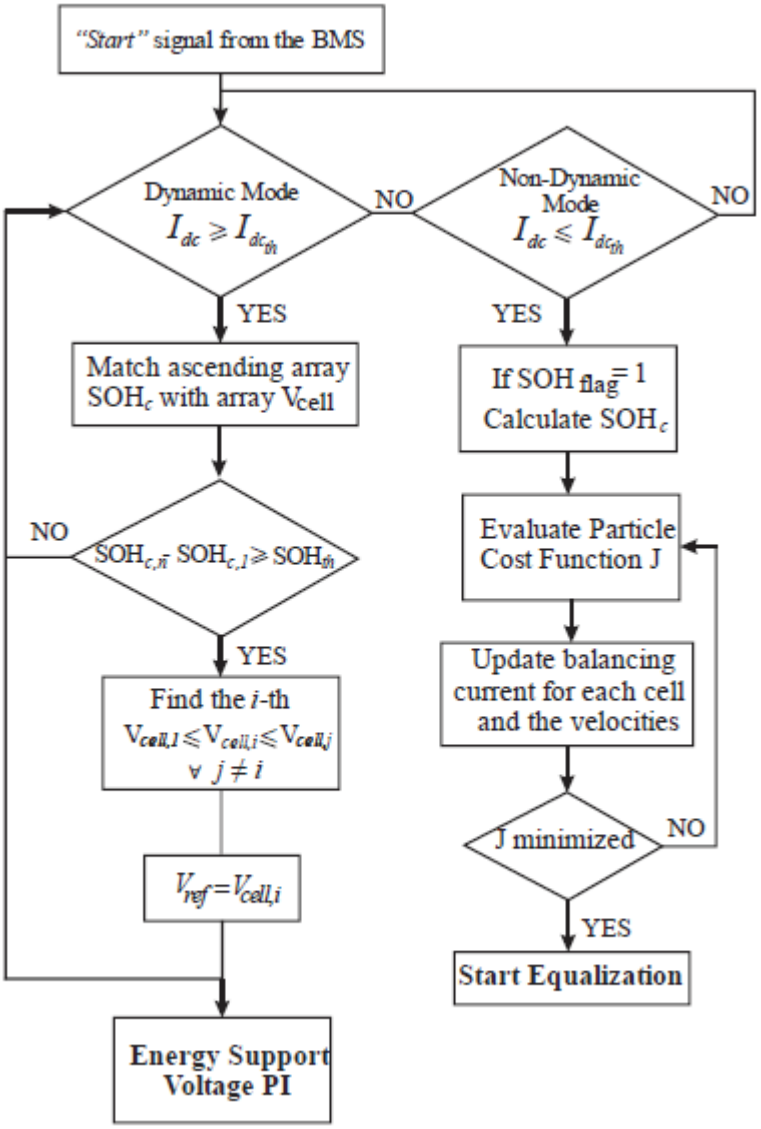


Figure 2: (Flowchart of the proposed equalization and energy support algorithms.

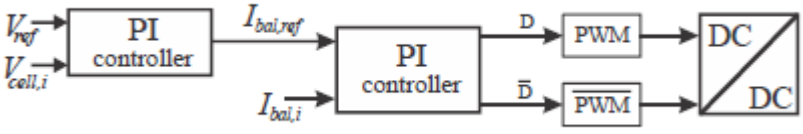


Figure 3 The energy support control algorithm that regulates the *dc/dc* converter.

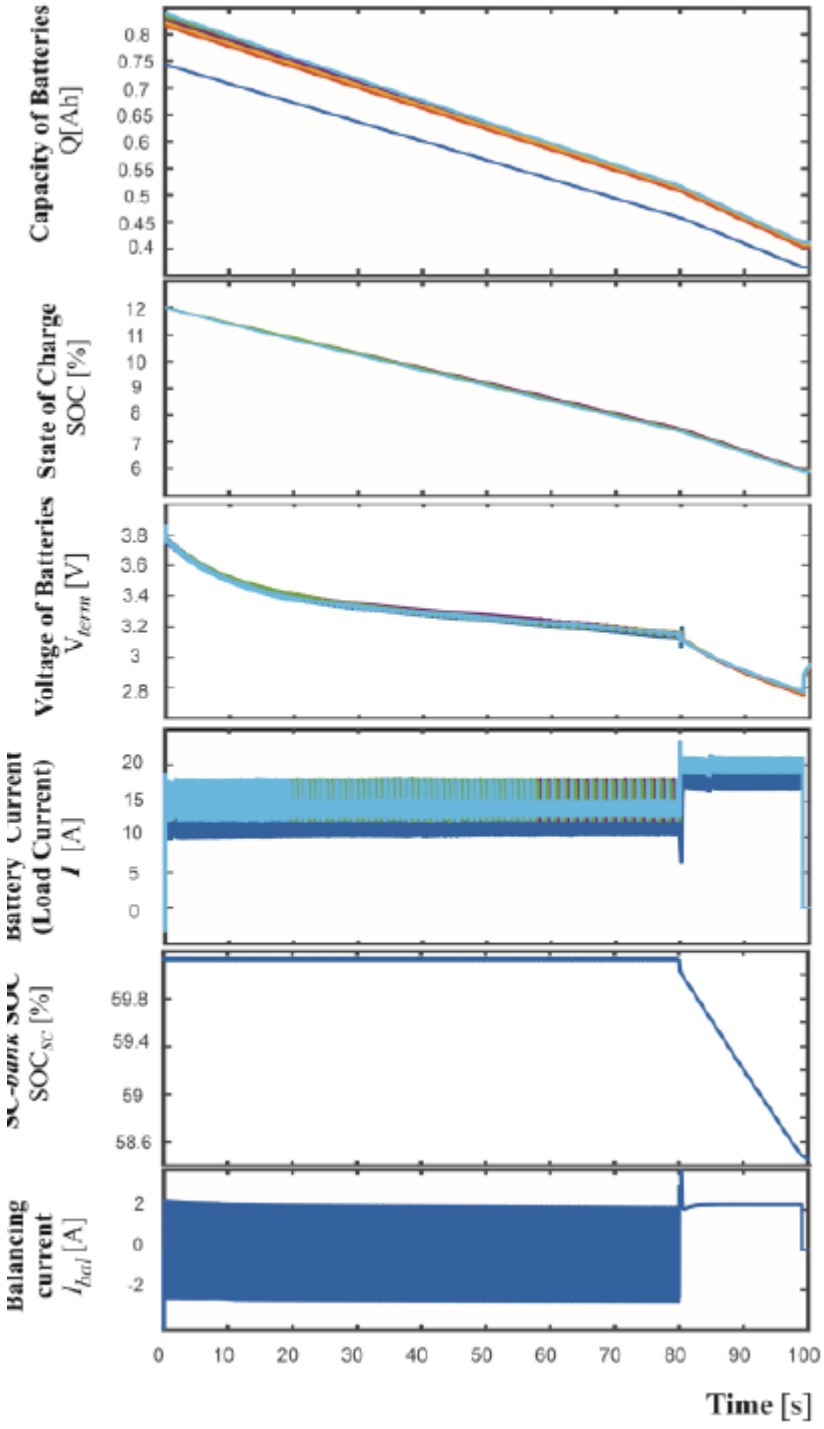


Figure 4 Performance of the proposed algorithm at a 6-cell lithium-ion battery pack with a SC-bank as auxiliary energy storage system.

5 Conclusion

This paper presents an enhanced battery management system (BMS) for lithium-ion batteries in electric motor drives. It combines a non-dissipative equalization algorithm and an energy support algorithm to improve the battery performance, protect battery lifespan, minimize energy losses, and provide support to weak battery cells during high dynamics. Selective simulation results have been presented to validate the effectiveness and operational improvements of the proposed integrated BMS method.

6 Acknowledgments

This work has been funded within the framework of the Operational Programme “Central Macedonia” of the PA 2014-2020, Innovative Investment Plans, and co-financed by the Greek State and the European Union and, in particular, by the European Regional Development Fund (ERDF).

7 References

[1] N. Ghaeminezhad, Q. Ouyang, X. Hu, G. Xu and Z. Wang, "Active Cell Equalization Topologies Analysis for Battery Packs: A Systematic Review," IEEE Trans. Power Electron, vol. 36, no. 8, pp. 9119 - 9135, 2021.

[2] X. Hu, W. Liu, X. Lin and Y. Xie, "A Comparative Study of Control-Oriented Thermal Models for Cylindrical Li-Ion Batteries," IEEE Trans. Transp. Electrific., vol. 5, no. 4, pp. 1237 - 1253, 2019.

[3] A. E. Mejdoubi, H. Chaoui, H. Gualous, P. V. D. Bossche, N. Omar and J. V. Mierlo, "Lithium-ion batteries health prognosis considering aging conditions," IEEE Trans. Power Electron., vol. 34, no. 7, p. 6834–6844, 2019.

[4] M. Koseoglou, E. Tsioumas, N. Jabbour and C. Mademlis, "Highly Effective Cell Equalization in a Lithium-Ion Battery Management System," IEEE Trans. Power Electron., vol. 35, no. 2, pp. 2088 - 2099, 2020.

[5] F. Altaf, B. Egardt and L. J. Mårdh, "Load Management of Modular Battery Using Model Predictive Control: Thermal and State-of-Charge Balancing," IEEE Trans. Control Syst. Technol., vol. 25, no. 1, p. 47–62, 2017.

[6] N. Jabbour, E. Tsioumas, M. Koseoglou and C. Mademlis, "Highly Reliable Monitoring and Equalization in a Hybrid Energy Storage System with Batteries and Supercapacitors for Electric Motor Drives in Building Applications," in IEEE 4th Southern Power Electronics Conference (SPEC), Singapore, 2018.

[7] S. Yarlagadda, T. T. Hartley and I. Husain, "A Battery Management System Using an Active Charge Equalization Technique Based on a DC/DC Converter Topology," IEEE Trans. Ind. Appl., vol. 49, no. 6, pp. 2720 - 2729, 2013.

[8] Z. Wei, F. Peng and H. Wang, "An LCC-Based String-to-Cell Battery Equalizer With Simplified Constant Current Control," IEEE Trans. Power Electron., vol. 37, no. 2, pp. 1816 - 1827, 2022.

[9] M. Raeber, A. Heinzelmann and D. O. Abdeslam, "Analysis of an Active Charge Balancing Method Based on a Single Nonisolated DC/DC Converter," IEEE Trans. Ind. Electron., vol. 68, no. 3, pp. 2257 - 2265, 2020.

[10] S. Holland, "16-cell lithium-ion battery active balance reference design," Texas Instrum., Dallas, TX, USA, 2016.

[11] S. Jinlei, LiuWei, T. Chuanyu, W. Tianru, J. Tao and T. Yong, "A Novel Active Equalization Method for Series-Connected Battery Packs Based on Clustering Analysis With Genetic Algorithm," IEEE Trans. Power Electron., vol. 36, no. 7, pp. 7853-7865, 2021.

[12] A. Hauser and R. Kuhn, "12 - Cell balancing, battery state estimation, and safety aspects of battery management systems for electric vehicles," Advances in Battery Technologies for Electric Vehicles, Woodhead Publishing Series in Energy, pp. 283-326, 2015.

[13] J. Kennedy and R. Eberhart, "Particle swarm optimization," in Proc. IEEE Int. Joint Conf. on Neural Networks, Perth, WA, Australia, 1995.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 240 – 246

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Design of an Autonomous Wireless Electric Field sensor for maritime applications: the EFOS project

Ioannis Masklavanos[a]*, Vasiliki Naskari[a], Christos Koutsos[a], Fotios Vartziotis[a], Gregory Doumenis[a], Stelios Siskos[b], Achilleas Bardakas[c], Apostolos Segkos[c], Christos Tsamis[c], Christos Papakis[d] and George Koukas[e]

[a]*Autonomous Systems Laboratory, Dept. Informatics and Telecommunications, University of Ioannina, Arta, Greece*
[b]*Electronics Lab, Physics Dept., Aristotle University of Thessaloniki, Thessaloniki, Greece*
[c]*Institute of Nanoscience and Nanotechnology, NCSR "Demokritos", Athens, Greece*
[d]*METIS Cyberspace Technology, Athens, Greece*
[e]*VECTOR Technologies Ltd, Athens, Greece*

i.masklavanos@uoi.gr

Abstract

In this paper we describe the design and development of an electrostatic field sensor, applicable for short-term forecasting of weather anomalies in naval environments. We present the problem and describe the development approach, focusing on the hardware integration of the developed sensor into a fully functional and energy autonomous naval weather station. The system can scavenge ambient energy for its operation -using commercial PV cells and experimental triboelectric generators. Further, it can connect wirelessly to the ship's communication infrastructure and relay data to an indigenous IoT platform for logging and analysis. The presented work has been carried out in the context of the EFOS RTD project (T2EDK- 00350).

1 Introduction

Weather forecasts are usually derived from intricate arithmetic models, meticulously developed by researchers, and employed by international agencies and corporations. These sophisticated models seamlessly assimilate and process an extensive array of data originating from various sources, including local weather stations, satellites, radars, weather balloons, and even man-made observations. By ingeniously integrating these diverse datasets, the forecasting process can effectively predict weather conditions for a significant duration of up to ten days into the future. (Florita & Henze, 2009)

Nevertheless, such models often fail to predict short-term and/or local events -unless dense spatial and temporal measurements are available. Short-term-forecasting (Nowcasting) relies -among other- on real-time data obtained from local sources. By analyzing this up-to-the-minute information,

However, the energy that is equalized is wasted, leading to the reduction of the overall meteorologists can make precise and localized forecasts for phenomena like thunderstorms, heavy rainfall, snow

squalls, and other rapidly evolving weather events. Nowcasting is important for the prediction of adverse phenomena that might affect operations and even endanger lives. In the terrestrial context, this can be achieved for areas of interest provided that abundant data are furnished by the numerous land-based weather stations. (Cuomo & Chandrasekar, 2021; Price, 2008)

However, in the vast expanse of the open seas, the availability of weather data is usually limited only to satellite data, sparse airborne and floating weather stations and passing ships' reports. In the context of the EFOS project, we investigate the applicability of the atmosphere's electrostatic field as an additional parameter which may help pinpoint adverse weather phenomena in the vicinity of the measurement. In this paper we present the concept and describe the development of an energy autonomous wireless system, capable of measuring the electrostatic field as well as other meteorological variables.

2 Project description

The main objective of the EFOS project is the design and implementation of a state-of-the-art multi-sensor system capable of short-term forecasting weather conditions. The research study primarily focused on:

- Design and implement a low-cost measurement system capable of recording measurements of atmospheric static electric fields, in combination with measurements of atmospheric pressure, temperature, humidity and wind characteristics.
- The system will be installed in ocean liners, so it should be as compact as possible and operate in an environmentally friendly way.
- The system should be powered by a state-of-the-art power management unit, based on renewable energy sources (photovoltaic cells, triboelectric and thermal generators)
- Create a database of behavioral data that are enriched with every new system installed. These data will be integrated and compared with other known arithmetic models and create the baseline for more efficient weather forecasting models.

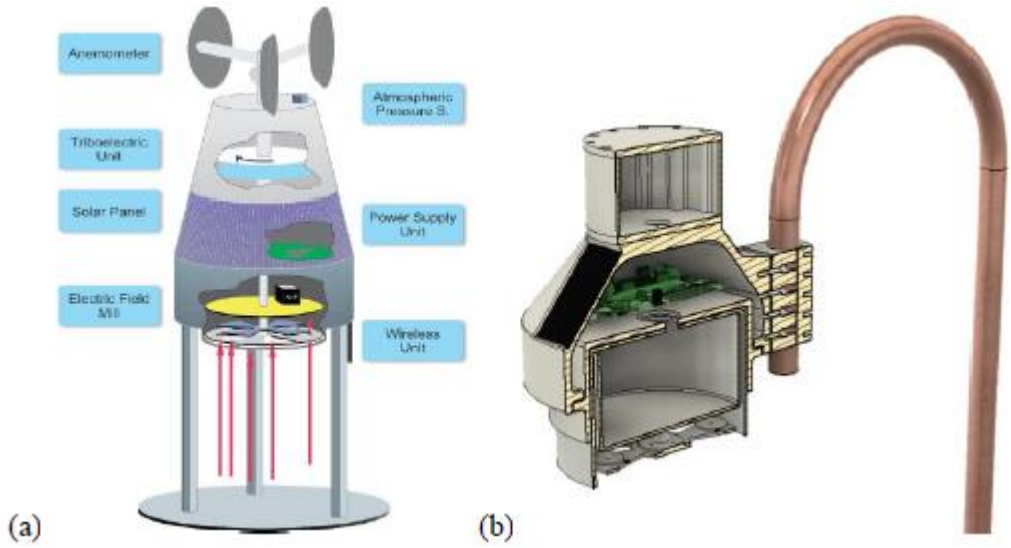


Figure 1: (a) EFOS System concept design **(b)** EFOS System final design

The sensor is portrayed in Figure 1. The project entails the development of an advanced power management unit (PMU) which intelligently optimizes the utilization of renewable sources, based on demand and availability. The PMU simultaneously captures solar energy through photovoltaic cells, converts mechanical movements with triboelectric generators, and recovers waste heat using thermal generators.

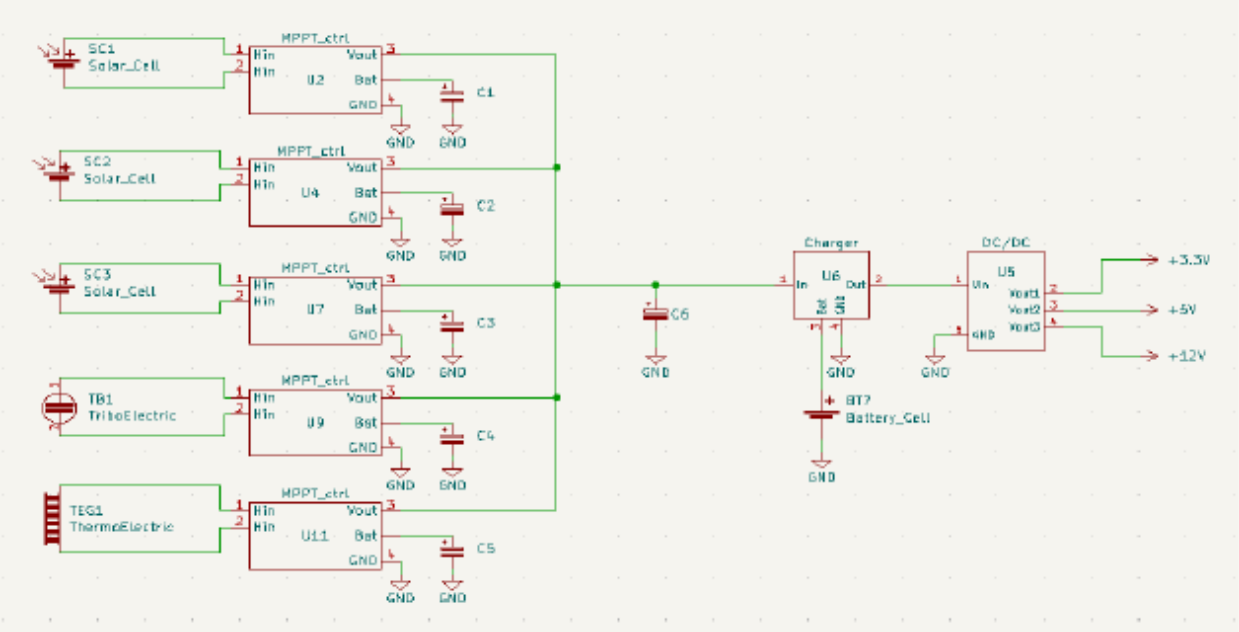


Figure 2: Carrier board circuit schematic

The carrier board serves as a centralized and vital component within the system, encompassing all previously mentioned harvesters and essential elements required for efficient energy management, including the secondary battery to store harvested energy. The main microcontroller unit (ESP32-S3 with RISC-V architecture) and the RF (Digi Xbee) transceiver are also integrated into the carrier board, serving as the computation and communication hub of the system. Furthermore, the carrier board accommodates various sensors responsible for collecting environmental data, creating a comprehensive and interconnected platform that maximizes energy utilization and enables seamless communication and monitoring.

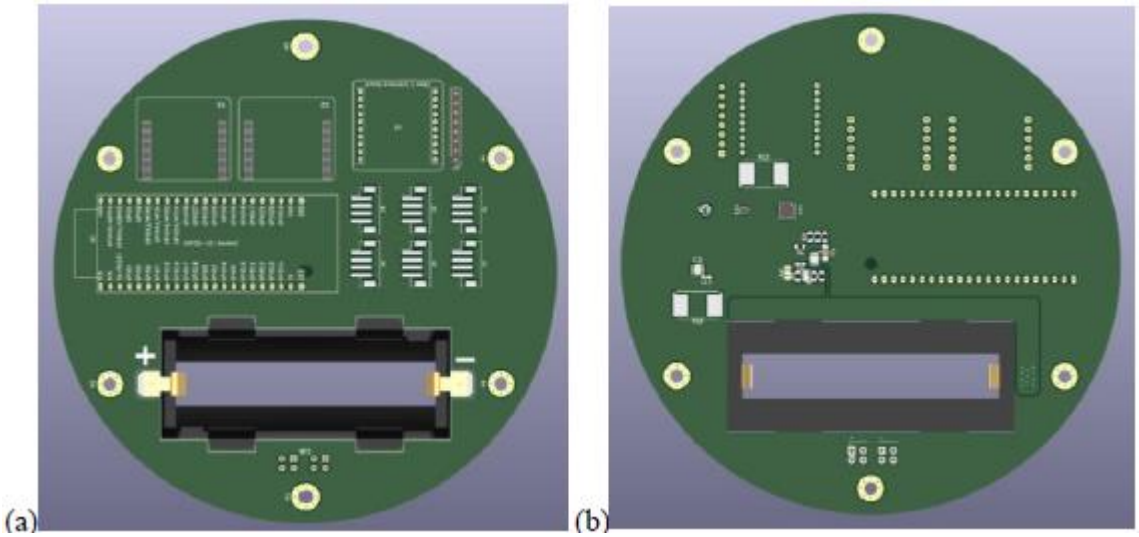


Figure 3: (a) Carrier board top side **(b)** Carrier board bottom side

Experimental Triboelectric generator

The triboelectric phenomenon has been identified as an attractive mechanism for the conversion of mechanical energy to electrical energy. Triboelectric generators (TENGs) have been developed with an energy efficiency of up to 80% and output power of up to 500 W/m2. Triboelectricity is based on contact electrification and electrostatic induction between two surfaces that are in relative motion. In this work we developed a triboelectric generator that is attached to an anemometer and can convert the mechanical energy of the rotational motion of the anemometer, due to the wind, to electrical energy. The electrical energy can be stored and used for supplying power to the electric field sensor.

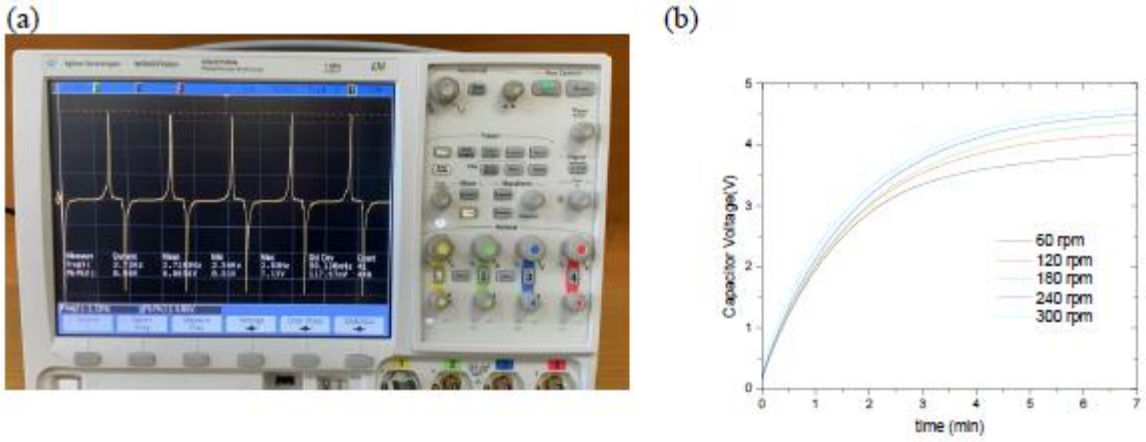


Figure 4: (a) Triboelectric signal as a function of time, due to the rotational motion of the anemometer

and **(b)** Capacitor charging as a function of rotational speed of the anemometer.

3 Project validation

To validate our project, we conducted two essential tests. Firstly, we assessed the accuracy of the EFM (Electromagnetic Field Monitor). Secondly, we evaluated the complete system's performance by connecting the device to an IoT platform.

In the first phase of validation, we focused on evaluating the accuracy and precision of the Electromagnetic Field Monitor (EFM). To perform this test, we deployed the developed EFM in a laboratory environment, with minimal electromagnetic interference.

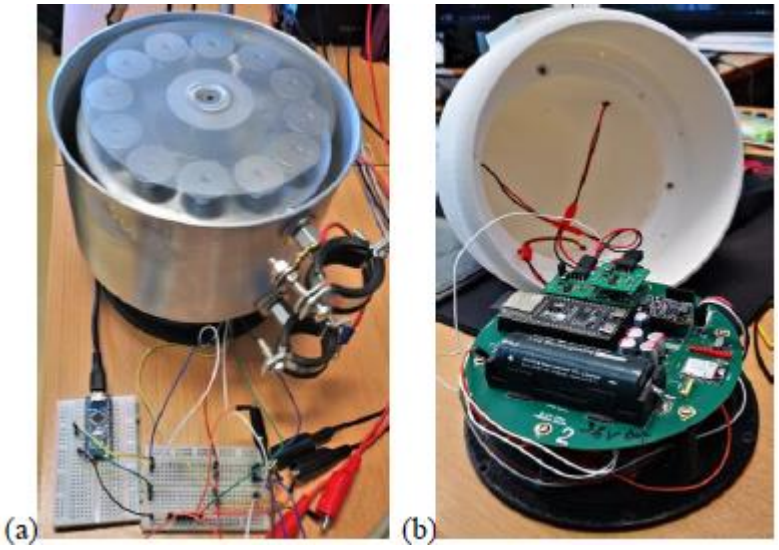


Figure 5: (a) Electric Field Mill **(b)** EFOS System board

During the evaluation, we recorded and analyzed data from the EFM, comparing its readings with well-established reference instruments used to measure electromagnetic fields. The comprehensive comparison allowed us to calibrate the EFM, quantify any discrepancies, and gauge the EFM's reliability in capturing accurate measurements.

In the second phase of our project validation, we conducted an assessment of the complete system's performance. During this test, our primary objective was to establish a seamless integration of the system with an Internet of Things (IoT) platform, enabling remote monitoring and data management. We also aimed to verify the system's ability to operate continuously by harnessing solar energy through Photovoltaic (PV) cells, ensuring sustainable and autonomous power supply. During testing, the system operated perpetually for 48hours without energy input.

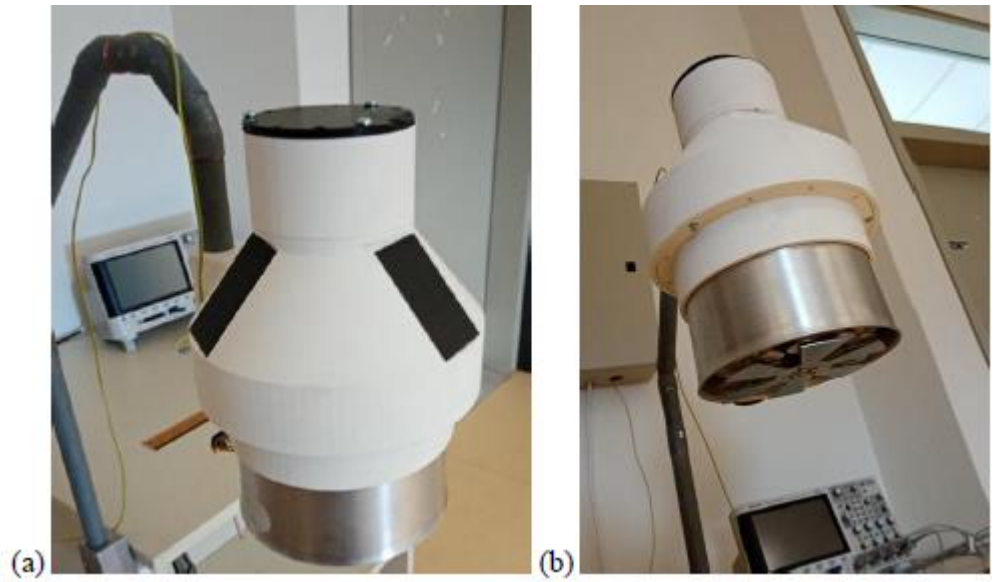


Figure 6: Complete EFOS System

The integration facilitated remote access to the system's parameters, enabling us to monitor and control its functioning from a central location. Moreover, we investigated the system's responsiveness and efficiency in adapting to changing environmental conditions while operating solely on solar power. Furthermore, we examined its ability to effectively manage, and store large volumes of data generated by different sensors, ensuring data integrity and consistency.

Figure 7: EFOS System dashboard

4 Conclusion

By successfully completing the EFOS sensor development, we demonstrated the feasibility of the concept of an autonomous weather sensor system with electrostatic field sensing. The seamless integration with IoT technology and reliable operation powered by renewable energy, ensure that our project can efficiently function as an autonomous and data-driven environmental monitoring solution

with wide-ranging applications in research, environmental protection, and various industry sectors.

Figure 8: EFOS System installed alongside a commercial EFM system

In parallel, preliminary analysis of the captured dataset, indicates a strong potential for forecasting adverse weather phenomena in the short term -using advanced analytics based on machine learning methods. This observation provides a strong path for the commercial exploitation of the system.

5 Acknowledgments

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE - INNOVATE (project code: T2EDK- 00350).

6 References

[1] Cuomo, J., & Chandrasekar, V. (2021). Use of Deep Learning for Weather Radar Nowcasting. *Journal of Atmospheric and Oceanic Technology*, *38*(9), 1641–1656. https://doi.org/10.1175/JTECH-D-21-0012.1

[2] Florita, A. R., & Henze, G. P. (2009). Comparison of Short-Term Weather Forecasting Models for Model Predictive Control. *HVAC&R Research*, *15*(5), 835–853. https://doi.org/10.1080/10789669.2009.10390868

[3] Price, C. (2008). Lightning Sensors for Observing, Tracking and Nowcasting Severe Weather. *Sensors*, *8*(1), Article 1. https://doi.org/10.3390/s8010157

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 247 – 252

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Area Allocation for Coverage Path Planning Using Affinity Propagation Clustering

Nikolaos Baras[1‡‡‡‡‡‡‡] Antonios Chatzisavvas[1] Irene Tabakis[1] and Minas Dasygenis[1]

[1] *Department of Electrical and Computer Engineering, University of Western Macedonia, Kozani 50100, Greece*

nbaras@uowm.gr, achatzisavvas@uowm.gr, i.tabakis@uowm.gr, mdasyg@ieee.org

Abstract

Coverage path planning, also known as CPP, is an essential task in many different industries, including agriculture, robotics, and surveillance. The Comprehensive Path Planning (CPP) project's goal is to locate a route that not only traverses the entire area of interest but also shortens the amount of time it takes to do so. In this paper, we propose a method for allocating space for CPP that makes use of Affinity Propagation Clustering (APC). The area of interest is clustered based on characteristics that are similar thanks to the utilization of APC in the proposed method. After that, a robot is assigned to each cluster, and a route for each robot to follow in order to cover its assigned cluster is planned. The approach that has been suggested intends to optimize the distribution of clusters among robots by reducing the amount of communication overhead as well as the path length. To assess how well the proposed method works, we will compare it with various other approaches by performing simulation experiments. According to the findings of the simulation, the approach that was suggested performs better than other methods in terms of the total path length and the communication overhead. The method that has been suggested is effective in terms of computation, and it can be utilized to solve a wide variety of CPP problems that involve varying area shapes and sizes.

1 Introduction

CPP is a fundamental problem in numerous fields, including robotics, agriculture, environmental monitoring and more [1]. The main objective of CPP is to generate a route that traverses the entire area of interest while minimizing its length or duration. CPP has been extensively studied over the past few decades, and numerous algorithms to solve the problem have been proposed. However, the complexity of CPP increases as the size and shape of the target area and the number of robots involved in the task increase. It is evident that efficient and effective algorithms for CPP is an ongoing and important challenge.

Even though many scholars have given solutions to the single-robot CPP problem, the multi-robot CPP problem is still under research [2]–[5]. The primary benefit of multi-robot CPP over conventional single-robot CPP is its capacity to divide the area of the environment at once, using numerous robots speeds up the covering of huge regions. For the majority of application domains, this is the main

‡‡‡‡‡‡‡ Corresponding author

advantage of multi-robot CPP. Typically, the usage of smaller and less expensive robots that can cooperate to finish the task instead of bigger, more complicated robots is preferred.

Effectively allocating the area of interest to several robots is one of the major issues in CPP. The area allocation problem specifies how to divide the target area into clusters of related objects and assign a robot to each cluster. Since an inefficient allocation can lead to longer path lengths and increased communication overhead between robots, the area allocation problem has a substantial impact on the CPP system's overall efficiency. The performance of CPP systems must therefore be improved by determining an area allocation that is optimal.

In this paper, we propose a method for area allocation in CPP that uses affinity propagation clustering (APC). Area points can be grouped using the APC clustering method according to how similar and dissimilar they are. Each cluster is assigned to a robot for coverage after we utilize APC to group the area of interest into groups of similar features. Our suggested strategy tries to reduce communication overhead and path length in order to maximize the allocation of clusters to robots. We use simulation tests to assess the performance of our suggested methodology in comparison to other approaches.

2 Literature Review

Over the past few decades, a number of strategies for coverage path planning (CPP) have been put forth. Researchers have concentrated on creating algorithms for CPP area allocation in recent years. Voronoi diagrams are one of the often employed techniques for allocating space. The area of interest is divided into regions using Voronoi diagrams, and each zone is assigned to a robot for coverage. However, there are drawbacks to Voronoi-based approaches, including their high computational cost and sensitivity to the beginning point choice. K-means clustering is a different method for allocating space that seeks to organize data points into k clusters according to how similar they are. Although the k-means algorithm has been employed in CPP for area allocation, it is not capable of handling non-convex shapes or irregular distributions of data points.

One other popular multi-robot CPP algorithms is the MSTC* presented by Tang et. al. in [6]. The primary goal of this algorithm is to generate coverage paths for a number of robots while taking into consideration actual physical limitations like obstacles and communication paths between robots. The MST technique [7], which is used to break down the target environment into smaller sub regions, served as the foundation for the algorithm. Based on their skills and the work needs, the robots are subsequently assigned to these sub-areas. This algorithm has the advantage of being able to deal with physical restrictions, which is crucial in real-world circumstances where robots must navigate through intricate and dynamic surroundings. The disadvantage of this algorithm, however, is that it makes use of an arbitrary number of robots to determine the number of robots and sub-areas. Therefore, the number of generated sub-areas is not calculated by the algorithm and cannot be guaranteed to be good. It is worth noting that the problem Tang et.al attempt to solve is similar to the problem formulated in this paper only with respect to area division and allocation. The exact details of the problem the authors at [6] are suggesting a solution for, are vastly different.

3 Problem formulation

Before we present the details of the proposed algorithm, it is important to define the actual problem of area division for multirobot CPP. The objective is to divide a grid-based area of interest into several sub-areas and assign a robot to each sub-area for coverage. The distribution should ensure that each sub-area is completely covered while minimizing communication overhead and path length. The issue

can be formalized as follows:

Let A represent the N-point grid-based area of interest. The area A must be divided into K sub-areas ($S_1, S_2, \dots, S_K$), each of which must be assigned to a robot for coverage. Both the total path length for all robots and the communication overhead between robots should be kept to a minimum. Each robot should only travel to the sub-areas that it is assigned, and each sub-area should be completely covered. By defining an objective function that represents the communication overhead and path length, the issue can be expressed as an optimization problem. If each sub-area is completely covered and each robot only visits the sub-area that is allotted to it, the objective function can be minimized. The CPP problem can be successfully solved by using the APC algorithm to determine the best distribution of robot sub-areas.

4 Implementation

The most important stage of the proposed model is the area allocation and division. We suggest using an affinity propagation algorithm [8], [9] to carry out this task. AP is a clustering algorithm that organizes data points into clusters. This is accomplished by sending messages between data points until agreement is reached on the number of clusters and which points belong to which cluster. The Affinity Propagation clustering algorithm is utilized in the context of this paper's focus on multirobot CPP to resolve the issue of segmenting the area into numerous sub-areas without predetermining the number of robots or the size of the clusters in an arbitrary manner. Fig. 1 presents an overview of the proposed algorithm.

The procedure that the algorithm follows to generate and output the final clusters, given as input a grid based area can be expressed as a 6 step process:

1. Convert given environment to a set of data points, suitable for APC.
2. Construct the similarity matrix: The first step is to establish the similarity matrix for the data points in the study area. The similarity matrix specifies the pairwise similarity between the data points, and each point in the region is taken to be a data point.
3. Calculate the responsibility matrix: The responsibility matrix, which shows whether each data point is suitable

   to serve as an example, is the next step. The similarity matrix and a damping factor are used to compute the responsibility matrix.
4. Calculate the availability matrix: The availability matrix shows whether each data point that will be assigned to an exemplar is available. It is calculated using a damping factor and the responsibility matrix.
5. Update the availability and responsibility matrices: Until convergence, the availability and responsibility matrices are updated iteratively. The updating procedure entails calculating the updated matrix values from the old values and the damping factor.
6. Find the exemplars: The availability and responsibility matrices are used to find the exemplars. If a data point has high availability and responsibility, it is regarded as an exemplar.

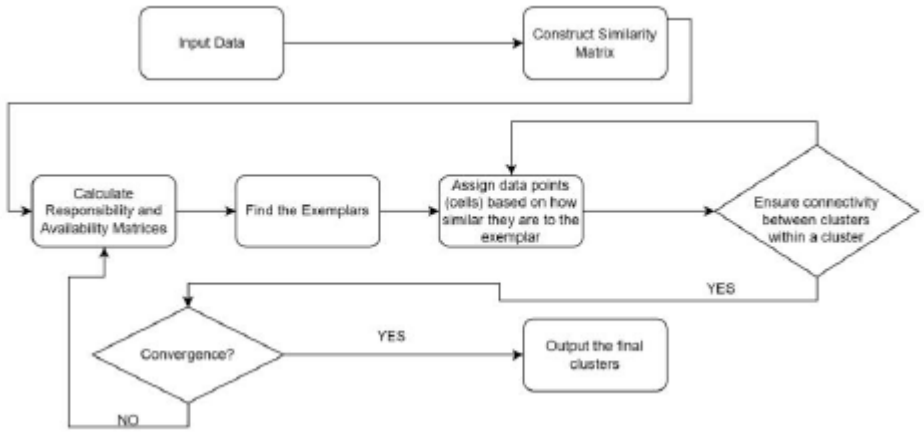


Figure 1: Graphical representation of the proposed algorithm.

Finally, the data points are grouped based on how similar they are to the exemplars. Each cluster has a designated sub-area that can be given to a robot to cover.

5 Experimental Results

To evaluate the performance of the proposed algorithm, we used three different environment setups: (A) 20x20, (B) 50x50 and (C) 100x100. Each environment was pseudo-randomly assigned 20% of its cells as obstacles and 80% of its cells as accessible area. The environments were given as input to each algorithm in the form of a binary matrix. The algorithm that we chose for comparison was the k-means clustering algorithm [10]. Figure 2 depicts an example 10x10 environment, which is converted from a regular grid to a set of data points.

In a typical dataset, a metric that is commonly used to evaluate the clustering performance is the Silhouette Score (SS). SS measures the quality of a clustering algorithm by computing the mean distance between each data point and other points in the same cluster (cohesion) and the mean distance between each data point and other points in the nearest cluster (separation). In our case, we used a customized SS function that considers the 4neighbor connectivity between cells and their normalized distance based on Lee's algorithm.

$ss(i) = (b(i) - \alpha(i)))/max(\alpha(i), b(i)$ (1) where:

- $ss(i)$ is the Silhouette Score for data point i.
- $\alpha(i)$ is the average distance between i and all other points in the same cluster.
- $b(i)$ is the smallest average distance between i and all points in any other cluster.
- $ax(\alpha(i), b(i)$ is the maximum of $\alpha(i)$ and $b(i)$.

The experimental results for the environments (A), (B) and (C) using the proposed algorithm and the traditional k-means algorithm (for cluster sizes 2, 3 and 5) are presented in Figure 3. The exact number of generated clusters using the proposed APC algorithm was based on the exact variables during execution and the environment size.

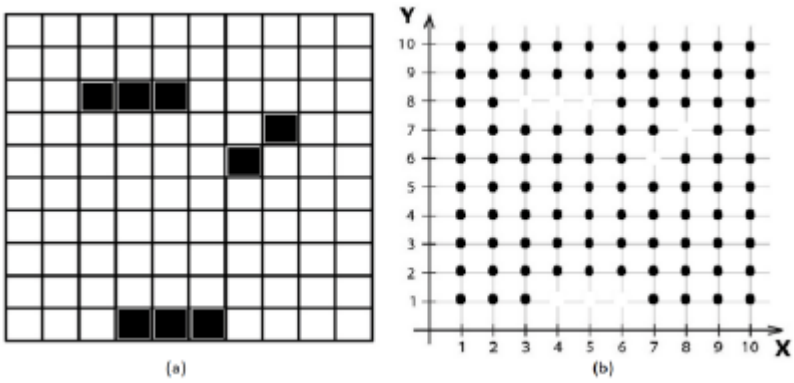


Figure 2: Initial environment **(a)** converted to data points **(b),** suitable for the proposed APC algorithm.

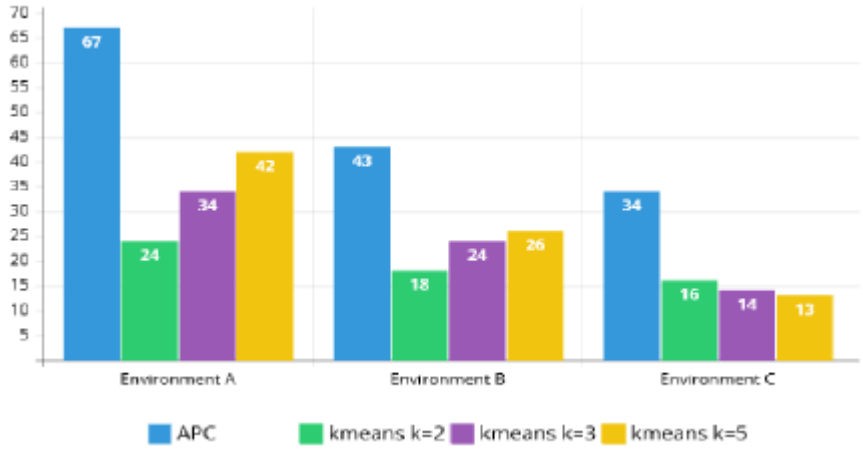


Figure 3: Performance evaluation of the proposed APC algorithm.

It is worth noting that even if the k-means algorithm *can* be faster and less computationally expensive than APC, it has the overhead of executing it multiple times to find the optimal number of clusters k. The proposed APC, on the other hand, does not require the arbitrary choice of clusters (sub-areas) and this is undoubtedly one of its main advantages.

6 Conclusions

In this research, we suggested an affinity propagation clustering-based method for area allocation in coverage path planning. The region of interest was divided into clusters of comparable features using the APC method, and each cluster was assigned to a robot for coverage.

The simulation results demonstrated that, in terms of overall path length and communication overhead, our suggested technique performed better than alternative approaches. Our method proved computationally effective and applicable to a variety of CPP issues with diverse area sizes and shapes. The simulations that we conducted used three different environment types with randomly allocated obstacles. We compared the proposed APC algorithm to the traditional k-means algorithm, using an arbitrary number of clusters. The experiments showed that the clustering offered by the APC algorithm was consistently superior compared to a typical k-means implementation.

In conclusion, the APC method we have suggested for CPP is a promising method that can enhance efficiency in multi-robot systems. Our study's findings showed how effective the suggested method is, making it a potential remedy for coverage path design in a variety of contexts. Future research could examine the viability of our strategy in practical applications and further improve the algorithm to handle more complicated cases.

7 References

[1] E. Galceran and M. Carreras, "A survey on coverage path planning for robotics," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1258–1276, Dec. 2013, doi: 10.1016/j.robot.2013.09.004.

[2] R. Almadhoun, T. Taha, L. Seneviratne, and Y. Zweiri, "A survey on multi-robot coverage path planning for model reconstruction and mapping," *SN Appl. Sci.*, vol. 1, no. 8, p. 847, Jul. 2019, doi: 10.1007/s42452-019-0872-y.

[3] V. G. Nair and K. R. Guruprasad, "GM-VPC: An Algorithm for Multi-robot Coverage of Known Spaces Using Generalized Voronoi Partition," *Robotica*, vol. 38, no. 5, pp. 845–860, May 2020, doi: 10.1017/S0263574719001127.

[4] G. Sun, R. Zhou, B. Di, Z. Dong, and Y. Wang, "A Novel Cooperative Path Planning for Multi-robot Persistent Coverage with Obstacles and Coverage Period Constraints," *Sensors*, vol. 19, no. 9, Art. no. 9, Jan. 2019, doi: 10.3390/s19091994.

[5] A. V. Le, V. Prabakaran, V. Sivanantham, and R. E. Mohan, "Modified A-Star Algorithm for Efficient Coverage Path Planning in Tetris Inspired Self-Reconfigurable Robot with Integrated Laser Sensor," *Sensors*, vol. 18, no. 8, Art. no. 8, Aug. 2018, doi: 10.3390/s18082585.

[6] J. Tang, C. Sun, and X. Zhang, "MSTC∗:Multi-robot Coverage Path Planning under Physical Constrain," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, May 2021, pp. 2518–2524. doi: 10.1109/ICRA48506.2021.9561371.

[7] Y. Gabriely and E. Rimon, "Spanning-tree based coverage of continuous areas by a mobile robot," *Annals of Mathematics and Artificial Intelligence*, vol. 31, no. 1, pp. 77–98, Oct. 2001, doi: 10.1023/A:1016610507833.

[8] "Clustering by Passing Messages Between Data Points | Science." Accessed: Oct. 09, 2023. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1136800

[9] C.-D. Wang, J.-H. Lai, C. Y. Suen, and J.-Y. Zhu, "Multi-Exemplar Affinity Propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 9, pp. 2223–2237, Sep. 2013, doi: 10.1109/TPAMI.2013.28.

[10] H. Guo, J. Ma, and Z. Li, "Active Semi-supervised K-Means Clustering Based on Silhouette Coefficient," in *Advances in Intelligent, Interactive Systems and Applications*, F. Xhafa, S. Patnaik, and M. Tavana, Eds., in Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2019, pp. 202–209. doi: 10.1007/978-3-030-02804-6_27.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 253 – 256

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Crosstalk exploration between PA and LNA in-ductors

Dimitrios Samaras, Georgios Chararas, Athanasios Stefanou, Rafaela Themeli, Vasileios Pavlidis and Alkiviadis Hatzopoulos

Department of Electrical and Computer Engineering Aristotle University of Thessaloniki

samadimi91@gmail.com,chararasg@ece.auth.gr, stefanou@ymail.com, rthemeli@physics.auth.gr, vpavlid@ece.auth.gr, alkis@ece.auth.gr

Abstract

The crosstalk between inductors of the analog front-end in a transceiver is present-ed. Various distances between the inductors are simulated to analyze the effect of magnetic and substrate coupling. A GF22nm SOI technology library is used, with a supply voltage of 0.8V for the LNA and 1.8 V for the PA. The best possible outcome in terms of low crosstalk effect and total noise is when distances above 300 μm are used between the “aggressor” and the “victim” inductors.

1 Introduction

Power amplifiers (PA) and low noise amplifiers (LNA) are crucial components in an RF transceiver. LNA receives a very weak signal which cannot be used as is in the receiver (RX). Therefore, the LNA amplifies the input adding the lowest noise possible [1]. The main characteristics of LNAs are Noise Figure (NF), IIP3 (3rd order Input Intercept Point), Power Gain (PG) and current consumption.

On the other hand, a PA as part of the transmitter (TX) block amplifies the signal transmitted to other RXs. PAs are usually power-hungry circuits as the output power must be sufficiently high to ensure a proper transmission range. Both PAs and LNAs are area consuming blocks due to the inductors these blocks contain [2]. The GaAs technology is highly suitable for a PA exhibiting the desired frequency characteristics and high supply voltages, but leads to large silicon area, thereby hindering the use of this technology in integrated systems. In modern days, every component of a transceiver chip is implemented in CMOS technology due to continuous decrease MOS channel length [3].

PAs consume significant power and operate in high voltage supply domains to transmit a powerful signal in electromagnetic terms. This behavior creates several issues mainly to adjacent blocks that also require inductors. On the contrary, LNAs receive and amplify a weak signal, sensitive to various noise factors. When integrating an transceiver circuit in a single die, the inductor of the PA can generate serious crosstalk and inject unwanted tones in the receiving path. Therefore, LNA’s inductor is called the “victim” and the PA’s inductor is called the “aggressor” respectively.

The effect of crosstalk between the “victim” and the “aggressor” is investigated in this work, using Ansys electromagnetic tools such as RaptorX. In Section 2, a description of the testbench is

presented. In Section 3, the simulation results are presented and Section 4 concludes this work.

2 Testbench setup

An LNA has been designed with a gain of 13.7dB at 28GHz to explore the effects of the crosstalk between PA and LNA inductors. The transmitted signal is generated by a port and fed to an inductor, thereby generating the aggressor part for the LNA's inductors. The input of the LNA is grounded via a resistor and a capacitor in series while the output is ac-coupled to a second port. The schematic of the testbench is depicted in Figure 1.

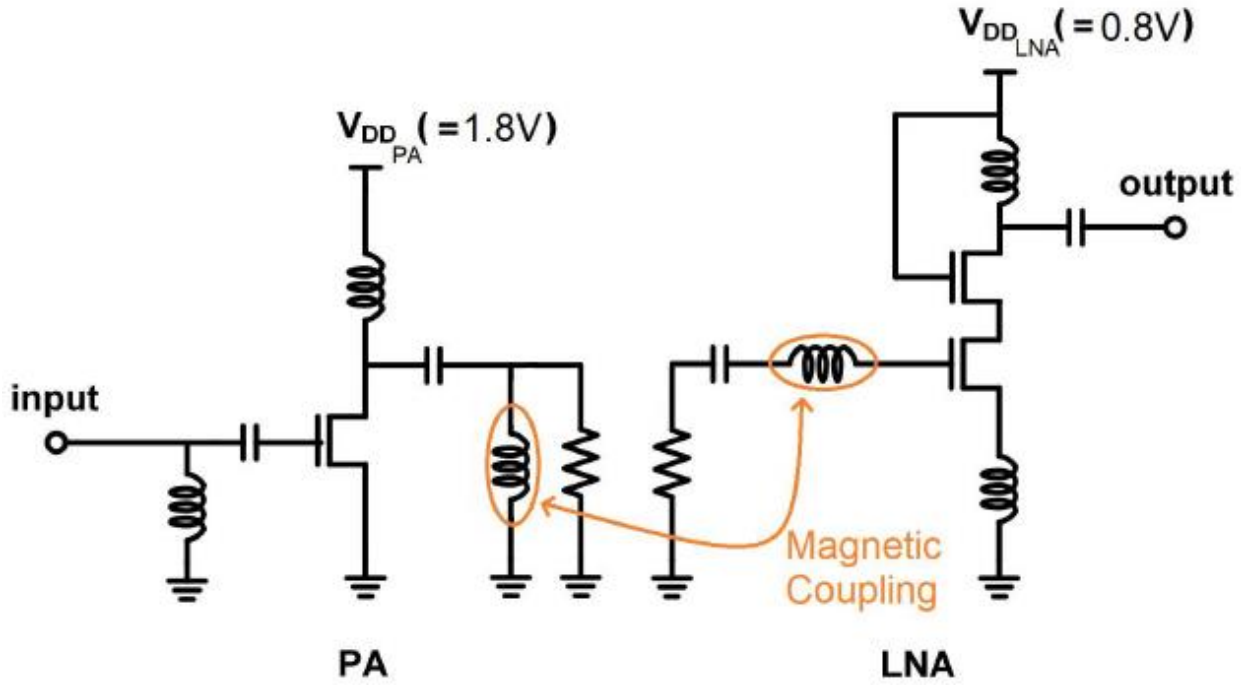


Figure 1. Testbench of the crosstalk study between the PA's and LNA's inductors

All the inductors are placed in a layout view and were extracted using RaptorX tool in order to ob-tain the mutual inductance, substrate capacitance, and sheet resistance. The main contributors to the crosstalk between the PA and LNA are the two inductors highlighted in Figure 1. This is because the input inductor of the LNA is the most susceptible to noise sensitivity and the PA inductor creates a strong electromagnetic field. Using the setup of Figure 1, PA is simulated by a port, and the related metrics are observed at the output of the LNA. When strong magnetic interference is present, the signal of the PA is observed at the output of the LNA, since the LNA has no input. The extracted views of all the inductors contain the s-parameters information. Finally, it must be noted here that the information regarding the substrate is not available.

3 Simulation Results

In this Section we observe the effect of electromagnetic interference and substrate coupling in terms of the distance between the inductors of the PA and the LNA circuit respectively. As illustrated in Figure 2 the distance of the "aggressor" inductor is gradually increased. The simulated distances are 100μum, 200μm, 300μm and 1000μm respectively. All views are extracted and Spectre simulator has

been used. In Figure 3, the simulation results for the distances are presented. The S_{21}parameters are measured where the effect of the crosstalk from the "aggressor" to the "victim" inductor is observed.

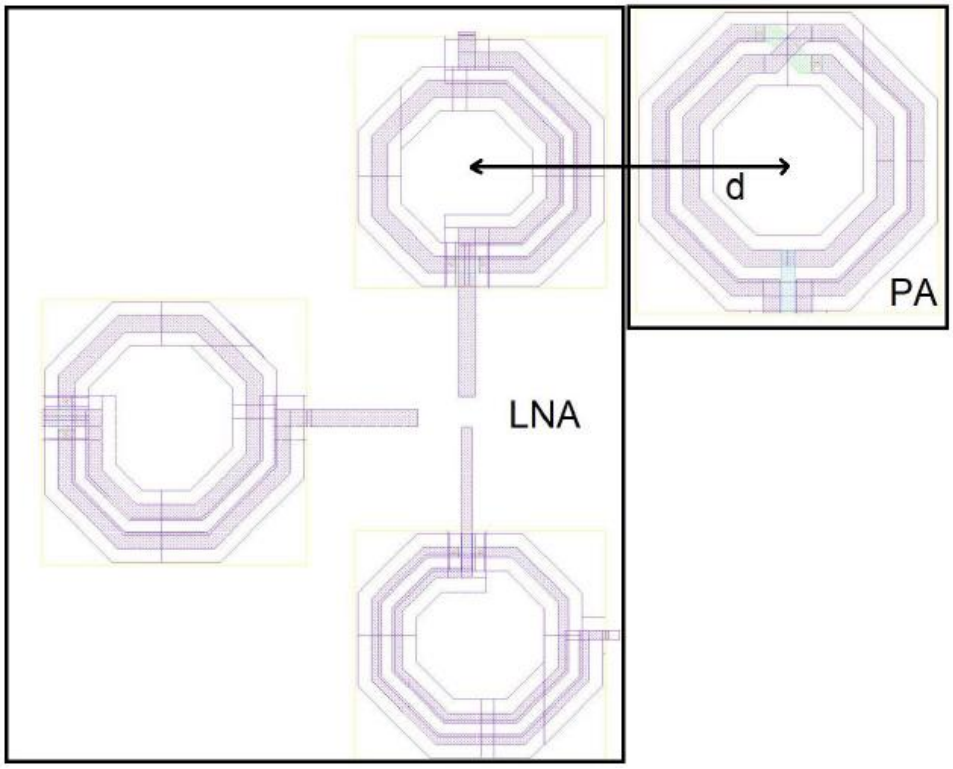


Figure 2: Layout placement of the inductors with increasing distance at every simulation test

The black trace is the dominant curve with -22dB at 28GHz, which is expected, considering that the "aggressor" is very close (100μm) to the "victim". As the distance of the "aggressor" increases we the effect decreases. More specifically, when the distance is 200μm, the S_{21} is -32dB lower by 10dB as compared to the distance of 100μm. For the distance is 300μm the S_{21} is -38dB and finally when the distance is much higher (1000μm) the effect of the crosstalk is minimal at the LNA's output with -50dB. The results in Figure 3, show that the simulated S_{21} tends to lower when the distance increases, proving that the isolation is a matter of overall distance between the inductors.

The desired distance of the "aggressor" and the "victim" inductors depends on each application. If the overall quality of the signal reception and transmission is state-of-the-art, then the distance should be several hundreds of μm, but it comes with the cost of die size which increases dramatically. On the other hand, if the application does not have specifications for crosstalk interference, then the distance could be smaller, but the quality of the signals will be impaired by the proximity of the "aggressor" inductor. However, there are some techniques, which can reduce the overall crosstalk between the inductors of interest, such as using twisted pair inductors [4]. In the work presented, the best isolation in terms of S_{21}parameters and overall crosstalk, comes with a distance greater than 300μm.

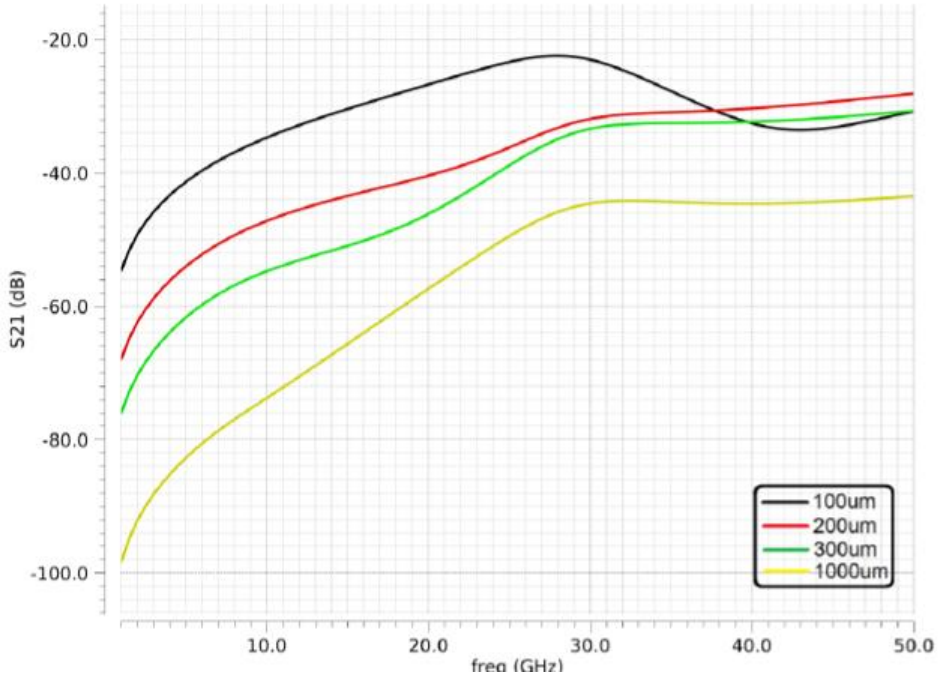


Figure 3: S_{21} parameters for different distances

4 Conclusion

Crosstalk between PA and LNA is a critical and unwanted phenomenon in transceiver blocks. Crosstalk can affect the performance of both the receiver and transmitter blocks. An investigation of the crosstalk between PA's ("aggressor) and LNA's ("victim") inductors is presented. A 28GHz LNA is used including the extracted views for the transistors and the electromagnetic models for the inductors using the RaptorX tool. S_{21}parameters are simulated and described to determine the total isolation between the inductors. The best possible outcome in terms of low crosstalk effect and total noise is when distances above 300μm are used between the "aggressor" and the "victim".

5 References

1 F. Vecchi et al., "A Simple and Complete Circuit Model for the Coupling Between Symmet-rical Spiral Inductors in Silicon RF-ICs", *Proceedings of the IEEE Radio Frequency Inte-grated Circuits Symposium*, 2008.

2 R. Minami et al., "Measurement of Integrated PA-to-LNA Isolation on Si CMOS Chip", *Proceedings of the Asia-Pacific Microwave Conference*, 2010.

3 J. N. Burghartz et al., "RF Circuit Design Aspects of Spiral Inductors on Silicon", *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 12, December 1998.

4 K. Ben Atar, P. Martin and R. Horn, "A Multi-Turn Twisted Inductor for On-Chip Crosstalk Reduction", *Proceedings of the ISCEE International Conference on the Science of Electrical Engineering*, 2016.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 257 – 260

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Auditory Scene Profile Adaptation for ANC Headphones

Dennis Tsoukalas[1] and Fotios Kontomichos[1]

[1] *Renesas Design Greece*

dennis.tsoukalas.fv@renesas.com , *fotios.kontomichos.xw@renesas.com*

Abstract

Adaptive Active Noise Cancellation (ANC) is a key feature for cutting edge and next generation headphones. Renesas DA74xx Codec family is a configurable ANC device that can be optimized for different ambient noise characteristics. This work, extends the integrated DA74xx ANC functionality to an adaptive solution, optimizing the ANC performance in different auditory scenes, with a novel minimal resources approach that enhances the user experience.

1 Introduction

ANC headphones, typically utilizing two microphones (Feedforward and Feedback), are commonly configured with a static ANC filter tuning. The filter design is typically a process involving a single stationary noise source extending in the full audible frequency range.

The subjective performance of a headphone device featuring static ANC filter design may be sub-optimal in real-world conditions for several reasons, among them:

1. Poor attenuation in the main frequency band of the noise
2. High attenuation in frequency areas outside the dominant noise band
3. Overshoots in frequency bands that sound annoying
4. High self-noise of the device in quiet environments

It is evident that a one-size-fits-all is not possible for static ANC solutions because the environmental noise characteristics, that the listener is exposed to, are varying continuously.

This challenge can be most commonly addressed either via the utilization of multiple specialized ANC profiles for manual selection by the user or via a real-time ANC filter adaptation algorithm. The first solution is sub-optimal and bound to be abandoned by the user. The second solution requirements are high in processing/memory resources and power.

Here, we propose a method that uses minimal resources to achieve such an adaptation to the external acoustic conditions.

2 Adaptive ANC

This work introduces an Auditory Scene based, perceptually-weighted solution that triggers an automated profile switch to achieve matching of auditory scene characteristics with an ANC profile that optimizes the noise cancellation performance. A minimal resources algorithm has been developed for DA74xx chip variants (Renesas, 2019) together with a library of auditory scene based ANC profiles. A Proof of Concept application has been implemented and verified.

As a first step, we analyzed a broad range of noise recordings from standardized and non-standardized databases. Twenty-eight (28) auditory scenes representing airplane, car, bus, train, café, kindergarten, nature (forest, rain) (ETSI, 2017) have been grouped in three main categories following their dominant acoustic energy frequency distribution. Following this analysis, the target ANC curves have been determined for each noise group and have been utilized as a template for the auditory scene based ANC profiles tuning.

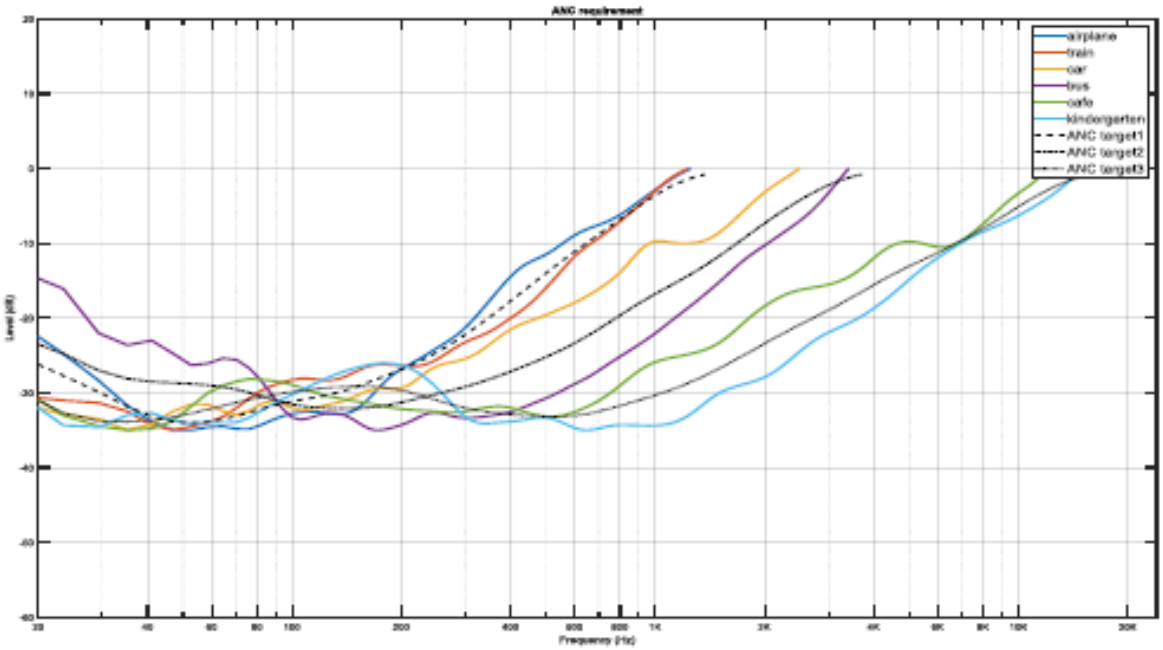


Figure 1. ANC target curves for different auditory scenes

The auditory scene analysis inside the chip is realized with a proprietary algorithm that continuously monitors the auditory scene spectral characteristics. The performance of the ANC library profiles is simulated based on a perceptual loudness measure. The profile that is a best match to the auditory scene is selected and applied to the Codec, subject to time constraints, stationarity of the auditory scene, and match between left and right channels. Figure 2 shows the attenuation measurements for three (3) different library profiles.

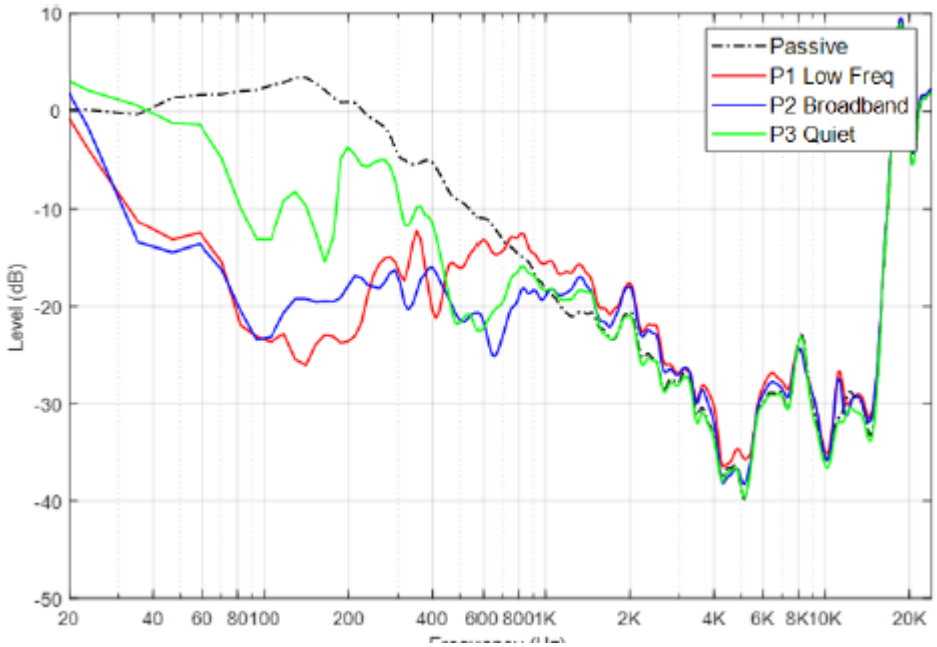


Figure 2: Evaluation of noise reduction for tuned ANC profiles in PoC application

The algorithm has been implemented as a Proof of Concept using the Renesas DA7401 DevKit (SmartBeat™ DA7401 mono codec with wideband digital ANC, 2019), utilizing the sophisticated supportive software and a specially developed user-friendly Android App for controlling the device. The User Interface of the mobile application is shown in Figure 3.

Figure 3: Android App GUI; Manual ANC profile selection **(a)**, Automated Auditory Scene Identification algorithm enable and profile switch time setup **(b)**

The hardware setup is shown in Figure 4. It consists of two Renesas DA7401 Codecs on an Evaluation Kit connected to an Android phone. On the EVK a reworked headphone mockup is attached.

Figure 4: Setup implementation of the Adaptive ANC method

A subjective evaluation of the reference design has been conducted to verify the algorithm profile selection against the user preference. Six (6) expert listeners have been involved to rate each profile for the auditory scenes used for the ANC target curve analysis. The final ANC profile tunings in the integrated reference design application and the effectiveness of our method has been also confirmed by another round of listening tests. It has been found that the automatically chosen profile would be the preferred profile for the majority of the listeners.

3 Conclusions

An adaptive ANC method has been developed and tested for Renesas DA74xx chip family. The solution extends the integrated DA74xx ANC functionality to adaptively match the spectral characteristics of the auditory scene. The method has been implemented with limited resources and negligible increase in power consumption. Subjective evaluation has proven the effectiveness of the adaptive ANC solution.

As future improvements we consider a more user-personalized approach with respect to the subjective match of the ANC profiles to the auditory scene. In addition, we are working to combine together more adaptive features into a single solution that covers different ANC perspectives for an integrated user experience.

4 Bibliography

[1] ETSI. (2017). ETSI EG 202 396-1 (V1.1.2): "Speech Processing, Transmission and Quality Aspects (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise database". ETSI.

[2] Renesas. (2019, 5 23). SmartBeat™ DA7401 mono codec with wideband digital ANC. Retrieved from www.renesas.com: https://www.renesas.com/eu/en/document/prb/da7401-product-brief?r=1563356

[3] Renesas. (2019). *SmartBeat™ DA7402 stereo codec with wideband digital ANC*. Retrieved from www.renesas.com: https://www.renesas.com/eu/en/document/prb/da7402-product-brief?r=1563306

Author Index