

UNIMODAL AND MULTIMODAL SENSOR FUSION FOR WEARABLE ACTIVITY RECOGNITION

Dissertation

Thesis approved by the Department of Computer Science University of Kaiserslautern-Landau
for the award of the Doctoral Degree Doctor of Engineering (Dr.-Ing.)

to

Hymalai Yenireth Bello Valera

Date of Defense	18.12.2024
Dean	Prof. Dr. Christoph Garth
Reviewer	Prof. Dr. Paul Lukowicz
Reviewer	Prof. Dr. Kristof Van Laerhoven
Reviewer	Prof. Dr. Karsten Berns

D-386

Abstract

The “Expressiveness of Human Body Movements” inspires this dissertation. People naturally synchronize hand, body movements, and facial expressions to create a cohesive nonverbal message. To understand this communication, measuring and quantifying it in natural settings is essential. Consequently, the focus of this work is on designing versatile wearable solutions, taking the situational context of body actions into account. The dissertation introduces a set of measurement tools to the wearable community, aiding in expanding the understanding of body movement expressiveness. It designs experimental scenarios with typical gestures associated with body language. It is important to note that the evaluations in this thesis primarily assess hardware capabilities and do not aim to evoke genuine emotions in participants. The work also proposes a variety of multipositional and multimodal wearable prototypes. The idea is that different sensor positions and multiple sensing modalities help to form a unified perception and understanding of complex situations. Another relevant aspect is to recognize human behavior/activities pervasively. Wearable devices are the most promising ubiquitous human activity recognition (HAR) option. Creating wearable-based HAR solutions that are both small and widely accepted by users presents a significant challenge. In this context, a multidisciplinary approach is required. This includes expertise in sensor technologies, signal processing, data fusion algorithms, and domain-specific knowledge. One way to gain user acceptance is to deploy the HW/SW systems in the most common wearable accessories on the market. Hence, the designs presented here are based on wristbands, goggles, headwear (helmet and sports cap), and clothing (jacket and gloves). This work focuses on HW/SW co-design systems for HAR in the context of **Hand Position and Gesture Estimation, Head and Facial Muscle Movements Recognition, and Body Postures and Gesture Classification**. Considering these three scenarios, the thesis explores customized smart-wearable design with application-specific goals. The decision criterion in this work is based on two factors: how relevant the scenario is to understanding human behavior and how innovative the sensing technology is within the wearable community. Overall, the designs have been tested in various experimental settings with evaluation based on mimicked gestures. The experiments were designed to test the feasibility of the proposed hardware for solving specific scenarios. The main goal is to provide tools that can be used in the future to understand the “Expressiveness of Human Body Movements” in a ubiquitous way. Nonetheless, the experiments should be extended to include the emotional element of expressions. This is beyond the scope of this dissertation, which focuses on mimicked experiments.

Acknowledgements

First, I want to thank God, who has always given me peace, even when my mind is at war.

Thank Prof. Dr. Paul Lukowicz for the opportunity and freedom to develop my wildest ideas without judgment.

Thank my colleagues for making every day a constant search for exciting ideas and adventures to pursue.

Special thanks to Jane Bensch, who always ensures good coffee and plenty of water and that we do not forget which days are holidays.

To Bo Zhou, who taught me that the figures in a research paper are the key to encouraging reviewers to read the paper.

To Sungho Suh, who always gives me opportunities to learn and explore new horizons, always supporting me even when I am very slow.

To Red, my dog, who has been forced to be with me as I write this dissertation.

To the Sánchez Marín family, who always try to support me, even by teaching me to be healthy and fit, what they call “tough love”.

Thank you to all those who are not mentioned by name but know that they have been part of my laughter and happiness, thank you all and I hope to return the support to all of you someday.

Contents

1. Introduction	1
1.1. Motivation	1
1.2. Contribution	6
1.3. Related Work	8
1.4. Outline	15
2. Hand Position and Gestures Recognition with Smart-Wearable Devices	17
2.1. Hand Vertical Position Estimation with Atmospheric Pressure and RFID synchronization	18
2.2. Capacitive and Inertial Fusion-Based Glove for Real-Time on Edge Hand Gesture Recognition	28
3. Facial Muscle and Head Movements Estimation with Mechanomyography and Inertial Fusion	39
3.1. Facial Muscle Activity Recognition with Reconfigurable Differential Stethoscope Microphones	40
3.2. InMyFace: Inertial and Mechanomyography-Based Sensor Fusion for Wearable Facial Activity Recognition	58
3.3. Meciface: Mechanomyography and Inertial Fusion-based Glasses for Edge Real-Time Recognition of Facial and Eating Activities	77
4. Body Posture and Gestures Recognition with Multipositional Capacitive Fusion	89
4.1. Problem Statement	90
4.2. Contributions	91
4.3. Electronic and garment prototype	91
4.4. Experiment Design	93
4.5. Signal and Data Processing	95
4.6. Results	98
4.7. Discussion	101
4.8. Conclusion	103
5. Conclusion	105
5.1. Limitations and Future Work	106

Bibliography	109
A. Curriculum Vitae	131

List of Figures

1.1.	Simplify Diagram of the Unspoken Expressiveness of Human Body Movements with Specific Example Scenarios Studied in Thesis.	1
1.2.	Wearable Devices and Applications Proposed in this Dissertation.	7
2.1.	Left The Hardware Prototype Shows the Antenna, Microcontroller (STM32L475), the RFID-tag, RFID-Reader(PN532), and the Barometer (BMP388). (Right) Depicts The Sample Rate Distribution for the Pressure Data in The Real-Time Operating System MBED.	20
2.2.	Three Days Static Pressure Variation Capture by the Customized Hardware for 20 to 45 minutes.	22
2.3.	Top Depicts the Pressure of the Reference Device in Red and the Pressure of the Wearable Device in Blue. Bottom Shows the Coherence of the Signals.	23
2.4.	Altitude Values Level Distribution (Vertical axis in Meters) from the Warehouse Scenario	25
2.5.	Left Confusion Matrix of the Best Case from the Warehouse Scenario Using RFID-Monitor Error Method. Right Best Case Confusion Matrix Without using RFID Synchronization Method	26
2.6.	Confusion Matrices of Three Participants from The Around the Body Position Scenario for the Naive Bayes Classifier Trained on Data From Participant Three.	27
2.7.	Hand Gestures Dictionary for Drone Control [98]	29
2.8.	(A) The Hardware Prototype Shows the Capacitive Channels and IMU Positions on the Sports Glove. (B) Depicts The Hardware Block Diagram with the Sensors Connections to the Main Board (Portenta H7) and PC.	30
2.9.	Real-Time and on-the-Edge Implementation for Hand Gesture Recognition	31
2.10.	Lightweight Neural Network Structure for Real-Time and on-the-Edge Inference of Hand Gestures using CaptAinGlove. Top: Inertial Model Structure. Bottom: Capacitive Model Structure.	32

2.11. Results of the offline Capacitive Model; Null(0), Up(1), Down(2), Back(3), Forward(4), Land(5), Stop(6), Left(7), Right(8) and F1-score=80%(A). Real-Time on the Edge Results of Capacitive Model; F1-score=67%(B).	33
2.12. Example of Smoothing Temporal Windows for Continuous Recognition [141]	34
2.13. CaptAinGlove V2 Hardware Description with Volunteer Wearing the Prototype (Left and Right Hand)	36
3.1. Pairs of Reconfigurable First Order Differential Stethoscope-Microphone Arrays To Detect Facial Expressions	41
3.2. Facial Expressions/Gestures Mimicking Set; Happiness, Upset, Sadness, Surprise and Disgust [145], Angry [84] and Gestures Blinking[52], Tongue Out[1], Kissing[34] and Taking a Pill.	42
3.3. Inside View of the Helmet with Four Fixed Stethoscope Microphones in Elastic Band (Numbers 1-4) and Two Loose (Numbers 5-6) Top Left . Stethoscope Microphones' Head with Leather Cover and Size Comparison Top Right . 3D Cone Connector Between Electret Microphone and Nurse Stethoscope Head Bottom Left . Stethoscope Microphones Distribution on Volunteer's Face Bottom Right	46
3.4. Apparatus Block Diagram from the Stethoscope Microphones and the Hardware Prototype to the Data Reception by a Python-Based GUI.	47
3.5. Left First Order Differential Microphone Array Block Diagram. With G as a gain factor, d space between microphones, and T time-delay between microphones. Right Polar Plot of first-order Differential Microphone Array (Frequency vs Angle of Arrival). With an Inter-microphone space of 16mm, distance from the source of 50 cm, and frequencies 0.5,1,3,5,10 and 20KHz and sound rejection at ± 90 degrees(red boxes). Diagram by STMicroelectronic[194]	47
3.6. Single Microphones Discrete Frequency Response. Phone as Sound Source and Stethoscope's Head at 3 Centimeters Separation with the Frequency Range of The Square Wave with 20 Seconds Duration at each Frequency.	49
3.7. Frequency Response of Individual Microphones. For the Low-Frequency Spectrum (21-751 Hz), Middle-Frequency Spectrum (1001-1751 Hz) and High-Frequency Spectrum (2001-2501).	50
3.8. DMA Discrete Frequency Response Setting. Phone as Sound Source and Stethoscope's Head at a 3 Centimeters Separation with the Frequency Range of The Square Wave with 20 Seconds Duration at each Frequency. Microphone "X" and Microphone "Y" with a Variable Inter-Microphone Distance "d" between 5 to 12 centimeters. The Sound Source Positions are Back, Right, Left, Top, and Front.	51

3.9.	Differential Microphone Array Discrete Frequency Response for Microphone Pair Distances Between Five, Seven, Nine, and Twelve Centimeters and Sound Source Positions, Back, Right, Left, Top, and Front.	52
3.10.	Flow Diagram for the Facial Muscle Movement Recognition with Differential Microphone Stethoscope	54
3.11.	Left Average Results of Eight Volunteer User-Dependent Evaluation with Ensemble Model in Matlab. Right Cross-User and Leave-One-Session-Out Evaluation Results of Eight Volunteers with SVM in Python.	56
3.12.	Facial Muscle Activities Dictionary with Sensor Signal Examples; 7 Facial Expressions from Warsaw Photoset [145] and 2 Gestures from [21]. Two Channel Raw Audio Data, Thirteen Mel Frequency Cepstral Coefficients (Two Channel Audio-Monophonic), Force Sensitive Resistor, Piezoelectric Film, Orientation and Acceleration.	60
3.13.	Hardware Prototype and Data Collection Diagram. A Sports Cap with Sensors Distribution. B Sensor Placement on the Frontalis and Temporalis Muscles of the Participant. C Data Acquisition Diagrams with the Custom Printed Circuit Board.	62
3.14.	Comparison between Synchronized Sensors Signals for the Dictionary of Facial Movements in Fig. 3.12(without neutral) Versus Activities, such as; Clapping, Walking, Checking Emails and Talking. Joy is yellow, Surprise is pink, Anger is red, Disgust is green, Sadness is blue, Winking is magenta, Fear is purple, Taking a Pill is grey, and the in-between (white spaces) are the noise factors activities.	64
3.15.	Left Ensemble Multimodal Sensor Fusion Model Overview; Fusion in the Prediction Phase. Right Hybrid Multimodal Sensor Fusion Model Overview; Fusion Within Hidden Layers, and Before Prediction.	65
3.16.	Sensor Dependent Neural Network Models. Top Neural Network for FSR-PEF (PMMG) and IMU(Orientation and Acceleration) Information. Bottom Neural Network for Mel Frequency Cepstral Coefficients of Audio (AMMG).	66
3.17.	Modified Inception Block with Dimension Reduction Locally Connected per Sensing Modality	68
3.18.	Multimodal Hybrid Fusion Model	69
3.19.	Results Sensor Dependent and Ensemble Neural Networks A F1-Score of Sensor Dependent Neural Network per Volunteer Comparison. B Recall Results of Ensemble Model per Participant Avg F1=85.00%. C Recall Results One Ensemble Model for All Participants F1=79.00%	73
3.20.	Left Recall Results of Hybrid Fusion Model Leave Out Session for the Eight Best Imitators F1 = 82%. Right Data Analysis Techniques Applied in the Hybrid Fusion Modeling Pipeline.	74

3.21. Facial Muscle Activities Dictionary; 6 Facial Expressions from Warsaw Set of Emotional Facial Expression Photoset [145] and 2 Gestures from [15]. Taking a Pill Facial Muscle Movement is Included to Differentiate Eating/Drinking Episode with the Sporadic Gesture of Touching Face/Mouth.	78
3.22. Meciface Prototype A . Hardware Connections Blocks: Motion and Environmental Station on The Glasses' Nose Bridge with BNO085 (IMU), SPH8878LR5H (Microphone) and BME688 (Barometer). On The Temples are The Force Sensitive Resistor (FSR), Piezoelectric Film (PEF), and QtPy ESP32 (MCU) B .	80
3.23. Real-Time and on-the-Edge Flow Diagram Implementation for the Eating/Drinking Scenario with the Two Stages Hierarchical Modeling; First Stage is the Mechanomyography-based Model (MMG-Model) to Detect Null/Activity. The Second Stage is the Inertial-Model to Classify Eating and Drinking Episodes by Window Size of One Second and Window Step of Half a Second A . Real-time and on-the-Edge Flow Diagram Implementation for the Facial Expressions Scenario with Motion Threshold Detection and Two Stages Hierarchical Modeling; The First Stage is the MMG-Model to detect Null/Activity. The Second Stage is the Inertial-based Model to Classify the Facial Movements Dictionary in Fig. 3.21 B	81
3.24. Results of the offline MMG-Model with Five Volunteers (Leave-one-session-out cross-validation) in Lunch/Dinner Scenario; F1-score=83 %(A). Results of the offline Inertial-Model with Five Volunteers (Leave-one-session-out cross-validation) Lunch/Dinner Scenario; F1-score=88 %(B). Real-Time on-the-Edge Recognition Results for Five Unseen Volunteers (User-independent) in Snacking Scenario; F1-score = 94 %(C).	83
3.25. Results of the offline Inertial-Model for Ten Sessions on Different Days with Leave-One-Session Out Cross Validation for the Recognition of the Dictionary in Fig. 3.21; Joy/Surprise(1), Anger/Disgust/Anger(2), Winking(3), Fear(4) and taking a pill(5) and F1-score=95%(A). Real-Time and on-the-Edge Results of the Inertial-Model for Three Sessions on Different Days for the Recognition of the Facial Activities in the Dictionary in Fig. 3.21; F1-score=86%(B).	84
4.1. Electronic Garment Design "the MoCaBlazer", A Is The Back Part of The Blazer, B Are The Textile Cables Sewn Inside The Garment, C Is The Front Part of The Blazer, D Is The Circuit Simplification Design of The Clapp Oscillator with The Antennas, E Are The Two Options for Collecting Data Coming From The Blazer; with a UART-Wired Option, and a Bluetooth-Based Android Application (Flutter framework) as The Wireless Option.	92

4.2.	Twenty General Upper-Body Gestures/Postures Dictionary with Example Signals. $\mathbf{x} = (0,400)$ Time Steps, \mathbf{y} :Norm.	94
4.3.	Eight Dance Movements Dictionary with Example Signals. $\mathbf{x} = (0,400)$ Time Steps, \mathbf{y} :Norm(0,1).	94
4.4.	Data Partition Scheme to Train and Test the Deep Learning Models. A Shows the Leave-Recording Out (LRO) and Leave-Person Out (LPO) Paradigms Used for the Data of the Twenty General Postures. B Shows the Leave-Recording Out (LRO) Scheme Employed for the Data of the Eight Dance Movements.	96
4.5.	Structure of the 1DConv Neural Network Model Used for The Data of The Eight Dance Movements. Input Shape(time-Steps,Channels,1) = (400,4,1) and Output Shape = 8 Classes.	97
4.6.	Real-Time Recognition System. A The "MoCaBlazer" with the RFID Reader-Tag Pair Positions. B Volunteer Wearing the "MoCaBlazer" with the RFID Synchronization System. C Flow Diagram of the Real-Time Recognition Python Script.	99
4.7.	Confusion Matrices for the Data of the Twenty General Gesture Dictionary. A Results for the Leave-Recording Out (LRO) Scheme. B Results for the Leave-Person Out (LPO) Scheme.	100
4.8.	Individual Models Confusion Matrices for the Data of the Eight Dance Movements Dictionary. A Results for the Leave-Recording Out (LRO) Scheme for Participant One. B Results for the Leave-Recording Out (LRO) Scheme for Participant Two. C Results for the Leave-Recording Out (LRO) Scheme for Participant Three.	100
4.9.	Group Model Evaluation Results. A Confusion Matrix for the Offline Test Results with Leave-Recording Out (LRO) for Three Participants. B Confusion Matrix for the Online Results Using the RFID Synchronization Method for One Volunteer.	101

Chapter 1

Introduction

Ph.D Forum H. Bello, “Unimodal and Multimodal Sensor Fusion for Wearable Activity Recognition” 2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (Per-Com Workshops), Biarritz, France, 2024, pp. 364-365, doi: 10.1109/Per-ComWorkshops59983.2024.10502797.

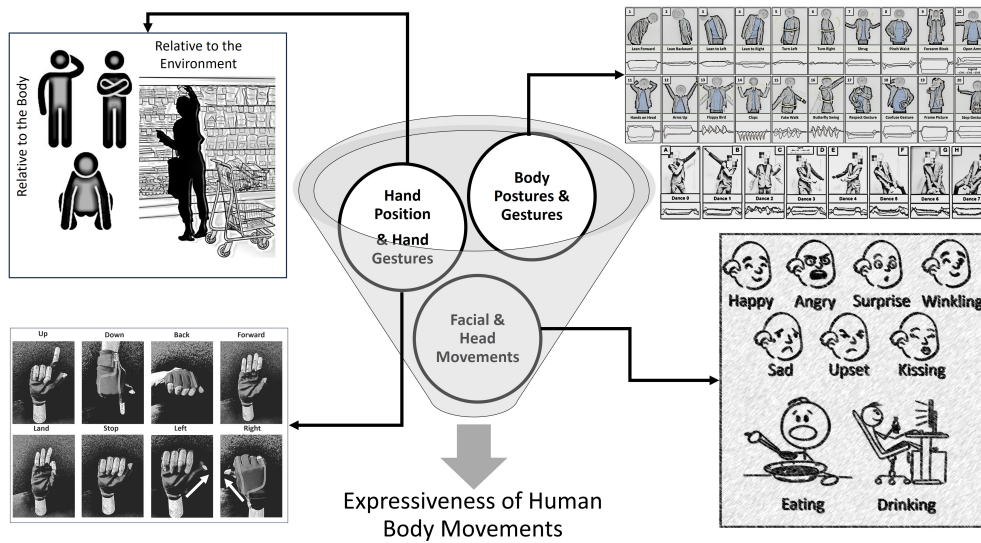


Figure 1.1.: *Simplify Diagram of the Unspoken Expressiveness of Human Body Movements with Specific Example Scenarios Studied in Thesis.*

1.1. Motivation

The “Expressiveness of Human Body Movements” inspires this dissertation. People naturally synchronize hand, body movements, and facial expressions to create a cohesive nonverbal message. To understand this communication, measuring and quantifying it in natural settings is essential. Consequently,

the focus of this work is on designing versatile wearable solutions, taking the situational context of body actions into account. The dissertation introduces a set of measurement tools to the wearable community, aiding in expanding the understanding of body movement expressiveness. It designs experimental scenarios with typical gestures associated with body language. It is important to note that the evaluations in this thesis primarily assess hardware capabilities and do not aim to evoke genuine emotions in participants.

In addition to the selection of scenarios, this work also proposes a variety of multipositional and multimodal wearable prototypes. The idea is that different sensor positions and multiple sensing modalities help to form a unified perception and understanding of complex situations. Another relevant aspect is to recognize human behavior/activities pervasively. Wearable devices are the most promising ubiquitous human activity recognition (HAR) option. Creating wearable-based HAR solutions that are both small and widely accepted by users presents a significant challenge. In this context, a multidisciplinary approach is required. This includes expertise in sensor technologies, signal processing, data fusion algorithms, and domain-specific knowledge. One way to gain user acceptance is to deploy the HW/SW systems in the most common wearable accessories on the market. Hence, the designs presented here are based on wristbands, goggles, headwear (helmet and sports cap), and clothing (jacket and gloves). This work focuses on HW/SW co-design systems for HAR in the context of **Hand Position and Gesture Estimation**, **Head and Facial Muscle Movements Recognition**, and **Body Postures and Gesture Classification**. Considering these three scenarios, the thesis explores customized smart-wearable design with application-specific goals. The decision criterion in this work is based on two factors: how relevant the scenario is to understanding human behavior and how innovative the sensing technology is within the wearable community. Overall, the designs have been tested in various experimental settings with evaluation based on mimicked gestures. The experiments were designed to test the feasibility of the proposed hardware for solving specific scenarios. The main goal is to provide tools that can be used in the future to understand the “Expressiveness of Human Body Movements” in a ubiquitous way. Nonetheless, in future work, the experiments should be extended to include the emotional element of expressions. This is beyond the scope of this dissertation which focuses mainly on mimicked experiments without inducing emotional states in the participants.

1.1.1. Scenario Selection

1. **Hand Position and Gestures** provide valuable information for the recognition of human activity, mainly based on the fact that our hands are the main means of interaction with our physical environment, facilitating the manipulation of objects, as in robot-human interaction, and improving human-human communication among others. In particular, the vertical position of the hands around the body provides clues as to which group of human activities is taking place, since, when the hands are around the feet it is reasonable to filter out activities that

include hand-face interaction and vice-versa. Hence, hand-vertical position tracking can be the first step in a hierarchical scheme for HAR, reducing the solution complexity. Understanding the vertical position of the hands, along with other aspects of hand gestures and body language, is crucial for effective communication, especially in situations where verbal communication may be limited or ambiguous. This work tracks the vertical position of the user's dominant hand using atmospheric pressure sensors and radio frequency identification (RFID). The idea is evaluated in two scenarios. The first scenario tracks the person's hand (vertically) to monitor order-picking activities for stock management in warehouses; this is an example of the position of the hand relative to the environment and shows the precision of the method. The second one tracks the vertical position of the hand around the body, which can be used as contextual information for HAR. In addition, a subset of hand gestures is recognized with a multipositional and multimodal approach, using textile capacitive sensors and inertial information. Although the gesture dictionary is intended for drone control, the methodology is extensible to other applications in the fields of sign language, gaming, and robot control, among others.

2. **Head and Facial Muscle Movements** are delicate and valuable features for understanding human interaction with other humans/objects. The ability to move our face depends on the craniofacial muscles. The craniofacial muscles cooperate to regulate the movements of the cheeks, chin, eyebrows, eyelids, forehead, upper and lower lips, and nostrils. In particular, the temporalis and masseter muscles control facial expressivity and are activated by eating episodes. Thus, by detecting facial muscle movements, it is possible to monitor facial expression and, at some level, also monitor eating episodes such as chewing and swallowing. Facial expression monitoring can be used for various purposes, including early detection of facial muscle-related conditions and even assessing pain levels in non-verbal patients. Tracking the food intake helps to gain control over eating habits, and when facial expressions and eating monitoring are combined it would be possible to gain an understanding of stress-eating episodes. In Chapter 3, a set of publications for facial muscle monitoring is presented. The focus is on wearable-based designs with mechanomyography and inertial information fusion.
3. **Body Posture and Gestures** as control commands can improve accessibility for people with physical disabilities. It is a relevant area in a variety of applications, from improving human-computer interaction to improving health and fitness, security, education, entertainment, and collaborative work environments. The state of the art in the wearable community is primarily focused on motion capture with inertial measurement units (IMUs) deployed in elastic or fitted garments. As a counterpart, in this work, a loose-fitting clothing solution is proposed with the use of contactless, textile-based capacitive channels and RFID-based

synchronization.

In summary, with these scenarios, this work aims to provide tools that can be used in the future to understand the “Expressiveness of Human Behavior” in a ubiquitous way.

1.1.2. Modality Selection

Several sensing modalities were studied. The selection is based on the application and the novelty of the idea in the wearable community. Ubiquitous solutions for HAR must take into account power consumption, memory footprint, cost, dimensions, and user safety, among other aspects. In addition, a trade-off between the sensing modality and the position of the sensors on the body must be considered. This has led many solutions to recognize activity indirectly, which is the case of detecting facial muscle movement with sensors at discrete points on the face, as a result of which the entire facial surface is not detected. Moreover, it is a cumbersome or even impossible task to study all wearable sensing modalities. In this work, the scope of modalities includes atmospheric pressure, radio frequency identification (RFID), sound and pressure mechanomyography, and inertial and capacitive sensing. The selected modalities are energy-efficient, memory-efficient, low-cost, privacy-friendly, and easily integrated into commercial wearable garments/devices. In general, information fusion is performed in two ways, multipositional and unimodal fusion and multipositional and multimodal fusion. This section provides a summary of the selected modalities.

Atmospheric Pressure: Atmospheric pressure changes can indicate a vertical positional displacement. The relationship is described by the barometric formula, which quantifies the decrease in pressure with increasing altitude. To measure the atmospheric pressure a barometer is employed. A barometer measures the relative pressure concerning the prevailing atmospheric pressure. The altitude estimation using the barometric equation is influenced by factors such as temperature, humidity, and weather conditions. Usually, the barometer is placed in a stable position to provide a reliable measurement of altitude. Nowadays barometers are available in smartwatches and smartphones, providing contextual information about the user. Thus, they are suitable for low-cost and ubiquitous monitoring. This thesis demonstrates the feasibility of using barometric differences to determine the user’s vertical hand position. Barometer differential measurements are proposed. The difference between a semi-static barometer, the reference, and a moving target barometer is used to reduce the drift influence and the weather condition dependency on the measurements.

Radio Frequency Identification (RFID): Radiofrequency identification (RFID) is a fast and reliable option to automatically track the movements of goods through distribution. RFID systems consist of a reader and a tag. The reader generates radio waves to detect the presence of an RFID tag and then reads the data stored in it. Tags are embedded in cards, buttons or other small items. The RFID types are passive or active. An active RFID system

uses a tag with an embedded battery to increase the detection range, at the expense of higher power consumption, size, tag price, and varying report rates. A passive RFID tag does not include a battery, instead, it receives its energy to function from the reader, making it a low-power option. The tag gathers the electromagnetic energy from the reader to respond with its identification information. Passive tags have the benefit of being able to read at a fast rate, they are thin, allowing them to be placed between layers of paper, and cheap less than a dollar. All of the above makes the RFID system suitable for smart wearable designs. This work uses passive RFID technology to locate reference positions as a synchronization/calibration method for sensor nodes distributed over multiple positions.

Mechanomyography (MMG): MMG is a noninvasive method to examine muscle mechanical activity. The mechanical activity of the muscle is detected using specific transducers to record muscle surface oscillations due to muscle stretching/relaxation. As the MMG is a passive and mechanical sensing modality, it is intrinsically low power. IMUs, piezo-electric sensors, pressure sensors, and microphones can be used as MMG transducers. Pressure and Acoustic mechanomyography are the two main MMG methods of interest in this work. *Acoustic Mechanomyography (AMMG)* also called phonomyography is a technique to measure the force of the muscle contraction by recording the low-frequency sounds resulting from muscular activity. Typically, the signal is measured using condenser microphones attached to the skin. AMMG is also employed to measure muscle activities underwater with the use of hydrophones. Sound can be captured at a higher sampling rate compared to other transducers, leading to a reduced response time. A trade-off between response time and power consumption should be considered. *Pressure Mechanomyography (PMMG)* is another approach for MMG. In this work, piezo-electric film and force-sensitive resistor sensors are employed as transducers for this type of MMG. A key advantage of MMG is that it does not require precise positioning of the transducer to monitor the muscle activity, this is due to the propagation property of the muscle's mechanical activity. It is considered a low-power, low-cost technique and although it requires contact with the skin, it does not need to penetrate it, so it is also called superficial MMG. Flexible/elastic garments are a suitable option for MMG wearable-based designs.

Acceleration and Orientation: IMUs-based wearables are widely explored in the state of the art for HAR. The inertial information plays a crucial role in HAR. This is due to its ability to provide fine-grained, real-time, and versatile information about motion patterns, enabling applications across diverse domains. IMUs are considered energy efficient. This makes them suitable for battery-powered wearable devices that need to operate for extended periods without frequent recharging. In this work, the inertial information is then employed for multipositional and multimodal fusion to enhance the accuracy and robustness of HAR system by compensating for limitations and errors inherent in individual sensor types.

Capacitive Sensing: Capacitive transducers work on the principle of

change in capacitance between conductive plates. Capacitance is typically measured indirectly, by using it to control the frequency of an oscillator or to vary the level of coupling (or attenuation) of an AC signal. It is a versatile technology for HAR. Its touch and proximity detection capabilities make it valuable for creating interactive and context-aware systems in smart environments, wearable and other technological applications. Moreover, capacitance can be non-contact-based, which is a desirable characteristic for loose garment-based wearables. This work focused on using capacitive sensing for body motion detection with textile and loose garments.

This thesis explores a wide range of modalities for HAR. The selection criterion is based on the specific scenario and the novelty of the idea in the wearable community. The details are discussed in the chapters below with the respective publications.

1.2. Contribution

In this section, a general summary of the contributions made by this thesis is presented.

- This work demonstrated vertical position tracking of the user's hand with differential atmospheric pressure fusion (between two barometric sensors), where a reference barometer in the pocket simulates a smartphone and another barometer on the user's wrist simulates a smartwatch. This multipositional fusion is then multimodally fused with an RFID signal to synchronize the barometer pair and reduce signal drift, obtaining a vertical hand position tracking of $\leq 30\text{cm}$ range.
- Capacitive textile-based gloves for hand gesture recognition were introduced. The capacitive channels are fused with inertial sensing modality hierarchically to reduce power consumption and increase robustness against the null class, where the first stage detects movements and recognizes a non-null hand gesture using an inertial model. Then, using a capacitive model, the second stage classifies a set of hand gestures for applications such as drone control. The solution includes real-time and on-the-edge deployment, demonstrating the flexibility, low power consumption, and low price of the idea. The evaluation is based on gestures for drone control but has potential applications in sign language, gaming, and robot control.
- The thesis also proved the idea of using differential sound analysis to unobtrusively acquire information about facial muscle activity patterns that can be associated with facial expressions and actions. The acquisition of the sound signal is based on stethoscope microphones distributed around the masseter, temporal, and frontal muscles to perform acoustic mechanomyography (AMMG) on the face, which is a novel contribution in terms of application and detection modality. In addition, facial muscle detection was extended to a multimodal fusion between AMMG,

pressure mechanomyography (PMMG), and inertial data fusion with a sex-balanced dataset with eight nationalities to gain inclusivity and generability of the solution. Continuing the work, the multimodal fusion between PMMG and inertial data solution was run on the embedded hardware and evaluated in real-time to detect facial expressions and eating episodes.

- The work continued with the introduction of a wearable approach for detecting body postures and gestures (BPG) that does not require sensors to be firmly fixed to the body or integrated into a tight-fitting garment. Instead, the sensing is embedded into a loose-fitting garment (a formal men's jacket). For this purpose, the famous musical instrument theremin (capacitive sensing) was adapted as a sensor and integrated into the jacket. The theremin antennas, capacitive electrodes, were textile cables sawed into the jacket lining without altering the jacket design structure, for a complete wearable and loose-fitting garment solution for BPG. The idea is then multimodally fused with RFID synchronization to automatically capture the start and end of gestures hierarchically, reducing the power consumption and increasing robustness against the null class.

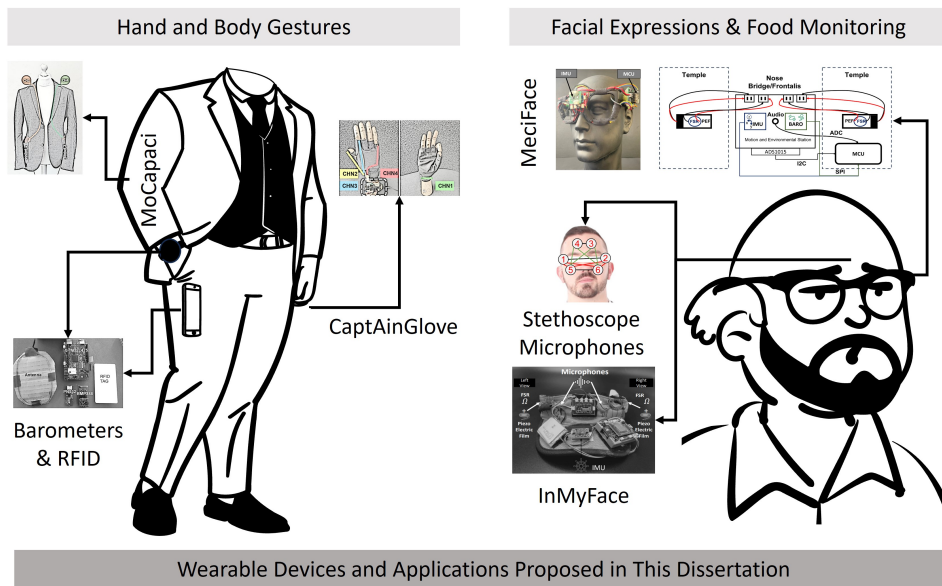


Figure 1.2.: *Wearable Devices and Applications Proposed in this Dissertation.*

Fig. 1.2 depicts an overview of the hardware designs and applications pursued by the work. The following chapters will expand on each of the above contributions with a more detailed list based on each of the publications in the thesis.

1.3. Related Work

1.3.1. Hand Position and Gesture Recognition with Smart-Wearable Devices

The position estimation of the user’s hands is a multidimensional, and complex problem. In this section, the focus is on a subset of the hand-tracking problem: the detection of the vertical position of the wrist. The state of the art in vertical position estimation is mainly based on IMU sensors, at least in the wearable community. IMUs can be used together with the appropriate biomechanical models to track the exact position and orientation of the hand. However, given the degrees of freedom of human joints, this cannot be achieved by a wrist-worn sensor alone but requires at least three sensors: one at the wrist, one at the upper arm, and one at the torso (at least for an exact solution), and under a stable magnetic environment. Nowadays, IMUs-based tracking has been combined with neural network-based algorithms, reducing the number of sensors and improving performance.

Table 1.1.: *Non-Visual and Wearable-Based Solutions for Hand Position Tracking*

Date-Work	Highlights
03-2008[48]	<ul style="list-style-type: none"> • Sensing: IMU and Ultra Wide Band (UWB) • Wearable: GypsyGyro-18 Suit and Ubisense Tag • Scenario: Body Position Tracking • Users: 1 Participant • Results: 0.14 m Error with Kalman Filter
03-2018[196]	<ul style="list-style-type: none"> • Sensing: RFID • Wearable: RFID Tags Around The Body • Scenario: Upper-Body Position Tracking • Users: 5 Participants. • Results: 4.4 cm Relative Position Error with Kinect 2.0 Testbed.
10-2019, Ours[16]	<ul style="list-style-type: none"> • Sensing: Barometer and RFID • Wearable: Wristbands • Scenario: Vertical Hand Position Tracking. • Users: 2 Scenarios, 3 and 5 Participants, respectively. • Results: 84.15 % and 60.25 % average accuracy.
10-2019 [26]	<ul style="list-style-type: none"> • Sensing: IMU • Wearable: Smartwatches (Left/Right) • Scenario: Moving, Upper, Lower, and Away Positions • Users: 6 Participants • Results: 97.00 % accuracy
04-2023 [136]	<ul style="list-style-type: none"> • Sensing: IMU • Wearable: Smartwatches (Left/Right). Smartphones(Left/Right Pockets). and Earbuds (Left/Right). • Scenario: Body Position Tracking. BLSTM • Users: 10 Participants • Results: 12.08 cm Mean Per Joint Vertex Error (MPJVE)

In the Table 1.1, a summary of wearable-based solutions for tracking hand position is compared. Multipositional IMU-based systems are the ones with higher accuracy in the classification task and lower error in the tracking tasks. The inertial information has also been employed in combination with Ultra-WideBand (UWB) in [48]. The authors employed a biomechanical model with a Kalman filter as the backbone of the multimodal fusion (IMU and UWB). This is an example of the classical approach to body tracking. It uses multiple

sensors around the body and relies on an accurate mechanical model, which is not always available and is usually oriented to linear systems.

To learn the non-linear patterns in the inertial data it is possible to use a neural network as in [136]. Recently, in [136], the authors estimate the full-body pose using three distributed IMUs. The IMUs' positions were in a smartwatch, smartphone, and earbuds. The system tracks which devices are present and uses the available IMU data. With one IMU device available, the Bidirectional Long Short-Term Memory-based (BLSTM) system has a Mean Per Joint Vertex Error (MPJVE) of 16.27 cm with a standard deviation of 9.93 cm. The error is further reduced when three IMUs are available, achieving an MPJVE of 11.1 cm with a standard deviation of 6.51 cm. Which are promising results toward an exact body position tracking with a reduced number of IMUs. In [196], a wearable RFID is proposed. The idea offers the advantage of using RFID Tags sticker to clothing around the body and it is also training free. The authors track the body movement in 3D space by analyzing the phase information of wearable RFID Tags combined with a human geometric model and a Kalman Filter. And obtained a 4.4cm relative position error for the position estimation of joints compared with the Kinect 2.0 testbed. The main drawback of the RFID solution presented in [196] and the IMU-UWB fusion presented in [48] is that both systems require a receiver/transmitter deployed in the environment, so they are not complete wearable approaches.

In summary, the literature can be split into mathematical-based modeling or data-based modeling. The former depends on how accurate the mathematical model is and the latter depends on big data and long training requirements. In [16], this work, demonstrated the use of relative height estimation between pressure differences and the RFID as a monitor-error method to detect the vertical position of the hand. Wrist elevation is tracked by comparing the signal from the wrist barometer(simulating a smartwatch) to the reference (simulating a smartphone) and using the barometric formula. This goal was achieved by only using a simple linear model and an RFID chip to calculate the initial offset and to reduce the impact of the drifting and offset in the vertical position estimation, in addition to the reduction of the effects of sudden changes on pressure, due to the opening of windows or doors around the devices. The details of the proposed solution are further discussed in Chapter 2.

Another focus of this work is on hand gesture recognition with wearable and textile-based solutions for drone control. We can find several textile-based sensing modalities on gloves for drone control in the literature, as shown in Table 1.2. Researchers have employed various sensing technologies, including textile pressure sensors, triboelectric nanogenerators (TENG), flexible capacitive pressure sensors, piezoresistive, and conductive fiber-based textile pressure sensors, among others [169], [5], [186], [111]. A common practice among state-of-the-art textile wearable glove alternatives is to use different and limited gesture dictionaries compared to camera-based solutions, which mainly include going forward/backward and going to the left/right classes. For the capacitive or textile pressure sensor options, the textile is used as a soft push button, where each textile patch is an on/off instruction. Textile as binary

actuators (push buttons) is a simple and effective way of Human-robot interaction (HRI). However, it lacks robustness against the null class, and the number of control instructions is limited to the number of textile patches. This is an issue if the drone/robot requires performing more sophisticated tasks [93]. For example, in [111], the patches are placed on the fingertips, leading to the incorrect recognition of null activities, such as checking a smartphone, typing on a keyboard, or touching/grabbing tools, as control instructions for the drone/robot. In [5], the authors employed a capacitive textile sensor on the back of the palm of the right/left hand, and with the other left/right hand, the user touched the patches to generate the control signal for the drone. However, this solution does not account for the potential extension of gestures by including both hands in the pipeline. In [169], the authors used TENG to fabricate a sensing glove for gesture recognition, including sign language and drone control applications. However, their focus was mainly on introducing the technology and its futuristic applications, thus neglecting null activities and lacking information about user experimental evaluations.

To overcome these limitations, we introduce a capacitive and inertial fusion-based glove-based design for real-time on-the-edge hand gesture recognition. Our design incorporates textile capacitive electrodes as sensing channels on the fingers and an IMU sensor on the wrist. Textile capacitive sensing has demonstrated its effectiveness as a low-power consumption, cost-effective, and scalable technology for movement tracking in gesture and activity recognition [22], [23], [220]. Furthermore, IMU sensors have been widely used to monitor wrist movements by researchers [121], [177], [89]. Our approach is an alternative solution for hand gesture recognition that addresses critical requirements, including minimal invasiveness, low power consumption, privacy preservation, flexibility, and scalability. While our focus is mainly on a gesture dictionary related to drone control, the same concept can be extended to hand-based HRI in various domains such as sign language, gaming, and robot control.

Table 1.2.: *Non-Visual-Based Hand Gesture Solution with Wearable Garments*

Date-Work	Highlights
02-2015 [91]	Sensing: Air pressure sensors Wearable: Air bladder band Scenario: 6 Hand gestures Users: 6 Participants Results: Custom Fuzzy Logic. 90.00%
05-2021 [158]	Sensing: 8 Nanocomposite Pressure Sensors Wearable: Stretchable Textile Tape Scenario: 0 American Sign Language Numbers Users: 10 Participants. Results: User-dependent. Extreme Learning Machine (ELM). 93.00 %
05-2021 [172]	Sensing: EGaIn-Silicone Soft: React to pressure or stretch Wearable: Glove Scenario: 12 Static hand gestures Users: 15 Participants. Results: User-dependent. Random Forest. 97.30%
06-2019 [217]	Sensing: Surface EMG Wearable: Armband, commercial Scenario: 5 hand gestures Users: 12 Participants Results: User-dependent. ANN 98.70 %
10-2023 and 02-2024, Ours [17, 18]	Sensing: Capacitive and IMU Wearable: Glove Scenario: 8 Hand Gestures for Drone Control Users: 1 Participant. User-dependent. Results: 80.00 % and 67.00% F1-Score, offline and Real-Time on-the-Edge, respectively.

1.3.2. Facial Muscle and Head Movements Recognition with Head-Mounted Accessories

Monitoring of head and face muscle activity has been explored with the use of several sensing modalities. Detection methods range from camera, light detection, piezoelectric phenomenon, sound, and electromyography (EMG) to textile pressure mechanomyography. Cameras are the most widely used solution to monitor facial activities with an expected accuracy of more than 90.00 %. In wearable applications, cameras are not always a desirable solution. The camera placed in front of the face brings privacy concerns for everyday use. The constant movements of the user, the light conditions, and occlusion can represent another problem for visual-based systems.

In the Table 1.3, the state-of-the-art visual-based wearable solutions for facial activity recognition are compared. The visual solutions presented in the Table 1.3 include cameras, infrared cameras, and photo-reflective sensors. The positioning of the visual-based sensors is highly relevant, thus the sensor captures the face area without hindering the user visibility. The photo-reflective approach is a clever idea deployed within the frame of smart glasses. Hence, the user visibility is not compromised, and it is a privacy-aware sensing modality. Still, the photo-reflective solution is sensitive to light conditions. Later on, an IMU is combined with photo-reflective sensors to improve the robustness of the system. However, it still relies on an active (photo-reflective) detection mode. Active sensing modalities are those that radiate signals around the human body. Therefore, an active detection method around the eyes for ubiquitous use may raise some concerns about the user's long-term health. Further studies should be conducted to ensure that targeted active detection is safe for the user.

On the other hand, there has been significant interest in the wearable community in non-visual-based sensing approaches for head and face movements monitoring. The main challenge faced by non-visual approaches is using limited areas to capture facial information, preferably around glasses, on the head (e.g., using a cap or helmet), and around the ears or nose. In the Table 1.4, the state-of-the-art wearable and non-visual-based solutions for facial activity recognition are compared. More than 46 % of the studies in the Table 1.4 are based on earpieces. Nowadays is considered normal to use headphones/earbuds every and almost all day, making the ear the new wrist in the wearable community. Around the ear is a non-obstructive position to acquire facial movement information. The earpiece and ultrasound-based systems are the most promising methods for facial movement tracking. Having multiple channels of ultrasound signals sent to the face and ear area every day and all day may compromise the user's acceptance of the method. Besides, experiments on animals have shown damage to internal organs from receiving different ultrasonic frequencies [138].

For the non-camera-based approaches, higher accuracy is obtained by the active sensing methods. This is the case for the photo-reflective signal and the ultrasound signals broadcasted around the user's face. In the present work, we focus on passive solutions such as the one presented in [221], where a textile

Table 1.3.: *Visual-Based Wearable Solutions for Facial Activity Recognition*

Date-Work	Highlights
03-2016 [129]	Sensing: 17 Photo reflective Sensors Wearable: Glasses Scenario: 8 Expressions. Support Vector Machine (SVM). Users: 8 Participants Results: User-dependent Average 92.80 % accuracy. One session
10-2020 [42]	Sensing: Cameras Wearable: 2 Ear Mounted Cameras Scenario: 8 Emojis. BLSTM Users: 9 Participants Results: User-dependent 88.60 % Accuracy
10-2020 [105]	Sensing: Camera Electrodermal Activity (EDA) Photoplethymogram (PPG) Wearable: Scenario: Recognition of the Quadrant of the Arousal-Valence Plane Support Vector Machine Users: 20 Results: User-dependent 76.09 % Accuracy
11-2020 [127]	Sensing: 16 Photo Reflective Sensor and IMU Wearable: Glasses Scenario: Eight Temporal Gestures. CNN Users: 13 Participants Results: User-dependent 91.10 % F1 Score.
02-2021 [10]	Sensing: Cameras Wearable: Glasses with 2 Cameras Scenario: 7 Pseudo Face Images Comparison with Stored Images Users: 6 Participants Results: User-dependent. Accuracy 87.50 % Neutral, 66.70 % Happy and 71.40 % Surprise
06-2021 [41]	Sensing: Infrared Cameras Wearable: Necklace and Neckband with Cameras Scenario: 8 Customized Facial Movements CNN. Real-Time Tracking Users: 13 Participants Results: MAE 30.29 Necklace and 25.61 Neckband
11-2021 [60]	Sensing: IR Sensor Array Wearable: Ear Accessories Scenario: 9 Facial Gesture. Support Vector Machine Users: 5 Participants. Results: User-dependent 97 % F1-Score

pressure matrix (mechanomyography) was introduced into a headband design to classify seven facial expressions, with intermediate results of up to 38.00% accuracy. Moreover, piezoelectric thin films (PEF) have been used in real-time [188] to detect and classify skin deformation to decode facial movements in patients with amyotrophic lateral sclerosis. Their work is intended to be used in clinical settings for nonverbal communication and neuromuscular monitoring conditions. PEF sensing technology is lightweight, customized, and with mechanical harvesting capability [168]; therefore, we could claim that PEF technology is worthy of research and study in specific applications.

The work in this thesis and for the scenario of head and facial muscle move-

Table 1.4.: *Non-Visual-Based Wearable Solutions for Facial Activity Recognition*

Date-Work	Highlights
10-2017 [9]	Sensing: Barometer Wearable: Earphone Scenario: 11 Face Related Movements Users: 12 Participants. Random Forest Results: User-dependent 87.60 % accuracy
09-2019 [7]	Sensing: Ultrasound Wearable: Earphones with Speaker and Microphones Scenario: 21 Customized Expressions. Support Vector Machine. Users: Eleven Participants Results: User-dependent Average 62.50 %
03-2020 [115]	Sensing: IMU and Electrooculography Wearable: Glasses Scenario: 9 Kissing Gestures and Walking Users: 5 Participants Results: User-dependent 74.13 % accuracy
08-2020 [221]	Sensing: Pressure Matrix Wearable: Headband with Pressure Sensors Scenario: 7 Facial Activities. Support Vector Machine. Users: Twenty Participants Results: User-dependent 37,80 % F1 Score. One Session.
10-2020 [21]	Sensing: 6 Microphones Wearable: Helmet with Stethoscope Microphones Scenario: 10 Facial Activities. Support Vector Machine. Users: 8 Participants Results: User-dependent 75.37 average % F1 Score.
10-2020 [188]	Sensing: Piezoelectric Films Wearable: Flexible Films Stickers Scenario: Smile, Twitch, and Pursed Lips. KNN-DTW Users: One healthy and One Patient with Amyotrophic Lateral Sclerosis Results: User-dependent Average 86.80 % and 75.00 % accuracy. One session.
02-2021 [130]	Sensing: Capacitive Sensors Wearable: Glasses Scenario: 12 Facial and Head Gestures Random Forest Users: 10 Participants Results: User-dependent 89.60 % accuracy
09-2021 [192]	Sensing: IMU Wearable: Earphones with IMU Scenario: 30 Action Units. Temporal Convolutional Network Users: 12 Participants Results: User-dependent 89.90 % accuracy
09-2021 [63]	Sensing: IMU Wearable: Earbuds with IMU Scenario: 3 Head Movements. Hierarchical classification. Users: 21 Participants Results: User-dependent 84.79 (smile, talk, and yawn)% accuracy
05-2022 [181]	Sensing: Ultrasound Wearable: Headphones Scenario: 7 Facial Expressions. LSTM Users: 5 Participants Results: User-dependent 80.00 % accuracy
07-2022 [113]	Sensing: Microphones Wearable: Microphones and Speaker in Earphones Scenario: Customized Facial Movements Users: 12 Participants Results: User-dependent. MAE 25.90
10-2022 [67]	Sensing: Surface Electromyography sEMG Wearable: Pico Virtual Reality Headset with EmteqPro Sensors. Scenario: Neutral, Negative, and Positive Valence and Arousal Scores. Users: 38 Participants. Pearson's Correlation Coefficient. Results: Significant Values -0.24 to 0.64.
06-2023 [15]	Sensing: IMU, FSR, Piezoelectric, and Microphones Wearable: Sportscap Scenario: 9 Head and Facial Gestures Users: 13 Participants Results: CNN User-dependent 85.00 % F1 Score
10-2023 and -2024 [19, 20]	Sensing: IMU, FSR and Piezoelectric Wearable: Glasses Scenario: 5 Head and Facial Movements and Eating Episodes Users: 1 and 6 Participants, respectively Results: CNN User-dependent 86.00 %. Expressions User-independent 94.00 % F1-Score. Eating Episodes

ment monitoring is presented in three main publications. In [21] acoustic mechanomyography (AMMG), a passive technique was employed with an F1 score of 75.37 % average in classifying ten facial muscle activities. In [15] a fusion between passive sensing is studied. The fusion consists of sensing such as pressure mechanomyography (PMMG) using a force-sensitive resistor (FSR) and piezoelectric film (PEF), inertial sensing based on orientation and acceleration, and acoustic mechanomyography (AMMG). In [19], [20] PMMG, PEF, and inertial information are fused to monitor facial muscle movements expected to appear in facial expressions and eating episodes. The wearable accessories used were a helmet, a sportscap, and glasses for each of the studies, respectively. The detailed information is presented in Chapter 3.

1.3.3. Body Posture and Gesture Recognition with Wearable

Body posture and gestures (BPG) are key components of human activities and are essential ways to convey emotion and personality, implicit social interactions, sign language, etc. As a result, BPG recognition has been one of the first wearable sensing applications, leading to many mature commercial applications for motion capture. BPG recognition methods are widely explored in the literature. Most popular wearable BPG sensing techniques use inertial measurement units (IMU) and, on the textile side, stretch sensors. While highly effective in many applications, most current systems share one limitation: they require sensors to be firmly fixed to the body through tight garments or dedicated accessories, such as bracelets and straps.

Inertial measurement units (IMU) distributed in clothing or accessories for BPG recognition is a widely used technique [32, 74, 164]. Another relevant approach for BPG analysis is called kinesiological electromyography (EMG) [47, 217]. Such approaches are reliable and robust solutions with accuracy above 90.00 %. However, both detection modalities require stable sensor positions to avoid the effect of noise and motion artifacts on the signals. Furthermore, the placement of discrete and rigid sensors around the joints could be uncomfortable for the user. In [119] the authors employed 100 microchips with memory and temperature sensors interconnected in a flexible fiber on a T-shirt, which is a solution to increase the flexibility and comfort of the user while wearing discrete sensors, a promising idea to explore in the future.

On the other hand, stretchable garments with strain-based or pressure sensing methods have been studied by many researchers [28, 91, 116, 134, 158, 172, 179, 221], which demonstrate their value in textile-based BPG recognition. Fiber optic embedded in a jacket and pants was proposed in a limited study (one person) [102]; the transmitted light changes with the wearer's movements, creating a time series pattern due to the bending of the fiber optics. The wearable optical technology is growing rapidly with multiple hardware designs being proposed [4, 101, 102, 107, 112, 189, 214]. A fabric-based triboelectric sleeve is proposed in Kiaghadi et al. [96]. Four Radio Frequency Identification (RFID) tags were proposed on the back, chest, and feet over the persons' clothes and shoes in [198] to recognize a total of eight activities (standing, sitting, walking, along with others). The piezoelectric effect was employed in

[35], where four flexible piezoelectric sensors were placed on the knee and the hip in slack pants to detect walking, standing, and sitting activities.

A comparison between the BPG recognition approaches is presented in Table 1.5. This work, introduced a quick and easy option to integrate e-textile components in loose-fitting garments [22]. The proposed idea uses commercial conductive textile parts as the antennas of the modified off-the-shelf theremin (OpenTheremin) based on capacitive sensing. In [23] the work is extended from multipositional fusion to multimodal fusion. The idea is then fused with Radio Frequency Identification (RFID) synchronization for real-time and wireless recognition of six classes of a dance movements dictionary. Further details are presented and discussed in Chapter 4.

1.4. Outline

The current thesis consists of five chapters. In this first chapter, the motivation and the contribution overview as well as the current state of research in the scientific area related to the thesis are presented. The motivation includes the reasoning about the scenario and the sensing modalities selection. The related work is separated by application-specific scenarios. Each of the three subsequent chapters (Chapter Two, Chapter Three, and Chapter Four) presents the contributions of each application-specific scenario. Chapter 2 focuses on hand position and gesture recognition with smart-wearable devices. Chapter 3 summarizes the contributions in the area of facial muscle and head movement estimation with mechanomyography and inertial fusion. Chapter 4 comprehends the works related to body posture and gesture recognition with multipositional textile capacitive sensing-based fusion. Finally, Chapter 5 summarizes the main conclusions, analyzes the research articles presented, and offers implications and recommendations for future research.

Table 1.5.: *Non-Visual based BPG Solution with Wearable Garments*

Author-Year	Highlights
02-2008[74]	Sensing: 3 Accelerometers Wearable: Long Sleeve Shirt Scenario: 12 Arm Movements Users: 8 Participants. Results: User-dependent. Nearest Centroid Classifier 95.00 % Accuracy
06-2015 [73]	Sensing: Capacitive Wearable: Leg/Chest Band, Insole Scenario: 5 Motor Activities Users: 10 Participants Results: User-independent. Bayesian Classifiers. 88.97 % Accuracy
03-2016 [198]	Sensing: RFID Wearable: 4 Antennas; Aack, Chest and Feet Scenario: 5 Motor + 3 Cleaning Activities Users: 4 Participants Results: User-dependent. SVM 93.60 % Accuracy
02-2018[35]	Sensing: Flexible Piezoelectric Wearable: Loose Pants Scenario: 5 Motor Activities + 8 Transitions Users: 10 Participants. Results: User-dependent. Rule-Based Algorithm [36]. 93.00% Accuracy
09-2018[96]	Sensing: Fabric-Based Triboelectric Joint Sensing Wearable: Sleeve Scenario: Brushing, eating, walking, idle Users: 14 Participants Results: User-dependent. SVM 91.30 % Accuracy
10-2018 [102]	Sensing: Hetero-core Fiber Optics Wearable: Jacket and Pant Scenario: 8 Motor Activities Users: 1 Participant. Results: User-dependent. SVM 98.70 % Accuracy
10-2018 [179]	Sensing: Textile Pressure Sensors. Wearable: Trousers (3 sizes). Scenario: 19 Sitting Postures/Gestures. Users: 6 Participants. Results: User-dependent. Random Forest. 99.18 % Accuracy
01-2020 [221]	Sensing: Textile Pressure Matrix (TPM) Wearable: Elastic Sport Band Scenario: 4 Gym Exercises + 3 Non-Exercises Users: 6 Participants. Results: User-dependent. ConfAdaBoost. 93.30 % Accuracy
04-2020 [116]	Sensing: Optical-strain sensor Wearable: Sweat Jacket Scenario: Standing, Sitting, Lying, Walking, Running Users: 12 Participants Results: User-dependent. CNN-LSTM 90.90 % Accuracy
06-2021[119]	Sensing: Flexible Fiber Wearable: Shirt with 100 Microchips with Temperature Sensors. Scenario: Sit, Stand, Walk and Run Users: 1 Participant. Results: User-dependent. CNN 96.40 % Accuracy
09-2021[22]	Sensing: Capacitive Wearable: Loose-Fitting Jacket Scenario: 20 Posture/Gestures Users: 14 Participants Results: User-dependent. Conv2D. 97.18%. User-independent 86.25% Accuracy
06-2022[23]	Sensing: Capacitive Wearable: Loose-Fitting Jacket Scenario: 8 Dance Movements Users: 3 Participants Results: User-dependent. 1DConv. 92.00% F1-Score

Chapter 2

Hand Position and Gestures Recognition with Smart-Wearable Devices

Contents

2.1. Hand Vertical Position Estimation with Atmospheric Pressure and RFID synchronization	18
2.1.1. Problem Statement	18
2.1.2. Contributions	19
2.1.3. Apparatus	19
2.1.4. Signal and Data Processing	19
2.1.5. Experiment Design	23
2.1.6. Results	25
2.1.7. Discussion	26
2.1.8. Conclusion	28
2.2. Capacitive and Inertial Fusion-Based Glove for Real-Time on Edge Hand Gesture Recognition	28
2.2.1. Problem Statement	28
2.2.2. Contributions	29
2.2.3. Apparatus	29
2.2.4. Signal and Data Processing	30
2.2.5. Results	33
2.2.6. Discussion	34
2.2.7. Conclusion	36

The author of this thesis has published the content, figures, and tables included in this chapter in the following publications:

Bello, H., Rodriguez, J., & Lukowicz, P. (2019, October). Vertical hand position estimation with wearable differential barometry supported by RFID

synchronization. In EAI International Conference on Body Area Networks (pp. 24-33). Cham: Springer International Publishing.

Bello, H., Suh, S., Geißler, D., Ray, L. S. S., Zhou, B., & Lukowicz, P. (2023, October). CaptAinGlove: Capacitive and inertial fusion-based glove for real-time and on-the-edge hand gesture recognition for drone control. In Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing (pp. 165-169)

Bello, H., Suh, S., Geißler, D., Ray, L. S. S., Zhou, B., & Lukowicz, P. (2024; February). Real-Time and on-the-Edge Multiple Channel Capacitive and Inertial Fusion-Based Glove. In: Mizmizi, M., Magarini, M., Upadhyay, P.K., Pierobon, M. (eds) Body Area Networks. Smart IoT and Big Data for Intelligent Health Management. BodyNets 2024. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 524. Springer, Cham. https://doi.org/10.1007/978-3-031-72524-1_13 **Best Paper Award**

2.1. Hand Vertical Position Estimation with Atmospheric Pressure and RFID synchronization

2.1.1. Problem Statement

Our hands are the primary means of interacting with our physical environment. Thus, the position of the user's hands is a crucial piece of information for a broad range of context recognition tasks. It is made difficult by two considerations. First, in many cases, to be meaningful, the tracking has to be accurate to within 10–50cm. This is, for example, the case when we need to know which object the user has picked from a shelf, which/how she/he has interacted with a household device, or when he/she has taken a piece of food into the mouth. Second, for many applications, the amount of instrumentation that can be introduced into the environment to facilitate the tracking is limited. Ideally, the tracking would be achieved by a sensor that can be easily integrated into a smartwatch or a fitness tracker without needing further environmental instrumentation. Furthermore, understanding the vertical position of the hands, along with other aspects of hand gestures and body language, is crucial for effective communication, especially in situations where verbal communication may be limited or ambiguous. To this end, this work tracks the vertical position of the user's dominant hand using atmospheric pressure sensors and RFID synchronization techniques. The idea is evaluated in two scenarios. The first scenario tracks the person's hand (vertically) to monitor order-picking activities for stock management in warehouses; this is an example of the hand's position relative to the environment and shows the precision of the method. The second one tracks the vertical position of the hand around the body, which can be used as contextual information for human activity recognition (HAR).

2.1.2. Contributions

The main contribution of the idea is summarized as follows:

- We demonstrated how a combination of wrist-worn and stationary barometer (differential pressure) can be used to track the vertical position of the user's hand with an accuracy of 30 cm.
- We propose a simple linear model based on the barometric formula to calculate the altitude. Differential atmospheric pressure sources from a static and a wearable position are used to calculate the relative altitude. And, an RFID-based synchronization technique is employed to obtain the initial offset between the wearable and stationary barometer. These techniques help reduce the impact of drift and the effects of sudden changes in pressure due to open windows or doors around the devices.
- We evaluated the system in two different scenarios: an order-picking scenario in a warehouse, and a movement of the arm to specific body locations scenario. In the first scenario, a six-level shelf is employed and the altitude of the user's hand is tracked. The altitude is later used to locate the user's position hand according to one of the shelf levels. In the second scenario, the categories to be classified were; hand on the head, chest, and feet. Despite the simplicity of our method, it shows initial results of 60.25% and 84.15% average accuracy, respectively.

2.1.3. Apparatus

The hardware design is based on the development board STM32L475 DISCOVERY with an Arm Cortex-M4, running the real-time operating system, MBED version 5.12. The sensor used to measure the atmospheric pressure was the barometer BMP388 from Bosch company. The radio frequency communication (RFID) system consists of a PN532 reader and a MIFARE tag (13,56 MHz). The barometer and the RFID reader use the I2C protocol at 400 KHz to communicate with the microcontroller (MCU). The MCU is connected to a PC via the ST-LINK-UART protocol at a baud rate of 1 Mbps. The complete system is depicted in Fig. 2.1**Left**. The sampling rate of the data acquisition system depends on the schedule of the RTOS MBED. Fig. 2.1 **Right** depicts the variability of the sampling time with 90.00% of time being less or equal to 62.5Hz (16 ms). Two copies of the hardware design were made, one for the person's wrist and one for a stationary reference system (on a table or the user's pocket).

2.1.4. Signal and Data Processing

In this section, the procedure to obtain the linear model for the barometric-based hand vertical position estimation is discussed. It also presents the details of the conducted experiment based on the two scenarios; an order-picking scenario in a warehouse, and a movement of the arm to specific body locations

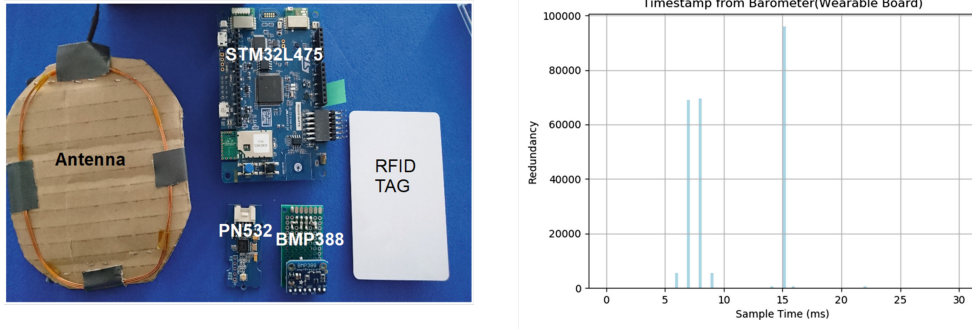


Figure 2.1.: *Left* The Hardware Prototype Shows the Antenna, Microcontroller (STM32L475), the RFID-tag, RFID-Reader(PN532), and the Barometer (BMP388). *(Right)* Depicts The Sample Rate Distribution for the Pressure Data in The Real-Time Operating System MBED.

scenario. Finally, the classification method selection as a naive Bayes classifier is justified.

Linear Model Procedure

The vertical hand position modeling is based on the barometric formula [54, 110, 200], as shown in Eq. (2.1). Where P is the pressure at a certain altitude H , P_0 is the pressure at a reference point, M is the molar mass of dry air, g is the gravitational field strength, R is the gas constant of air, and T_0 is the temperature [207]. From Eq. (2.1) it is possible to get the height for a given pressure as shown in Eq. (2.2). This is a non-linear relation between the pressure and the height. The linear version of this equation is in Eq. (2.3). Where, β is the temperature elapsing under 11km, with a value of $\beta = -6.5 \times 10^{-3}$ [200]. This is an equation that depends on a reference value. The sea level pressure is usually used as the reference pressure value, as in [200]. The authors used a base barometer and a rover barometer and made the height estimation relative to the sea level pressure.

$$P = P_0 * \exp \frac{-g * M * H}{R * T_0} \quad (2.1)$$

$$H = -\frac{R * T_0}{g * M} \ln \frac{P}{P_0} \quad (2.2)$$

$$H = \frac{T_0}{\beta} \cdot \left[\frac{P}{P_0} \frac{-\beta * R}{g} - 1 \right] + H_0 \quad (2.3)$$

Since the air pressure varies typically between 950 and 1050 hPa during a year, the expected variation in sea level due to air pressure is between +63 cm and -37 cm around mean sea level, which is a situation that we cannot quantify or control. In [151] the authors studied the sources of errors in barometer pressure measurements and concluded that barometers in differential mode would provide a very accurate altitude solution, but local disturbances in pressure must be taken into account in the design of the application. We proposed to

apply the formula 2.3 but, with a stationary reference pressure and a wearable pressure device, eliminating the dependence on sea level pressure, and having access to monitor changes in reference pressure values. The altitude equation also depends on the temperature. In [110], the authors set the temperature value as an average between the temperature at the standard atmospheric pressure and the current pressure point. Following this logic, we set the temperature value as the average between the stationary pressure point and the variable pressure point, which is the pressure given by the wrist-worn device.

The aim is to use the differential altitude between the stationary and variable pressure points to track the vertical hand position. However, even two barometers of the same version and on the same height position possess greater noise levels and variability. The idea is to find a linear relationship between the pressures to obtain the simplest differential solution possible. Thus, we proceed to do a set of experiments, including coherence values and stationary tests to guarantee that a linear fitting is suitable to represent the relation between the two devices.

Coherence Pressure Test: We developed two copies of the hardware design. One is used as a fixed reference point and the second one as a movable pressure point. Both devices were placed on the same table (same height) and close to each other to record the pressure values. This setting aims to obtain the time series static information. The experiment was repeated three times on different days and the duration of each recording was between 20 to 45 minutes as depicted in Fig. 2.2. These recordings were done indoors without interruptions and windows and doors closed.

Fig. 2.3 depicts the coherence between the pressures of the pairs of devices. The coherence results show a relation between the two pressures in the low-frequency band. Thus, an exponentially weighted low pass filter is applied. This technique increases the similarity between the pressures, improving the precision of a linear fitting between the two pressures.

Stationary Test: The precision of the tracking is required to be less than one meter to be relevant. However, the variation of only one pascal in the pressure means an eight-centimeter difference relative to the sea level. As a consequence, a linear fitting is only valid in the situation it was calculated. Hence it is relevant to perform the stationary test, also called the Dickey-Fuller test [76]. The Dickey-Fuller test was applied to pressure data coming from three different days in static conditions. The results of the stationary test of the pressure data from the reference and wearable boards are presented in Table 2.1. The reference pressure data is 95.00% to 99.00% stationary, with $Test \leq Critical-Value(1\%)$. And, the pressure data from the wearable board was 90.00% to 95% stationary, with $Test \leq Critical-Value(5\%)$. Both results indicate that the pressure values are semi-stationary. But still, the Fig. 2.2 shows that the pressure values vary per day. Moreover, the offset difference between the pressures of the reference and wearable board is not constant.

The coherence and stationary tests show that the simplest model version is a linear model, as the Eq. (2.4). Where W_p is the pressure at the wearable board and the R_p is the pressure at the reference board, a is the weight of the

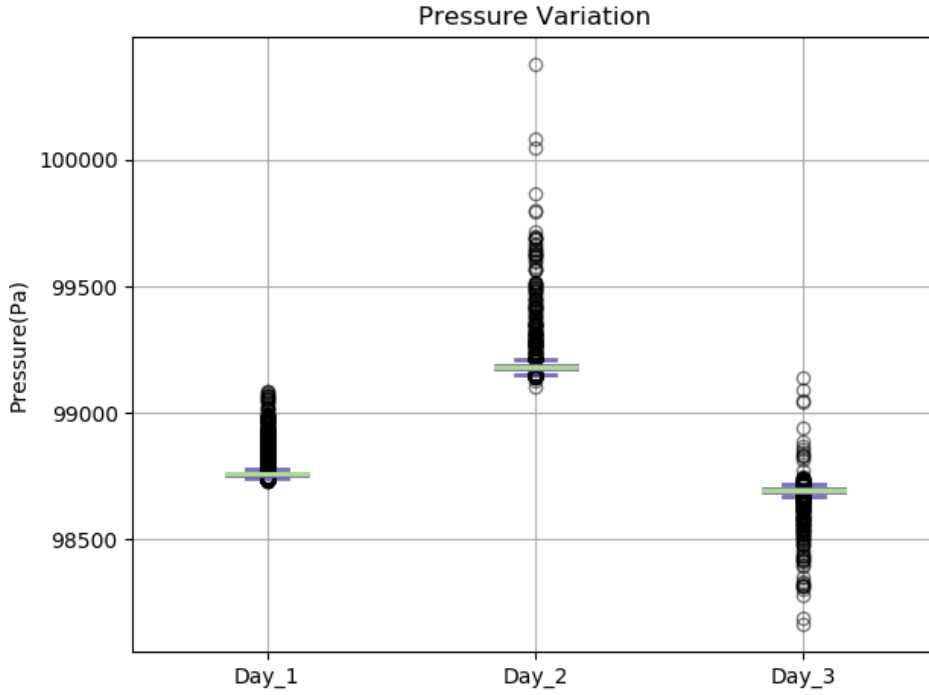


Figure 2.2.: Three Days Static Pressure Variation Capture by the Customized Hardware for 20 to 45 minutes.

similarity between pressures, and e is the DC offset difference between Wp and Rp .

$$Wp = a * Rp + e \quad (2.4)$$

Linear Model Fitting: The linear equation in Eq. (2.4) has two unknown variables, a and e . RANSAC and polynomial linear fit methods are used to determine the variable a , which is the similarity weight between the two pressure sources. The static pressure data (see Fig. 2.3) is divided into 70.00% training data and 30.00% testing data. The results from both methods are in Table 2.2. The minimum mean-squared error is 4.27 Pascal from the RANSAC method. This corresponds to a value in height of 33 centimeters, relative to the sea level pressure. The weight of similarity between the pressures (variable a) is then set to the suggested value of the RANSAC method of 0.97.

The DC offset variable e is initialized at the mean value difference between the reference and the wearable pressure values from the semi-static scenario. The DC offset error is then updated by the RFID synchronization method. The RFID synchronization method consists of recalculating e when the reference and wearable pressure devices are at the same height. The RFID method has two elements, a reader on the wrist and the tag on the reference position. Then, every time the reader captures the tag the e is adjusted accordingly. And, the RFID tag-reader combination selected for this work is the same used

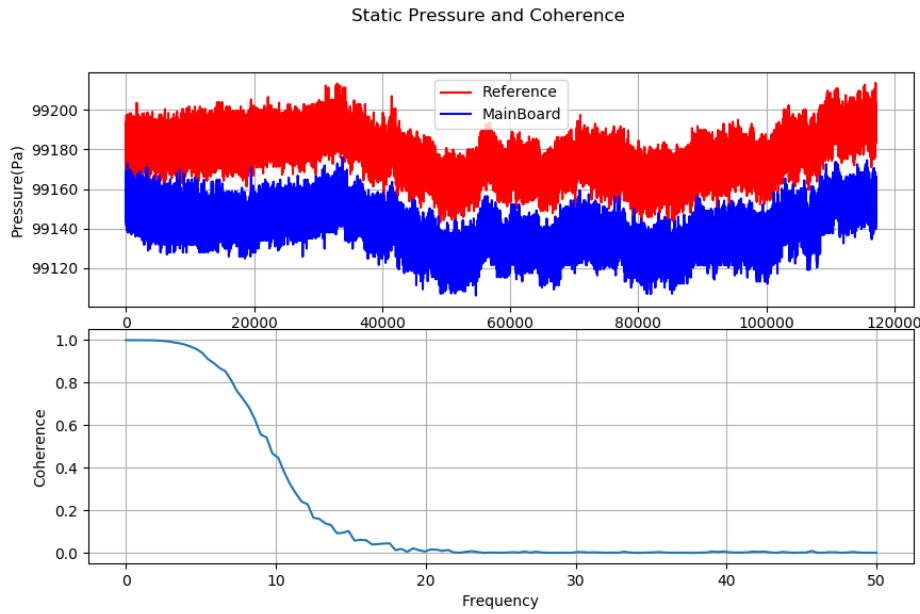


Figure 2.3.: *Top* Depicts the Pressure of the Reference Device in Red and the Pressure of the Wearable Device in Blue. *Bottom* Shows the Coherence of the Signals.

by researchers in [100, 159, 202].

2.1.5. Experiment Design

Two scenarios were tested. The first was an **order-picking scenario** in a warehouse. Order picking involves the movement of boxes/items between a compartment and a cart or picking table. The trolley or pick-up table is a suitable reference position to monitor height differences and determine in which compartment an object has been placed/taken. For this scenario, we placed the reference device (barometer + RFID tag) on a fixed table, simulating an order-picking car. The movable device (barometer + RFID reader) is set on

Table 2.1.: *Dickey-Fuller Test Results to Test Stationary Attributes of the Pressure Values in Three Different Days in Static Conditions for the Reference and Wearable Pressure-Based Devices.*

Reference Board	Wearable Board
Test = -3.44	Test = -3.17
p-value = 0.0096	p-value = 0.022
Lags = 71	Lags = 71
Observations = 118265.0	Observations = 118265.0
Critical-Value(1%) = -3.430405	Critical-Value(1%) = -3.43
Critical-Value(5%) = -2.86	Critical-Value(5%) = -2.86
Critical-Value(10%) = -2.57	Critical-Value(10%) = -2.57

Table 2.2.: Comparison of The Linear Fitting Results for RANSAC and Polynomial Linear Fit Methods of The Pressures Between the Reference and The Wearable Device.

RANSAC	Polynomial
Mean-squared-error = 4.27	Mean-squared-error = 4.47
Variance-score = 0.95	Variance-score: 0.94

the wrist of the participant to simulate the smartwatch position. In this scenario, a shelf of six compartments is used as the level categories to be tracked. Five participants were recruited to place/take objects to/from between the shelf and the table. They were asked to move a box between the reference table and each of the shelf's compartments randomly. The waiting period in each shelf's level was between 3 to 5 seconds. In total, each participant performed 10 picking actions per compartment, every time going back to the reference table and waiting there for 3 to 5 seconds. The maximum compartment height is 28 cm, and the reference table height is 85.4 cm. And, the heights of the volunteers in decreasing order were: 197,190,177,170,157 cm.

The second scenario was related to the **movement of the arm to specific body locations**. Three locations are defined as the upper, middle, and lower parts of the body. In this case, the reference device is placed on top of the pocket of the participant to simulate a common position for the smartphone. As in the first scenario, the movable device is set on the wrist of the volunteer. Three participants were asked to move their dominant arm starting from the pocket to three different positions. The levels around the body were to the head, to the chest, and to the feet. They performed the movements randomly and for 10 repetitions per location, returning to the pocket for every repetition.

The RFID synchronization method was validated in an additional experiment. Three volunteers were asked to move an object between the table and a shelf compartment of their choice. Two of the participants performed 30 repetitions of the gesture. The third performed 50 repetitions of the gesture. In all cases, the RFID captured 100% of the picking actions, which validated the RFID synchronization method.

Naive Bayes Classifier: In Fig. 2.4 is the altitude distribution (one hour per level) by each level of the shelf. The altitude distribution based on the barometric formula shows Gaussian behavior. Hence, a naive Bayes classifier is a suitable method to quantify the system's accuracy. First, the altitude is calculated using the barometric formula. The linear model is used to match the pressure from the wearable device to the reference pressure and then proceed to obtain the elevation values. Then, each picking action is divided into windows of 12 samples, 0.192 seconds in total with a sampling rate of 62.5Hz(16ms), and statistical features were calculated for each window. The statistical features were the mean, standard deviation, maximum, and minimum values of the pressure-based height measurements.

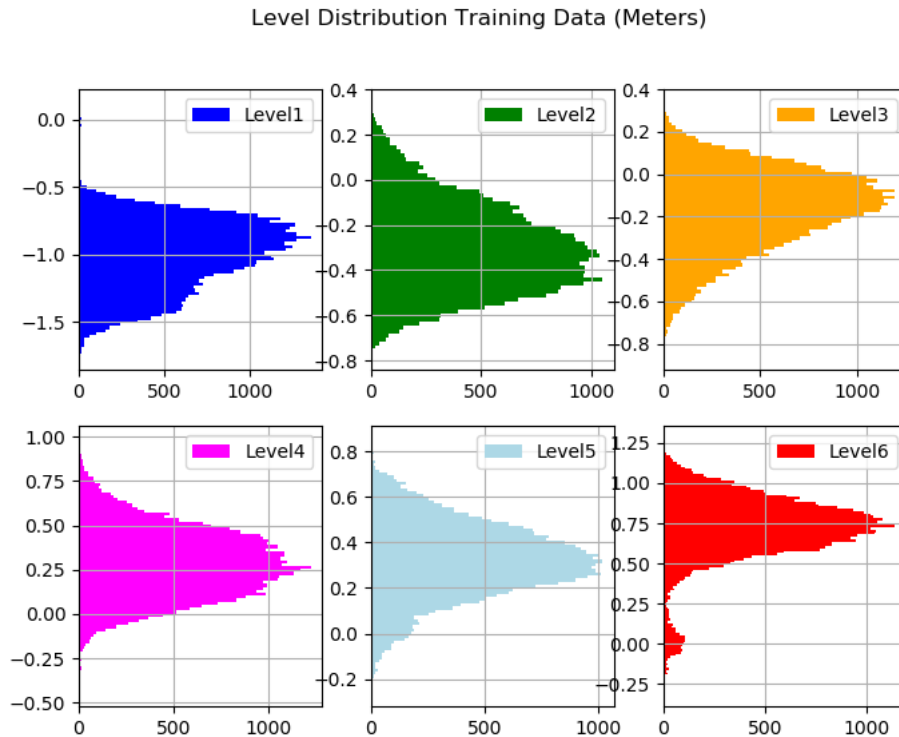


Figure 2.4.: *Altitude Values Level Distribution (Vertical axis in Meters) from the Warehouse Scenario*

2.1.6. Results

For the first scenario, the naive classifier was trained with one minute of altitude static data from each compartment. The testing data came from five participants for a user-independent evaluation scheme. The performance for the best case was 62.86% accuracy. And for the worst case, the accuracy was 56.81%. Fig. 2.5 **Left** shows the confusion matrix of the best performance for the pressure-RFID-based approach. In Fig. 2.5 **Right** shows the best results for the case of not using the RFID method for synchronization and error correction. In this case, the classifier was trained with the same one minute of altitude static data, but without applying the RFID error correction method. This means only using the differential barometric information with the linear model. Thus, the system will only reduce the impact of pressure signal drifting and sudden changes in the pressure. When the RFID method is not employed the offset error between the pressures is not updated. As a consequence, the accuracy of the system decreased to 48.61%, and the recognition performance of the shelf middle height levels was highly affected (see Fig. 2.5 **Right**).

For the second scenario, arm movement around the body, three different models were developed. One model is trained on 70% of the data coming from the participant with 170 cm in height, and tested in the remaining two participants. Another model was trained with 70% of the data from the volunteer

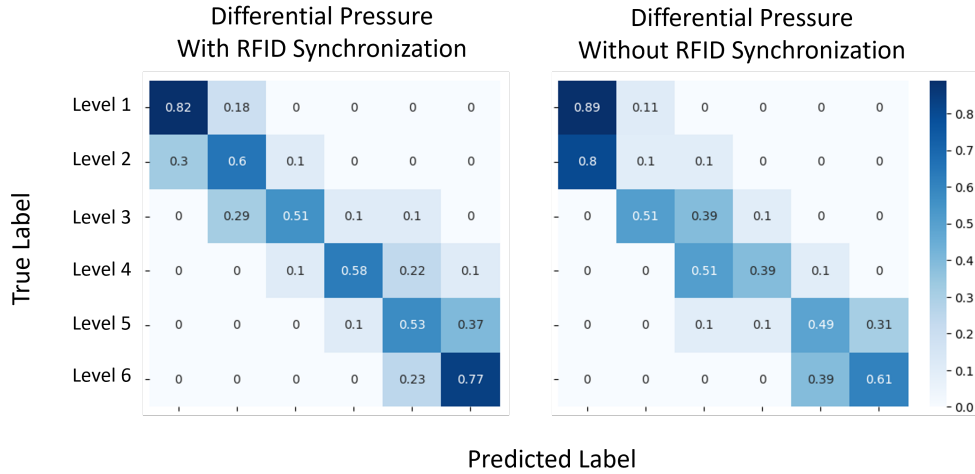


Figure 2.5.: *Left* Confusion Matrix of the Best Case from the Warehouse Scenario Using RFID-Monitor Error Method. *Right* Best Case Confusion Matrix Without using RFID Synchronization Method

with a height of 177 centimeters. And, a third model is trained on 70% of the data coming from the volunteer with 190 cm in height. The accuracy results are presented in Table 2.3. The third scenario is the one with less variation in accuracy. In Fig. 2.6, are the confusion matrices for the third scenario for the three participants. At this stage, it is almost impossible to justify these results. This is mainly due to the reduced number of participants.

User to Test	Height	User to Train	Accuracy
P1	170 cm	P1-170cm	87.28%
P2	177 cm	P1-170cm	81.82%
P3	190 cm	P1-170cm	78.81%
P1	170 cm	P2-177cm	82.03%
P2	177 cm	P2-177cm	91.48%
P3	190 cm	P2-177cm	73.05%
P1	170 cm	P3-190cm	86.39%
P2	177 cm	P3-190cm	85.23%
P3	190 cm	P3-190cm	84.25%

Table 2.3.: Comparison Between Accuracy Results of Three User-Independent Models using The Vertical Hand Position Estimation to Classify Body Positions (Upper, Middle, and Lower Body).

2.1.7. Discussion

Our work validates the idea of using differential atmospheric pressure to track the vertical position of the user’s hand. Two copies of the same hardware with a barometric sensor are deployed to obtain the reference pressure and variable pressure sources. One custom hardware was on the wrist (variable pressure)

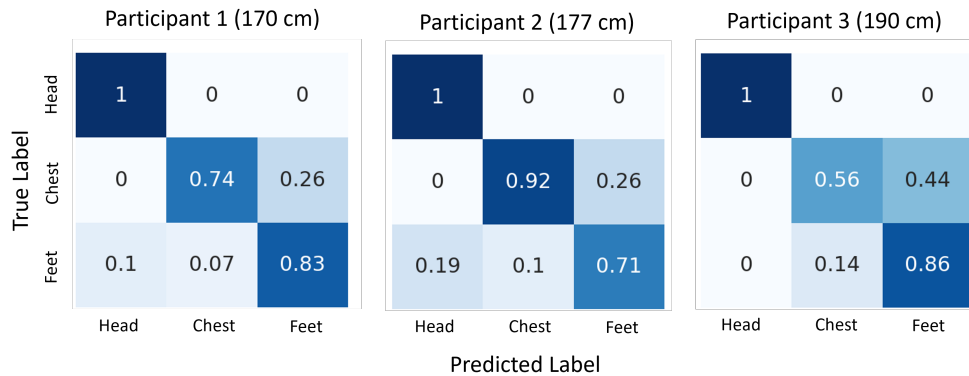


Figure 2.6.: Confusion Matrices of Three Participants from The Around the Body Position Scenario for the Naive Bayes Classifier Trained on Data From Participant Three.

and the other was static on a reference table or in the user’s pocket. The main idea is to use a linear model as the relationship between the reference pressure and the variable pressure source at the wrist. This technique mitigates the effects of pressure drift and reduces sensitivity to sudden pressure changes in the environment, such as opening/closing doors or windows. In addition, RFID synchronization is used as an error correction method for the initial offset between the two pressures. The RFID reader is placed on the wrist device and the RFID tag is placed on the reference device (on the table or in the pocket). Then, each time the RFID reader detects the tag it indicates that the two pressure sources are at the same height, and the initial offset is corrected. The system is evaluated in two scenarios: order picking experiments and arm movement around the body. In general, the results with the RFID-based approach are significantly improved ($\geq 48\%$). The idea offers a novel, simple but effective strategy for using differential pressure information. Moreover, in the case of the order-picking scenario with six levels as categories, the model is trained on one minute of pressure data in static conditions per level and tested on data from five participants in a user-independent evaluation scheme (accuracy of 60.25%). This shows the potential of the method.

Nevertheless, our system has several limitations. The size of the custom hardware is 11 cm in height. In the future, it is possible to reduce the size of the electronic design (smartwatch size). The experiment design can be extended in the number of participants to train a neural network that could improve the accuracy of the hand vertical position estimation. Also with neural network-based modeling, it is possible to learn the nonlinear influences on the pressure-height relationship, which is intrinsically nonlinear. Moreover, the system could be merged with technologies such as inertial sensing and Bluetooth low energy (BLE) to add wireless connectivity, both of which are available in recent smartwatch designs. On the other hand, the experiment results included user-dependent and user-independent evaluation schemes. In the case of the order-picking experiment, the evaluation scheme is completely

user-independent. And, for the case of the head, chest, and feet position categories, only 70% of the data from one of the participants (three in total) is used for training, which is minimal. The classifier is then evaluated with test data from the remaining two participants. This makes our evaluation fair but challenging to obtain high performance ($\geq 90\%$ accuracy) due to the small number of volunteers to learn from.

2.1.8. Conclusion

The use of relative height estimation using barometer pressure differences and RFID as an error monitoring method improved vertical position detection compared to using the barometer-height relationship without RFID, where the dependence on drift, humidity, and temperature was not taken into account. This objective was achieved by using only a simple linear model and an RFID chip to calculate the initial error and reduce the impact of drift and displacement on the estimation of the vertical position, as well as reducing the effects of sudden changes in pressure due to open windows or doors around the devices. This initial step could evolve into sensor fusion for more accurate results. An important point to note is that the prototyping hardware has a height of 11 cm, which means that the error will be in the measurement depending on the orientation of the wrist in the picking action. In addition, the linear fit has a minimum mean square error of 4.27 Pa. For the simplicity and limitations of the system, this first step achieved relatively good results.

2.2. Capacitive and Inertial Fusion-Based Glove for Real-Time on Edge Hand Gesture Recognition

2.2.1. Problem Statement

Human-robot interaction (HRI) has emerged as a significant field that focuses on optimizing the interaction between users and robots by designing interfaces that meet users' needs [78]. HRI is relevant in the smart factory because it improves efficiency and safety. Besides, it can empower workers with human-centered artificial intelligence [150]. The default and most accurate option to recognize hand gestures are camera-based solutions [142, 143, 152, 210, 211]. However, concerns regarding workers' privacy and technology protection in industrial environments are critical. In this context, alternative solutions that can provide good performance without the risk of technology leakage are highly encouraged. Non-camera-based wearables offer a convenient option for a privacy-aware robot control mechanism. The ideal wearable should be flexible and comfortable to ensure minimal disruption to the worker's schedule while maximizing efficiency. One of the challenges the wearable community faces is bridging the gap between the performance of camera-based solutions and the accuracy of a flexible, privacy-aware wearable device. Using gloves for smart garments particularly interests the wearable community [50]. Gloves are commonly used as protective equipment in various industries and offer flex-



Figure 2.7.: *Hand Gestures Dictionary for Drone Control [98]*

ibility and dexterity to generate many control patterns with the fingers and wrist. Textile sensors integrated into gloves provide additional advantages such as softness, comfort, lightness, and air permeability. Thus, a glove-based textile solution for HRI is a convenient wearable option. In this work, a subset of hand gestures is recognized with a multipositional and multimodal approach, using textile capacitive sensors and inertial information. Although the gesture dictionary is intended for drone control (see Fig. 2.7), the methodology is extensible to other applications in the fields of sign language, gaming, and robot control, among others.

2.2.2. Contributions

The main contributions of our approach can be summarized as follows:

- We present a real-time on-the-edge solution for drone control that utilizes textile-based sensing, providing flexibility, low power consumption, and cost-effectiveness, with potential applications in sign language, gaming, and robot control.
- We employ lightweight neural network models to ensure a low memory footprint, providing an embedded and sustainable solution.
- We propose a hierarchical multimodal fusion to reduce power consumption and increase robustness against the null class, where the first stage detects movements and recognizes a non-null hand gesture using an inertial-based model. Then, using a capacitive-based model, the second stage classifies the dictionary shown in Fig. 2.7.
- Experimental results demonstrate that our approach is a step towards a wearable, textile, and privacy-friendly alternative for hand gesture recognition.

2.2.3. Apparatus

Fig. 2.8A presents the prototype showing the IMU and the capacitive channels (textile electrodes) on a sports glove. The hardware block diagram is in Fig. 2.8B. In Fig. 2.7, the drone hand gesture dictionary is depicted. The dictionary comes from [98]. They presented a camera-based and real-time solution using MediaPipe. The gesture majority are dominant by finger patterns. To monitor finger movements, we proposed to use textile conductive

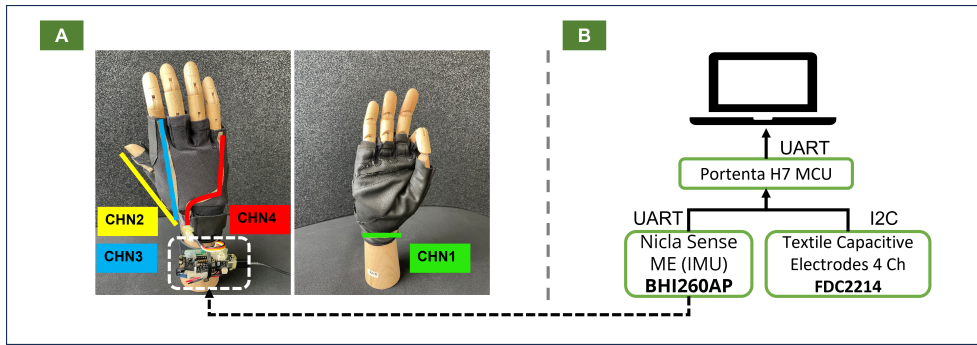


Figure 2.8.: (A) The Hardware Prototype Shows the Capacitive Channels and IMU Positions on the Sports Glove. (B) Depicts The Hardware Block Diagram with the Sensors Connections to the Main Board (Portenta H7) and PC.

thin patches as capacitive channels. Moreover, the IMU-selected placement is on the wrist. This approach reduces the number of connections, and flexibility and comfort are considered. Noticeably, our glove does not cover the entire area of the fingers, minimally affecting the user’s mobility.

The hardware has three blocks; a main board, an inertial and environmental sensing board, and a capacitive sensing board see Fig. 2.8B. The main board is a Portenta H7; the main processor is the dual-core STM32H747, including a Cortex M7 running at 480 MHz and a Cortex M4 running at 240 MHz. Portenta H7 offers 2MB flash and 8MB SDRAM and wireless data transmission options such as WiFi, Bluetooth classic, and BLE. The inertial board is a Nicla Sense with a 64 MHz ArmCortex M4 (nRF52832) and sensors such as; IMU, air pressure, humidity, temperature, and gas. The capacitive board is based on the state-of-the-art capacitance to digital converter FDC2214 with four channels. Four capacitive channels are distributed on the glove; channel one on the wrist, channel two on the thumb, channel three on the index finger, and channel four on the little finger. The capacitive channels are textile electrodes based on Shieldex Technik-tex P130+B. The dimensions of the electrodes are 0.55 mm in thickness and between 11-15 cm long. The FDC2214 offers single-end and differential sensing modes. We use single-end mode to reduce the number of capacitive patches on the glove. Furthermore, the FDC2214 is configured using an external inductor of 18 uH and a capacitor of 33 pf to operate with an average frequency of 13.7 Mhz [220]. The sampling rate for the sensors is around 50 Hz. The highlights of our prototype are in Table 2.4.

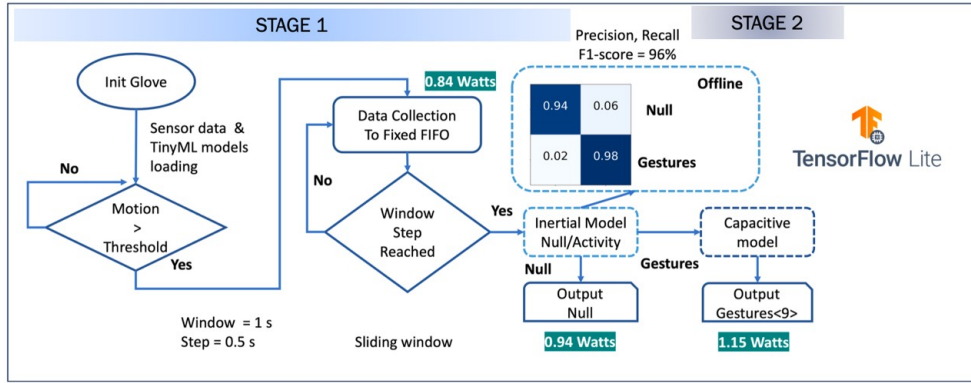
2.2.4. Signal and Data Processing

As shown in Fig. 2.9, two collaborative models were deployed for the real-time and on-the-Edge (RTE) recognition of the gestures in Fig. 2.7. A pre-normalization $((x - x_{min}) / (x_{max} - x_{min}))$ per window is applied to the inertial and capacitive signals. The window size is 2s, and the window’s step is 0.5s. The first neural network model (NN) is the inertial model with three channels as input (linear acceleration). This model is used to distinguish the null class

2.2. Capacitive and Inertial Fusion-Based Glove for Real-Time on Edge Hand Gesture Recognition

Table 2.4.: Glove Apparatus Component Highlights

Component	Benefits
Portenta H7	<ul style="list-style-type: none"> • Dual core STM32H747 • Graphics Accelerator • 2MB Flash, 8MB SDRAM • WiFi/BT Module • NXP SE050C2 Crypto
Nicla Sense ME	<ul style="list-style-type: none"> • BHI260AP IMU • BMP390 and BME680, Pressure, Humidity, Temperature, and Gas • AI self-learning sensor
FDC2214	<ul style="list-style-type: none"> • 4 Channels 28-Bit Capacitance to Digital Converter • Single or Differential mode • Proximity Detection • Liquids sensing (detergent, soap, ink) • Collision Avoidance • Rain, Fog, Ice, Snow Sensor
Textile Electrodes	<ul style="list-style-type: none"> • Shieldex Technik-tex P130+B • Knit type: Stretch-Tricot • Resistivity $\leq 2\Omega$ • Nitrile rubber protective coating • Double stretch direction • Temperature range -30 to $90^\circ C$



Real-Time and on-the-Edge Implementation for Hand Gesture Recognition

Figure 2.9.: Real-Time and on-the-Edge Implementation for Hand Gesture Recognition

from gesture detection. The null class includes activities such as; walking and standing/sitting down, among others. The output of the acceleration model served as a trigger for the second model, the capacitive model. If an activity is classified as non-null, the capacitive model is activated. The second model fused the four capacitive channels as four independent input channels. The outputs of the capacitive model are the eight classes defined in the dictionary in Fig. 2.7 plus the null class (total of nine classes). The hierarchical approach reduces the complexity of the models, leveraging the information fusion with lightweight NNs (0.10-1.23 MB) to be deployed in tiny MCUs. The intermediate tensor space (Arena) is 16.66 KB for the inertial model and 130.56 KB for the capacitive model.

The Inertial Model: The NN structure comprises three convolutional layers (filters=10, kernel=10, ReLu). For each convolutional layer, batch normalization, max-pooling ((5,1)), and dropout (0.5) are applied. Then it is followed by a flattening layer, a fully connected (FC) layer of 10, and an FC with softmax and two outputs. The training ran for 100 epochs with early stopping (patience 30 and restoring weights). The number of parameters of the inertial model is 2882; thus, it is a lightweight design and less susceptible to overfitting. The structure of the neural network for the inertial model is depicted in Fig. 2.10 **Top**.

The Capacitive Model: The NN structure comprises two convolutional layers (filters=40, kernel=10, ReLu). A normalization layer follows the first convolutional layers. For each convolutional layer, batch normalization, max-pooling ((5,1)), and dropout (0.3) are applied. Then it is followed by a flattening layer, a fully connected (FC) layer of 100, and an FC with softmax and nine outputs (gesture dictionary Fig. 2.7 plus null class). The training ran for 200 epochs with early stopping (patience 30 and restoring weights). The number of parameters of the capacitive model is 49890; thus, it is a lightweight design and less susceptible to overfitting compared to a small network such as MobileNetV2 with 3.5 Million parameters. For both models (inertial and capacitive), the NN optimizer is AdaDelta, with a learning rate of 0.9 and categorical cross-entropy as a loss function. The metric to monitor during training is accuracy. The NN models were trained using the TensorFlow/Keras 2.12.0 framework. The structure of the neural network for the capacitive model is depicted in Fig. 2.10 **Bottom**.

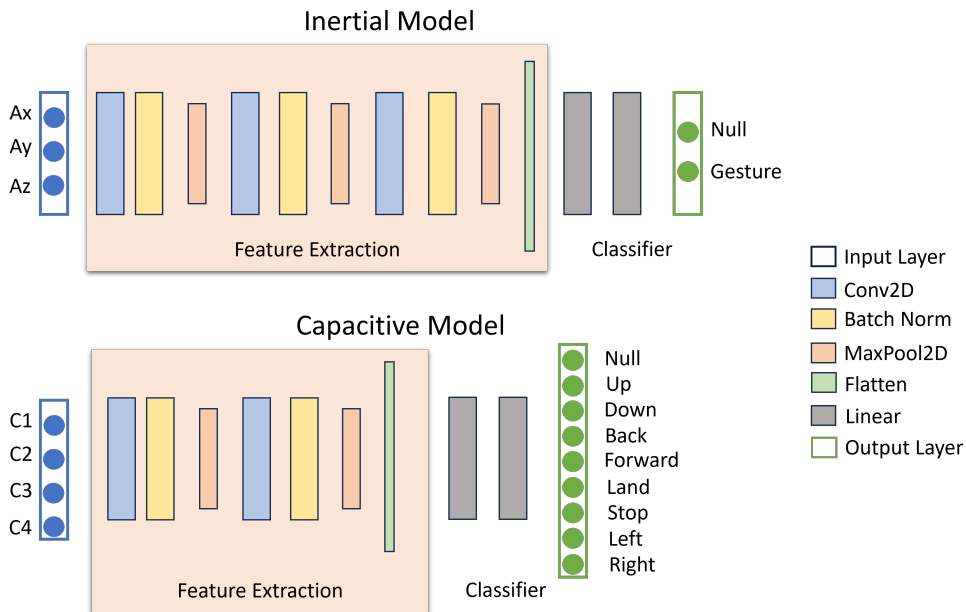


Figure 2.10.: *Lightweight Neural Network Structure for Real-Time and on-the-Edge Inference of Hand Gestures using CaptAinGlove. **Top:** Inertial Model Structure. **Bottom:** Capacitive Model Structure.*

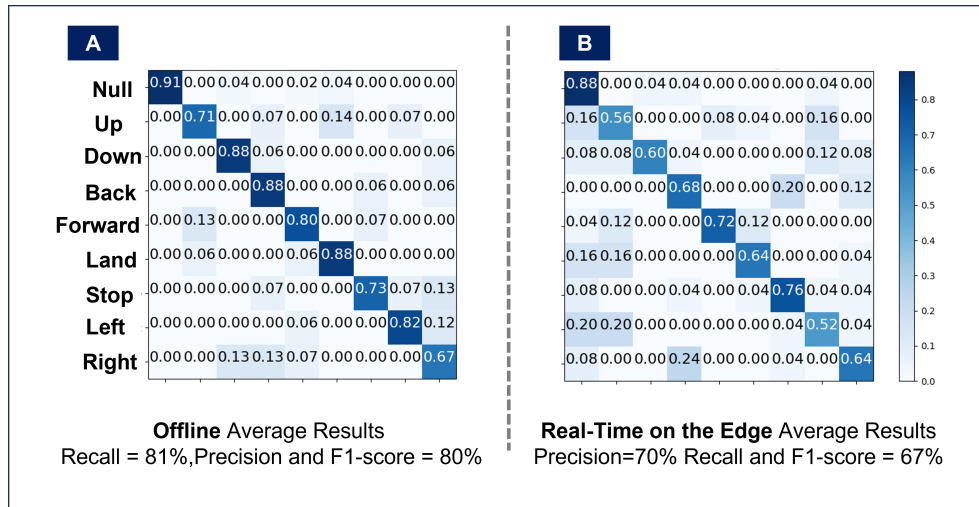


Figure 2.11.: Results of the offline Capacitive Model; Null(0), Up(1), Down(2), Back(3), Forward(4), Land(5), Stop(6), Left(7), Right(8) and F1-score=80%(A). Real-Time on the Edge Results of Capacitive Model; F1-score=67%(B).

Real-Time and on-the-Edge (RTE):

TensorFlow Lite for MCU was used to generate the embedded version of the NN models. For RTE recognition, a sliding window scheme is employed. A sliding window of 2s (100 samples) with a step size of 0.5s is used as an input data frame to the NNs. The Fig. 2.9 depicts the real-time on-the-edge procedure. The first step consists of movement detection (using acceleration), reducing power consumption by 10 %. The movement detection is based on a threshold condition ruled by $\sum_{n=0}^5 = |a_x|_n + |a_y|_n + |a_z|_n$. Then, the inertial model will run and detect null or gesture cases. In the case of $activity \neq Null$, the capacitive model will output the recognized drone control gesture in Fig. 2.7. The power consumption (PC) when only the movement detection is activated (only sensor data acquisition) is 0.84 W. Then, if a movement is detected, the inertial model is triggered, and the PC increases by 0.10 W (0.94 W). If the inertial model detects a gesture, the capacitive model runs, and the PC increases to 1.15 W. Hence, the NN model PC adds 0.31 W to the system pipeline.¹ For the real-time and on-the-Edge assessment, one volunteer performed five sessions with five repetitions of each gesture from our dictionary.

2.2.5. Results

Fig. 2.11 shows the offline results (10-fold cross-validation) for the collaborative approach; inertial model (Null vs. Activity) in Fig. 2.9 with F1-score = 96% and capacitive model (gesture dictionary) with an F1-score = 80% in Fig. 2.11A. For the training data (offline results), one volunteer (female) participated in mimicking (randomly ten sessions) the gesture dictionary in

¹USB Digital Power Meter: <https://www.az-delivery.de/en/products/charger-doktor>
DLA: January 2, 2025

Ground Truth → 00 111111100 Prediction → 00 110131100 After Smoothing → 00 111111100

Figure 2.12.: *Example of Smoothing Temporal Windows for Continuous Recognition [141]*

Fig. 2.7 while wearing the system. The sessions were recorded on different days to ensure our device was worn repeatedly. The offline evaluation scheme was 10-fold cross-validation with a leaving-one-session-out. Each session has four random tries per gesture. ²

In Fig. 2.11A and in Fig. 2.11B, the confusion matrices for offline and on-line recognition are presented. In the offline results, we can observe confusion between the gestures, Up and Land (14%) and Forward and Up (13%); both pairs mainly differentiate by how the finger’s upper parts move. The sports glove we employed does not cover the finger’s upper parts to allow flexibility/comfort for the user. There is also confusion for the case of the pairs; Stop and Right (13%), Right and Down (13%), Right and Back(13%), and Left and Right (12%). All these pairs have in common that the fist is closed, and their main difference is how the thumb and the index finger move. For specific applications such as sign language gesture recognition, the glove can be extended to cover the fingers completely to reduce confusion. For applications where the finger flexibility/freedom does not want to be reduced, we proposed as a future work that the inertial data (including orientation) could be fused with the capacitive information to add the wrist position in space (earth navigation frame). As an example of such applications, in an industrial environment, when the worker is focusing on order and picking or assembling tasks, the worker can benefit from the help of a drone/robot to ease the workflow but still needs to handle tools comfortably. These future solutions will positively impact the online results presented in Fig. 2.11B. The real-time on-the-edge results (5 sessions, re-wearing, one volunteer) in Fig. 2.11B gave an F1-score=67%. There was a reduction of 13% in the F1-score between the offline and the embedded solution. Noticeably, the RTE confusion matrix in Fig. 2.11B is based on cross-validation, shuffled, and without temporal smoothing between adjacent windows, which could improve the results in the future, as shown in Fig. 2.12.

2.2.6. Discussion

Our system combines inertial and capacitive sensing modalities to recognize hand gestures used for drone control using a sports glove. The inertial information is employed as a movement detector (using a threshold). Later, using an inertial model, the inertial information is used to recognize between null and gestures from the dictionary in Fig. 2.7. Then the inertial model triggers the gesture recognition with the capacitive information (nine classes). This is similar to the approaches applied in [23] and in [16], where the Radio Frequency Identification (RFID) signal is used as a trigger to begin gesture

²The participant signed an agreement following the policies of the university’s committee for protecting human subjects and following the Declaration of Helsinki.

detection, reducing power consumption and model complexity while improving accuracy. The set of gestures is not very intuitive. The selected dictionary comes from [98]. And it could be improved to suit the target. We aimed to have a fair comparison between a real-time camera-based solution and our proposed method.

It is important to note that our system was tested for one participant. Thus, more participants are still required to make it generalizable. The solid hardware components (MCUs and IMU) position was limited to the wrist to reduce the negative impact on the user's movements. The capacitive electrodes used are stretchable and soft. We selected a sports glove that does not cover the entire fingers to maintain the user's mobility. The hierarchical fusion of the inertial and capacitive information impacts the power consumption reduction by about 27%. The fusion method also helps reduce model complexity and parameters to obtain lightweight neural networks to be deployed in embedded devices. Our approach is a step toward a textile, flexible, embedded solution for drone control. The main idea can be extended to other hand gesture-controlled applications where comfort, power consumption, and privacy are desirable.

On the other hand, our design has several **limitations** and possibilities for improvement.

- The hardware prototype can be reduced in size by doing a professional encapsulated electronic design.
- The latency of the recognition in real-time is not optimized. The main board offers two MCUs that can work in parallel, but in our design, we allocate the entire flash (2MB) to the Cortex M7 core of the STM32H747, and the code is running only on the M7 at 480 MHz. The latency could be improved by using the M7 to run the neural network models and the Cortex M4 (at 240 MHz) for the sensor acquisition data. Special attention must be given to memory access to avoid collisions and bottlenecks.
- Additionally, our work transmits the recognition results to the PC using a universal asynchronous receiver/transmitter (UART) to focus on proving the idea. The main board can be configured to send the recognition results by wireless communication (Bluetooth or Wifi) to improve comfort and make the system ubiquitous. The Bluetooth/Wifi will require an external antenna and memory allocation for the wireless communication managing functions.
- For the case of the RTE results, the confusion matrix is calculated based on shuffled cross-validation over fine-granular windows, which does not consider continuous sequences of windows of a single gesture where the majority of windows are true positive with sparse false detection. Our result could be improved by merging temporal windows from simple gap filling and event-based smoothing to selective merging based on CNN [141]. Although temporal window smoothing techniques have been demonstrated in offline evaluations where the computation is performed

on the PC, edge adaptation with resource-constrained embedded hardware is a task we will investigate in future work.

- The power consumption reported in this work includes the complete pipeline. The pipeline comprehends sensor powering and data acquisition up to the RTE. For the worst-case power consumption, only 27% (0.31 Watts) is required for the real-time inference. Hence, the power consumption could be further reduced with techniques, such as lowering the data sampling rate and setting the sensors in sleep mode.
- In addition, the NNs can be pruned and trained with aware quantization to reduce size with minimal impact on the performance [175].

2.2.7. Conclusion

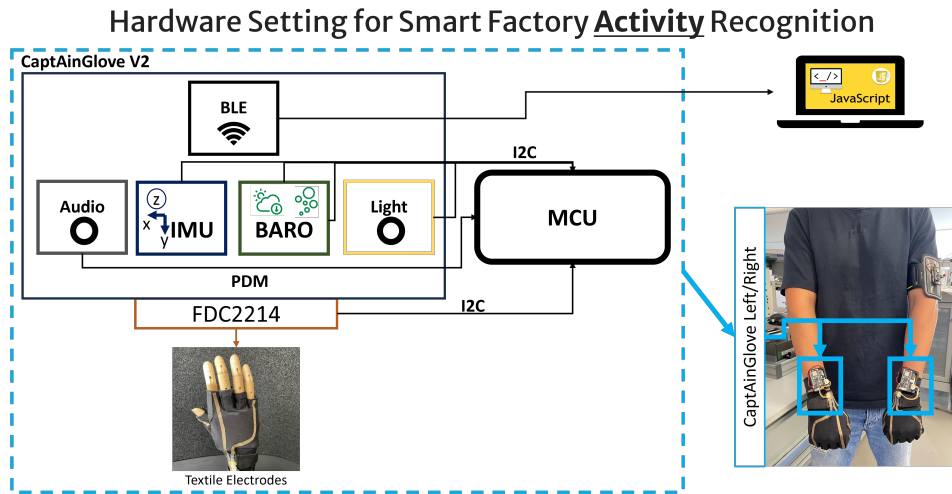


Figure 2.13.: CaptAinGlove V2 Hardware Description with Volunteer Wearing the Prototype (Left and Right Hand)

In this section, we have presented a glove-based design that provides a minimally obtrusive, low-power, privacy-friendly, flexible, and scalable solution for hand gesture recognition. By combining textile capacitive electrodes and inertial sensors, our system achieves real-time on-the-edge recognition of hand gestures for drone control.

A key contribution of our approach lies in the hierarchical fusion of inertial and capacitive information, which significantly reduces power requirements and enables the deployment of tiny memory models suitable for on-the-edge devices. This fusion technique improves the system's efficiency and performance, making it well-suited for practical applications.

Beyond drone control, our glove-based design holds great potential for various control-related applications, such as game control and assisting robot control in industrial settings. Moreover, its privacy-friendly nature and wear-

able form factor open up possibilities for broader adoption in human-robot interaction scenarios, sign language recognition, and gaming interfaces.

In the case of specific applications such as sign language gesture recognition, the glove can be extended to cover the fingers completely to reduce confusion. For applications where the finger flexibility/freedom does not want to be reduced, we proposed as a future work that the inertial data (including orientation) could be fused with the capacitive information to add the wrist position in space (earth navigation frame).

The extension of this work has been published in TSAK [14], where we used a pair(Left/Right Hand) of capacitive textile-based gloves fused with inertial data to recognize Smart Factory worker's activities, Fig. 2.13 shows an overview of the extended hardware design.

Chapter 3

Facial Muscle and Head Movements Estimation with Mechanomyography and Inertial Fusion

Contents

3.1. Facial Muscle Activity Recognition with Reconfigurable Differential Stethoscope Microphones	40
3.1.1. Problem Statement	40
3.1.2. Contributions	41
3.1.3. Approach	43
3.1.4. Apparatus	45
3.1.5. Frequency Response Analysis of The Stethoscope Microphones	45
3.1.6. Experiment Design	52
3.1.7. Signal and Data Processing	53
3.1.8. Results and Discussion	55
3.1.9. Conclusion	57
3.2. InMyFace: Inertial and Mechanomyography-Based Sensor Fusion for Wearable Facial Activity Recognition	58
3.2.1. Problem Statement	58
3.2.2. Contributions	59
3.2.3. Approach	61
3.2.4. Apparatus	62
3.2.5. Multimodal Sensor Fusion	65
3.2.6. Experiment Design	71
3.2.7. Results	72
3.2.8. Discussion	75
3.2.9. Conclusion	76

3.3. MeciFace: Mechanomyography and Inertial Fusion-based Glasses for Edge Real-Time Recognition of Facial and Eating Activities	77
3.3.1. Problem Statement	77
3.3.2. Contributions	78
3.3.3. Apparatus	79
3.3.4. Multimodal Sensor Fusion	80
3.3.5. Experiment Design	81
3.3.6. Real Time and On-The-Edge Recognition	82
3.3.7. Results and Discussion	83
3.3.8. Conclusion	86

The author of this thesis has published the content, figures, and tables included in this chapter in the following publications:

Bello H, Zhou B, Lukowicz P. Facial Muscle Activity Recognition with Reconfigurable Differential Stethoscope-Microphones. *Sensors (Basel)*. 2020 Aug 30;20(17):4904. doi: 10.3390/s20174904. PMID: 32872633; PMCID: PMC7506891. **Journal**

Bello, H., Marin, L. A. S., Suh, S., Zhou, B., & Lukowicz, P. (2023). InMy-Face: Inertial and mechanomyography-based sensor fusion for wearable facial activity recognition. *Information Fusion*, 101886. **Journal Impact Factor 18.6**

Bello, Hymalai, et al. "FaceEat: Facial and Eating Activities Recognition with Inertial and Mechanomyography Fusion using a Glasses-Based Design for Real-Time and on-the-Edge Inference." *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*. 2023.

Bello, H., Suh, S., Zhou, B., & Lukowicz, P. (2024). MeciFace: Mechanomyography and Inertial Fusion-Based Glasses for Edge Real-Time Recognition of Facial and Eating Activities. In: Bravo, J., Nugent, C., Cleland, I. (eds) *Proceedings of the International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI 2024)*. UCAmI 2024. *Lecture Notes in Networks and Systems*, vol 1212. Springer, Cham. https://doi.org/10.1007/978-3-031-77571-0_38 **Best Paper Award**

3.1. Facial Muscle Activity Recognition with Reconfigurable Differential Stethoscope Microphones

3.1.1. Problem Statement

The face plays a crucial role in many critical human actions and interactions. Through facial expressions, we show our feelings and communicate them to others. Our faces show when we are tired, stressed, engrossed in a task, or lost in thoughts. Eating, drinking, speaking, and breathing, the most elementary

actions of our lives, involve facial muscles. The same is valid for health-related activities such as sneezing, coughing, snoring, or various habits such as smoking. In the literature, face analysis is mainly pursued in computer vision, with relevant performance, but comparatively little in wearable sensing [11, 99].

This is because, for a long time, placing sensors on the user’s face was considered obtrusive to be practicable. Specially, for wearable systems that are meant to be widespread and for everyday use, rather than constrained lab settings. However, recently, more and more intelligent ”head-mounted” devices such as smart headphones, smart glasses, or smart hats have become available and have gained user acceptance. Such devices are an attractive platform for sensing face activity. Nonetheless, while facial activity affects nearly the entire face area, such devices only allow for placing sensors at particular locations (e.g., in the smart glasses frame). Consequently, sensing modalities are needed, which can infer overall facial activity from a few predefined locations.

In this section, we argue that differential body sound is a useful candidate modality. Thus, any time our facial muscles perform an action sound is generated. While the sound by itself may be challenging to interpret, differential analysis can pinpoint the sound source, which is correlated to the muscles that have created it[13, 40]. Patterns of differential sound correspond to the activation pattern of the different facial muscles (21 mimetic and masticatory muscles [126]). Furthermore, differential analysis helps mitigate noise.

Differential Pair Topology	Horizontal	Temple-Cheeks	Eyebrow-Cheeks
DMA pairs	Mic1-Mic2 Mic3-Mic4 Mic5-Mic6 ----	Mic2-Mic6 Mic1-Mic5 Mic2-Mic5 Mic1-Mic6	Mic4-Mic6 Mic3-Mic5 Mic3-Mic6 Mic4-Mic5

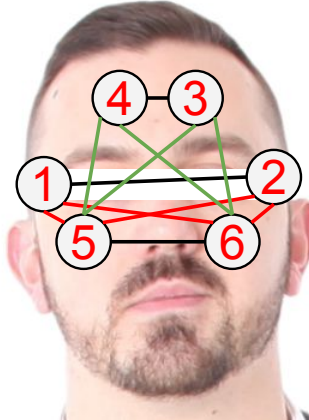


Figure 3.1.: Pairs of Reconfigurable First Order Differential Stethoscope-Microphone Arrays To Detect Facial Expressions

3.1.2. Contributions

This work explores in detail the potential of body sound to unobtrusively collect information about users’ facial activity and makes the following specific contributions:

1. We put forward the idea of using differential sound analysis as an unobtrusive way of acquiring information about facial muscle activity patterns

and the associated facial expressions and actions.

2. We present the design and implementation of a reconfigurable signal acquisition system based on that idea. It consists of six stethoscopes at positions compatible with a smart glasses frame (see Figure 3.1).
3. We present an in-depth analysis of the system's characteristics and the signals for various facial actions.
4. We describe the design and implementation of the entire processing pipeline needed to go from signal pre-processing to recognizing complex facial actions, including a study of the significance of different features, derived from combinations of six stethoscopes(at a set of locations inspired by a typical glasses frame). And the selection of best-suited ML methods.
5. We have conducted a systematic evaluation with eight users mimicking a set of 10 common facial expressions and actions (plus the null class of neutral face), as shown in Figure 3.2. Each user has recorded three sessions of 10 repetitions of each action for a total of 2640 events. Using a leave-session-out evaluation scheme across all users, we achieve an F1-score equal to 54%(9% chance-level) for those ten classes plus the non-interest class defined as "Neutral-face." In the user-dependent case, we achieved an F1-score between 60% and 89%(9% chance level), reflecting that not all users were equally good at mimicking specific actions.

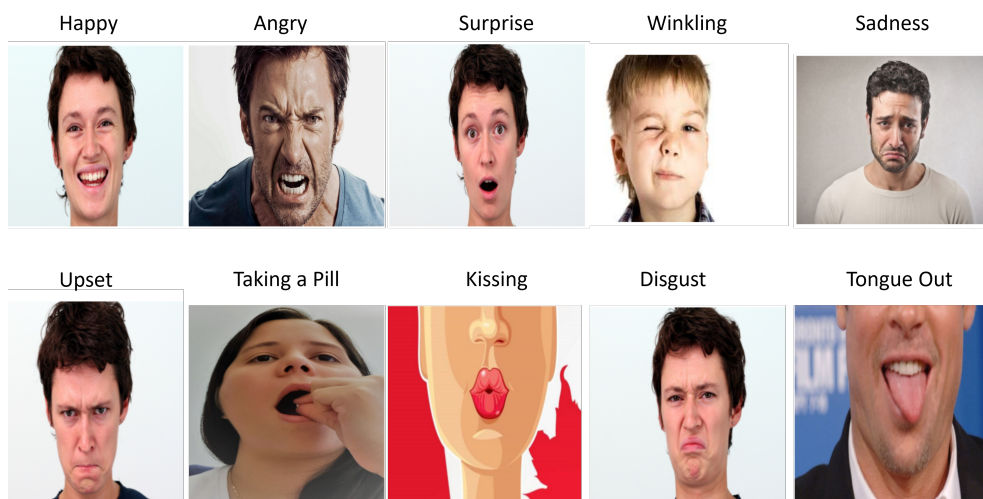


Figure 3.2.: Facial Expressions/Gestures Mimicking Set; Happiness, Upset, Sadness, Surprise and Disgust [145], Angry [84] and Gestures Blinking[52], Tongue Out[1], Kissing[34] and Taking a Pill.

3.1.3. Approach

Our approach includes wearable differential stethoscope microphones distributed around the face as depicted in Fig. 3.1. And the aim is to recognize the set of facial movements in Fig. 3.2. The idea is split into; microphone arrays, stethoscopes, and acoustic mechanomyography (A-MMG).

Microphone Arrays

Microphone arrays, commonly referred to as beamforming microphones, [85] are primarily designed to modify and control the directivity of the gain factor and noise robustness. Their functionality depends on the incident angle of the source [131]. Thus, the source's position will induce changes in how the device amplifies or reduces the gain of the captured signal. This brings the possibility of using them for position estimation of the sound sources [13, 40]. Typically, the structural design of microphone arrays can be categorized as Additive Microphone Arrays (AMAs) and Differential Microphone Arrays (DMAs)[31, 194]. There are also sub-categories based on the relationship shared by the intensity of the sound and the angle of incidence. Further specifications include the microphone arrays' configuration variations in space, denoted as first, second[33], third degrees, or higher. The structural design also varies from planar, circular[30, 40, 77], spherical[157] to hybrid structures[3, 163]. With DMA, it is possible to tune the gain in a range of frequencies or even generate a configurable resonance frequency peak. An advantage over a traditional amplifier is that there is no amplification of noise outside the selected frequencies. In the application, we use the DMA principle to focus our system on relevant sound sources inside the body (facial muscles).

Overall the choice of Differential Microphone Array (DMA) configuration is motivated by two considerations. First, DMA and Additive Microphone Array (AMA) are the most straightforward configurations of microphone arrays. Secondly, DMA allows noise resilience to common environmental sound interference (subtraction). With AMA, this would increase (addition), and we would need an additional technique for environmental sound removal. Eleven pairs and three categories are considered, as shown in Fig. 3.5. DMA-Horizontal, including $Temple_{left}-Temple_{right}$, $Eyebrow_{left}-Eyebrow_{right}$ and $Cheek_{left}-Cheek_{right}$, and the self-explanatory temples-cheeks and eyebrows-cheeks cases.

Stethoscopes

The combination of a stethoscope's head and a microphone is an established approach. In many cases the head is constructed using 3D printed stethoscope [154] together with an Electret microphone [6, 81, 187], mechanical microphones[103, 122, 148] or Piezo-Electrical Film[153]. Applications include automatic analysis of the cardiac, lung, and even fetal-heart rate signals[38]. Various improvements to the design have been proposed including frequency selection, noise filtering, wireless transmission, and real-time feedback [81, 88].

To exploit the advantages of the stethoscope, it is critical to tune its frequency response. Before using the stethoscope for our experiments, we must understand how the stethoscope's structural characteristics influence its acoustic properties. In simple terms, the stethoscope is an imperfect transducer of sound for most frequencies other than the resonance frequency. So, a frequency response analysis is necessary to understand our electronic stethoscope design. Changing the material of the head of the stethoscope [125] and adding/removing tubes will impact the resonance frequency [53]. Even a non-air-tight system impacts the frequency response, which is challenging to handle in a prototype. Likewise, the noise rejection depends on the construction. Thus, the stethoscope is a challenging device to design and use. However, for our purpose, it significantly improves the signal-to-noise ratio(SNR) if adequately tuned.

In [53], an evaluation of stethoscopes used by nurses, a Littmann[®] electronic stethoscope, and custom design was presented. The authors demonstrated that both types (electronic and passive) are resonant devices based on the experimental calculation of the frequency response using the step response analysis. A summary of this method: (1) generates a fast change in pressure on the stethoscope's head, (2) measures the input pressure and the output signal to obtain the impulse response (3) takes the derivative and (4) applies the Fourier transform. In [154] a validation of a 3D-printed stethoscope system was made. This time using a method called the "phantom method", which is based on a latex balloon filled with water used to simulate the skin. The balloon was stressed doing a sweep in vibration frequency to generate the response. Hence, there is a particular interest in frequency analysis and construction design as crucial parameters.

Acoustic Mechanomyography (A-MMG)

Many activities performed by our body are intrinsically related to muscular contractions; therefore, activity recognition based on those contractions, myography, is an active research topic [59, 68, 82]. In this area, our specific interest is in the subset of mechanomyography, which involves measuring the force contraction using low-frequency sounds/vibrations (2-200Hz) with a signal power below 50Hz [203]. We proposed to use this method to capture the facial muscle (and to a degree tissue) movements for a specific group of gestures/facial expressions. To the best of our knowledge, there was no known research on A-MMG for the facial muscles by the time of publication. In the literature, the authors in [209], investigate the replacement of electromyogram with sound myography to measure muscle fatigue by using different microphones on the middle forearm during lifting activities, finding that the Electret condenser microphone with a sampling rate of 44,1KHz was the most stable. In [203] sound was combined with the IMU(Inertial Measurement Unit), to monitor the muscle movement of patients under rehabilitation. The inspiration came from the high variability of the features of a person's actions during a typical day, in particular patients under-recovery from a neurological injury or an accident. According to [205], machanomyography (MMG) also can be com-

combined with CNN (Convolutional Neural Network) for feature extraction and SVM(Support Vector Machine) for regression to estimate the angle of the knee.

In summary, our evaluation aims to explore our hardware’s ability to distinguish different facial expressions and actions. It is motivated by the above understanding of the importance of facial expressions to assess users’ emotions and mental states. However, we understand that there is a big difference between users mimicking expressions and experiencing emotions. Thus, recognizing mimicked expressions is an important first step in identifying emotions, but the two are not the same.

3.1.4. Apparatus

Fig. 3.3 **Top Left**, depicts the wearable prototype. It consists of six stethoscope microphones placed inside a construction helmet, four of them fixed into an elastic band (numbers 1-4), to fix them around the temples and the eyebrows of the subject. The other two (numbers 5-6) are attached to the cheeks using construction goggles. These particular positions were selected to match a typical glasses frame. Fig. 3.3**Top Right**, shows the stethoscope’s head covered with a leather-like textile to reduce outside noise further. The 3D cone in Fig. 3.3 **Bottom Left** connects the Electret microphone to the stethoscope’s head. This provides an air-tight design. Fig. 3.3 **Bottom Right**, depicts the stethoscope microphones distribution around the face of a volunteer.

The electronic components were 6 Electret Microphone boards attached to low power, low-cost pre-amplifier (MAX446) with adjustable gain from the company Adafruit. The microphones have built-in amplification circuits and are easy to program. The development board is the ESP32Huzzah development board from Adafruit. The board has two cores running at 240 Mhz and 2 analog to analog-to-digital converters (ADC) with 12 bits resolution, a signal-to-quantization noise ratio (SQNR) $\approx 72\text{dB}$, and a DC-Bias of $V_{CC}/2$ at $V_{in}=3.3\text{V}$ with a precision of 0.805 mV. In addition, it has Bluetooth low energy (BLE), Bluetooth serial, and Wifi for easy wireless communication.

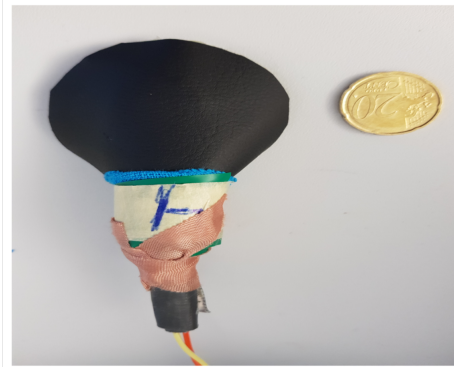
The sampling rate of each microphone was set to 200Hz. Although the hardware could handle up to 3 kHz, a lower sampling rate has several advantages. First, noise increases with a high sampling rate and we would require more complicated noise rejection techniques. Second, the signal of interest is localized in the low-frequency (lower than 200 Hz) [203]. The data transmission protocol to the PC was by Universal Asynchronous Receiver-Transmitter (UART) at 1 Mhz for robust data collection. The data was collected using a user interface developed in Python 3.6. Fig. 3.4, depicts the apparatus block diagram .

3.1.5. Frequency Response Analysis of The Stethoscope Microphones

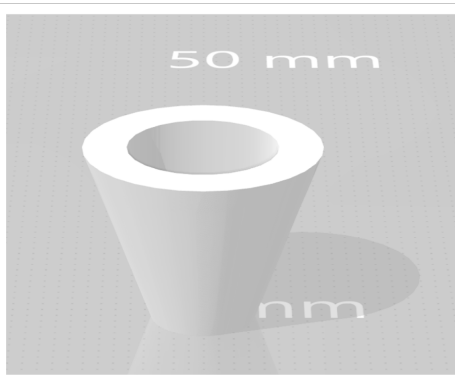
In this section, we discuss the Differential Microphones Array (DMA) design and calibration procedure to target acoustic mechanomyography(A-MMG), considering the frequency response of single and differential microphones.



Inside View of the Helmet



Stethoscope Microphones



3D Cone Microphone Connector



Volunteer Wearing the System

Figure 3.3.: *Inside View of the Helmet with Four Fixed Stethoscope Microphones in Elastic Band (Numbers 1-4) and Two Loose (Numbers 5-6) **Top Left**. Stethoscope Microphones' Head with Leather Cover and Size Comparison **Top Right**. 3D Cone Connector Between Electret Microphone and Nurse Stethoscope Head **Bottom Left**. Stethoscope Microphones Distribution on Volunteer's Face **Bottom Right**.*

For this study, the DMA configuration called first-order end-fire dipole was used. First-order means that only two elements are subtracted from each other, as shown in Fig. 3.5 **Left**. The subtraction works as a filter for environmental noise. End-fire connotation implies that the array will reject sound signals from sources in the ± 90 degrees and enhance sound signals coming from sources at 0 and 180 degrees. This characteristic depends not only on the spatial distribution of the array but also on the distances between each element and the geometry of the component itself. The sound wave can be simplified from a spherical wave to a plane wave when the source is in the far field, occurring when $r \geq 2 \cdot W^2/\lambda$ (Fraunhofer distance)[131, 167], where W is the largest dimension in the aperture (stethoscope-microphone head), λ is the wavelength and r is the distance from the opening to the source. Usually, the far-field is considered to start at a distance of 2 wavelengths away from the sound source.

Moreover, the dipole term describes the form of the polar graph of intensity

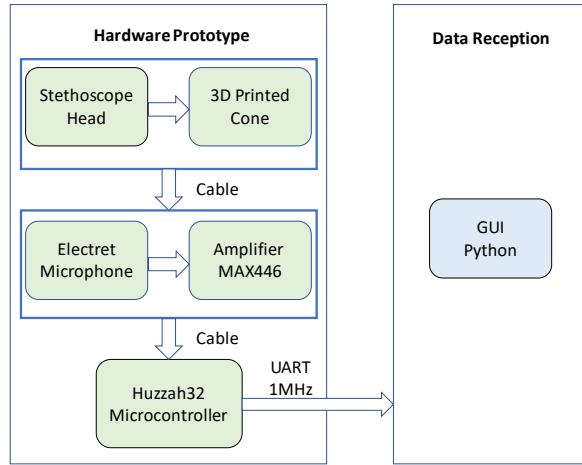


Figure 3.4.: Apparatus Block Diagram from the Stethoscope Microphones and the Hardware Prototype to the Data Reception by a Python-Based GUI.

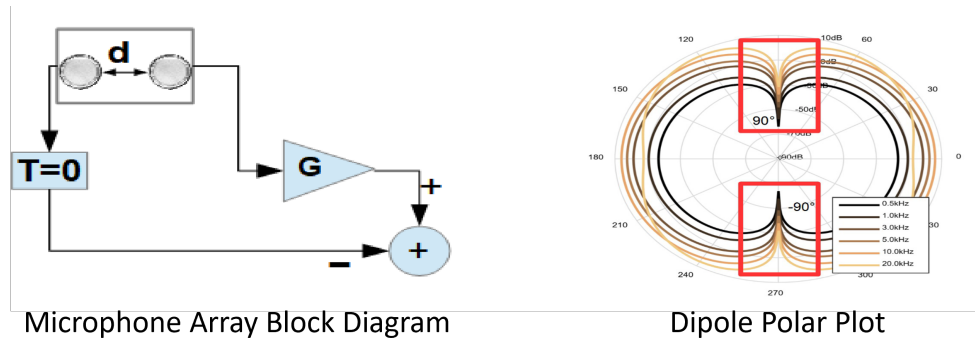


Figure 3.5.: *Left* First Order Differential Microphone Array Block Diagram. With G as a gain factor, d space between microphones, and T time-delay between microphones. *Right* Polar Plot of first-order Differential Microphone Array (Frequency vs Angle of Arrival). With an Inter-microphone space of 16mm, distance from the source of 50 cm, and frequencies 0.5,1,3,5,10 and 20KHz and sound rejection at ± 90 degrees (red boxes). Diagram by STMicroelectronic[194]

vs. angle of arrival for different frequencies, as depicted in Fig. 3.5 **Right**. The diagram in that figure assumes an inter-element space of 16mm, a distance from the source of 50 cm, and frequencies of 0.5,1,3,5,10 and 20KHz. In our case, the wave simplification is not valid, considering our sound sources as the face muscles' surface.

The End-fire design is shown in Fig. 3.5 **Left**. Where G is the gain factor difference of the microphones, d is the inter-spacing, and T is the time delay between the two microphones. We will assume $T = 0$, but this value can also be configured according to the formula $T = D/V$, where $V = 343 \text{ m/s}$ (sound

velocity). This adjustment will change the polar-plot (directivity pattern) in Fig. 3.5 **Right** and the dipole will change to cardioid (heart shape) representation [194]. This directivity pattern depends on the number of elements inside the array, the inter-spacing, the length of the aperture (Area of sensing), and their differences in the frequency response [85, 131, 194].

The frequency response and gain of the stethoscope microphone are highly sensitive to the mechanical design parameters, which are not precisely identical across the individual devices. Furthermore, in the DMA configuration, the spatial separation of the microphone is an additional influence. Thus, before our sensor configuration can be used for facial activity recognition, a calibration step is needed. This includes gain differences for each of the individual microphones and the first-order differential case. The calibration procedure is divided into; Single Microphones Discrete Frequency Response and Differential Microphones Discrete Frequency Response.

Single Microphones Discrete Frequency Response

Each stethoscope-microphone was exposed individually to a range of frequencies in the low, middle, and high spectra. In the low spectrum, the frequencies were 21, 41, 61, 81, and 101 Hz. In the middle spectrum, the frequencies were 501, 751, 1001, and 1251 Hz. And in the high spectrum, the frequencies were 1501, 1751, 2001, 2251, and 2501 Hz. The sound signal was a square wave and the duration was of 20 seconds. The signal was generated by an Android application in a Samsung Galaxy S8. The phone was placed on the top of the stethoscope head with 3 centimeters separation and the audio signal was at the highest available volume, as shown in Fig. 3.6 The experiment was conducted in a quiet room with only a single person present.

Fig. 3.7 shows the frequency response of the six microphones. The peak-to-peak value of the signal (ADC value) is not flat for the entire frequency range. It is possible to distinguish a low-frequency spectrum from 21-751 Hz, a middle-frequency spectrum from 1001-1751 Hz, and a high-frequency spectrum from 2001-2501 Hz.

If we define, $A = V \cdot G$. Where G is the gain of the microphone, and V is the response of the microphone to a sound input at $G = 1$. $A = V \cdot G$ depends on the frequency area and the stethoscope-microphone used. This result was to be expected due to a prototype design developed by hand.

In [203], the relevant frequency range is within the low-frequency range. The corresponding values of the minimum, maximum, and mean for each microphone are shown in Table 3.1. With the values in Table 3.1 and assuming ideal conditions is possible to estimate the differential gain minimum error. In the case of microphone five and microphone six, it will be 4.03. Hence, the minimum gain discrepancy between the stethoscope microphones is a factor of 4 in the low-frequency response range. From the results in Table 3.1 the matching factor between the microphones is obtained. The matching factor is $MF = \frac{A_y}{A_x}$. Where A_x is the ADC value of microphone X and A_y is the ADC value of microphone Y.

Next, we apply the 1st order DMA(Differential microphone array) with a

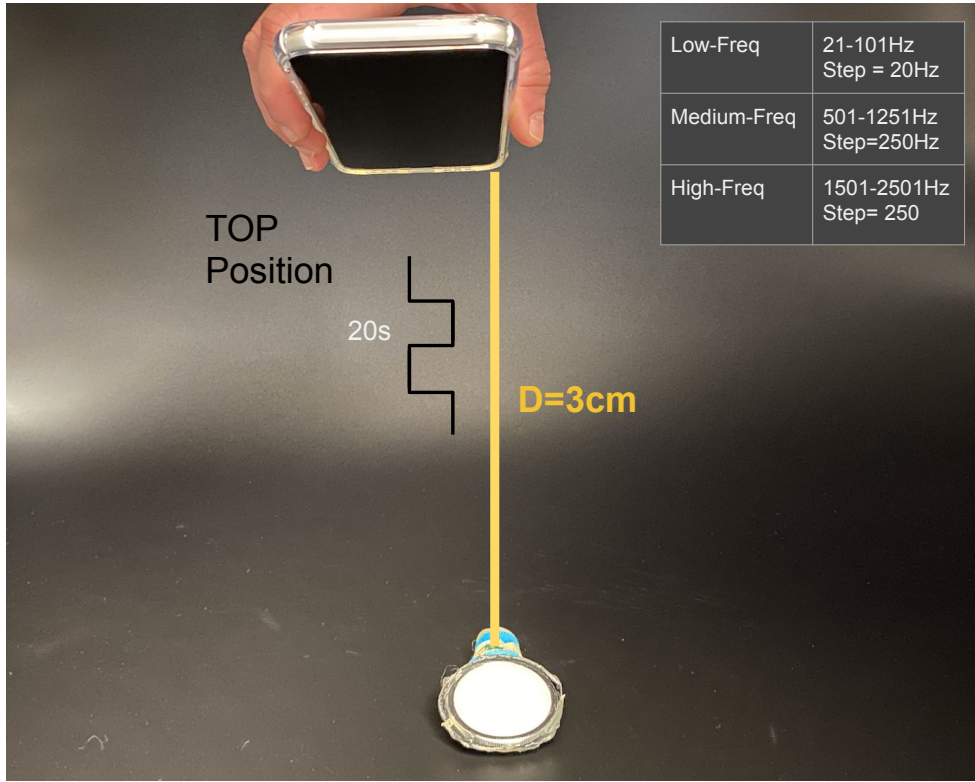


Figure 3.6.: Single Microphones Discrete Frequency Response. Phone as Sound Source and Stethoscope's Head at 3 Centimeters Separation with the Frequency Range of The Square Wave with 20 Seconds Duration at each Frequency.

Table 3.1.: $A = V * G/1000$ Values of Individual Microphones in the Low-Frequency Response Range.

M1	M2	M3	M4	M5	M6
min=2.372	min=1.486	min=1.217	min=1.108	min=1.071	min=1.015
max=3.219	max=2.254	max=2.122	max=2.737	max=4.095	max=1.811
mean=2.787	mean=1.953	mean=1.701	mean=2.103	mean=2.656	mean=1.451

matching factor $MF = \frac{A_y}{A_x}$ before the subtraction and test the influence of the gain. The matching factor MF is used in Eq. (3.1) to map the output of microphone X to the output of microphone Y, resulting in Equation (3.3).

- DMA with matching factor is cal_{dma} :

$$cal_{dma} = \frac{A_y}{A_x} \cdot M_x - M_y, \text{ where } A = V * G \text{ and M is the ADC value:} \quad (3.1)$$

- By substitution of $A = V * G$ in cal_{dma} :

$$cal_{dma} = \frac{V_y * G_y}{V_x * G_x} \cdot M_x - M_y \quad (3.2)$$

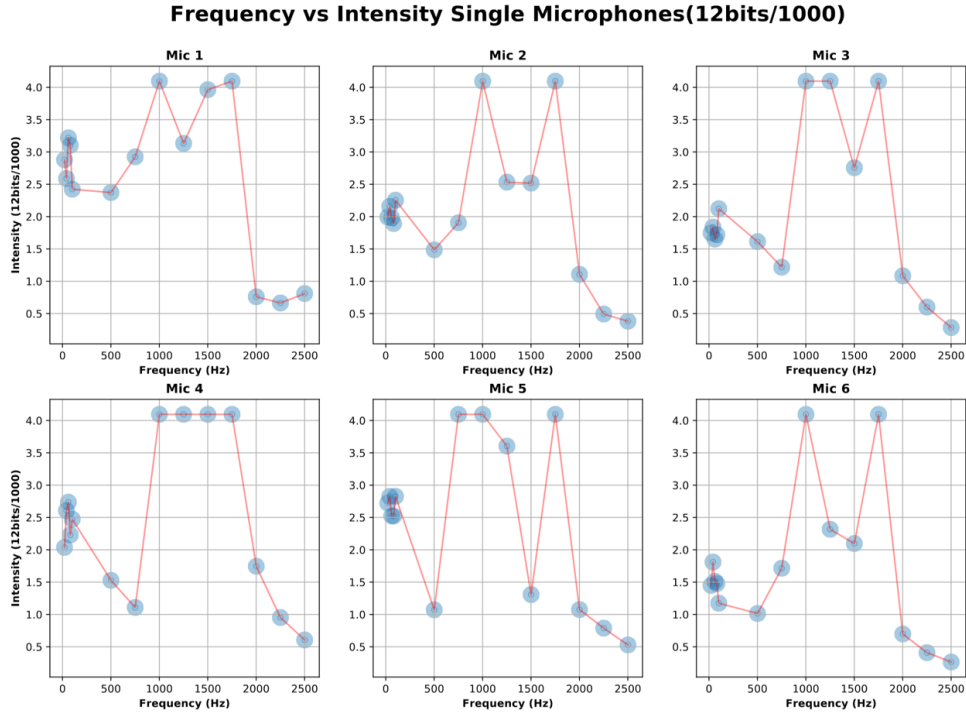


Figure 3.7.: Frequency Response of Individual Microphones. For the Low-Frequency Spectrum (21-751 Hz), Middle-Frequency Spectrum (1001-1751 Hz) and High-Frequency Spectrum (2001-2501).

- Assuming the same sound source position and geometry, only gain discrepancies.

$$V_y = V_x$$

- By substitution of $V_y = V_x$ in cal_{dma} .

$$cal_{dma} = \frac{G_y}{G_x} \cdot M_x - M_y \quad (3.3)$$

Differential Microphones Discrete Frequency Response

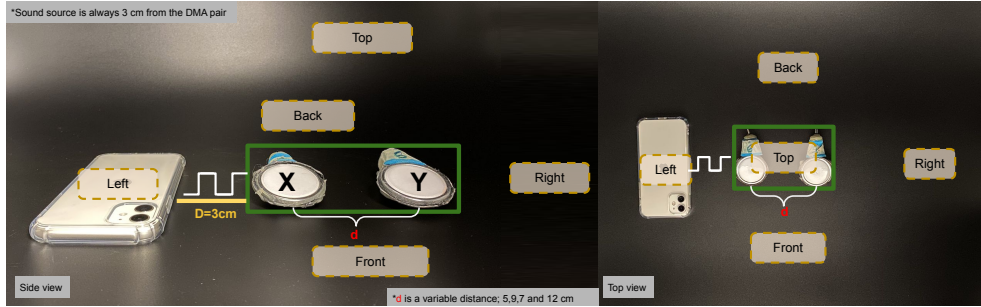


Figure 3.8.: DMA Discrete Frequency Response Setting. Phone as Sound Source and Stethoscope’s Head at a 3 Centimeters Separation with the Frequency Range of The Square Wave with 20 Seconds Duration at each Frequency. Microphone ”X” and Microphone ”Y” with a Variable Inter-Microphone Distance ”d” between 5 to 12 centimeters. The Sound Source Positions are Back, Right, Left, Top, and Front.

In a DMA, the separation between the elements of the array alters the gain frequency pattern. Thus, we conducted a discrete frequency sweep for four distances between the microphones-stethoscope pair. The selected distances were 5,7,9, and 12 centimeters. The distances were the average values of separation between the six stethoscope heads placed on the volunteers’ faces. DMA depends on the source position. Then, the sound source was placed in 5 different locations; top, back, front, right, and left, as shown in 3.8, at a 3cm fixed distance from the microphones. Fig. 3.9 depicts the results of the DMA frequency response. The signal’s gain of our system depends not only on the frequency range but also on source placement. In the low-frequency spectrum, and the source positioned on top of the DMA gives a higher gain. The frequency response in the low-frequency spectrum does not show a resonance peak for the nine centimeters, seven centimeters, and twelve centimeters separations between the microphone’s heads. This is in line with the formula $F_{null} = V/2 \cdot D$. Where, $V = 343 m/s$ is the sound speed, D is inter-spacing, and F_{null} is the tuned frequency [194]. In the case of DMA, end-fire dipole implies the need for $D = V/2 \cdot f_{null}$. Due to geometry reasons, it was not possible to achieve tuned directivity in the low-frequency range. The minimum distance of our design (5cm) has a tuning frequency of $f_{tune5cm} = 3.4KHz$ while the maximum (12cm) has $f_{tune12cm} = 1.4KHz$.

In our design, we employ the Eq. (3.3) as a calibrated first-order DMA to reduce gain discrepancies. This is applied for all the eleven combinations in Fig. 3.5. Assuming negligible difference in the geometry of the microphone-stethoscopes, and focusing on the low-frequency region between 21 and 751 Hz (see Fig. 3.9) and the sound source position on top of the DMA. The system remains with only the space position between the microphones’ dependency. Therefore, in the presence of a common sound source, the ADC-coded value will capture the space position relevance for such sound.

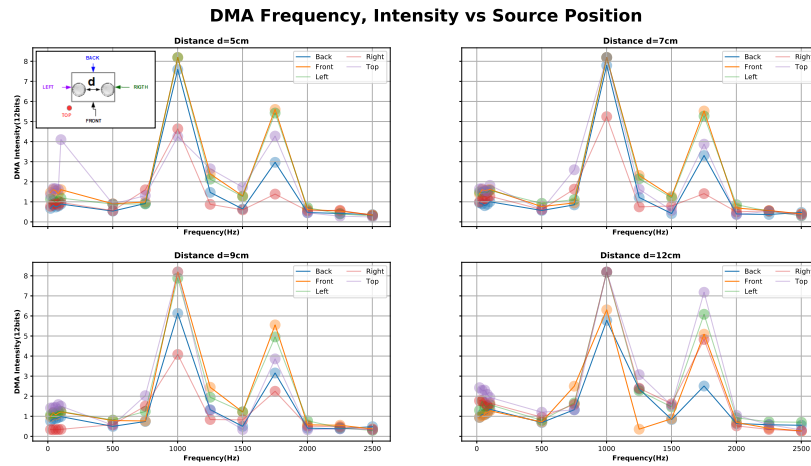


Figure 3.9.: Differential Microphone Array Discrete Frequency Response for Microphone Pair Distances Between Five, Seven, Nine, and Twelve Centimeters and Sound Source Positions, Back, Right, Left, Top, and Front.

3.1.6. Experiment Design

Eight volunteers were asked to mimic the dictionary in Fig. 3.2. The selection of categories is based on the Warsaw study [145]. Additionally, three flirting gestures such as winking, kissing, and tongue out are added. The gesture of taking a pill is purposely added. This is a highly relevant gesture in medication monitoring. Also, it is a gesture that represents the null class of sudden touches to the mouth, which are not related to facial expressions.

The participants were five women and three men between the ages of 24-29. They come from countries like Venezuela, Brazil, and India, and all of them were students at the Technical University of Kaiserslautern(TU-KL), Germany. This provides for reasonable ethnic diversity. To the best of our knowledge, all volunteers had normal eyesight and could perceive the presented expressions without prescription glasses. The participants did not have a problem identifying the facial expressions of themselves and others. All of them signed an agreement following the policies of the university’s committee (Technical University of Kaiserslautern) for the protection of human subjects, which approves experimental protocols at the university. The experiment was video recorded for further private analysis. There were no reported pandemics or any contagious disease outbreaks in the region during the experiment recording time.

We followed the protocol described in [165], asking the volunteers to emulate the facial actions with as little variability as possible. In addition to the pictures, the name of the expression was provided. While this clarified the action for some participants, it was, however, perceived as confusing by others. The experiments were performed in a closed office with a carpeted floor and only two persons inside, the person monitoring the experiment and

the participant.

The hardware was fixed on each volunteer’s head with all six microphone stethoscopes at the same time. The set of facial expressions was displayed in random order with ten repetitions per activity for eight volunteers. We used color-coded lights to prompt the subjects to start and stop mimicking each repetition of each action. So, when the graphical user interface (GUI) showed a green light, then the participant had to start making the respective action, and he/she stopped and went back to neutral expression when the red light went on. The duration per action was between two to three seconds, including the rest time (neutral event). This neutral event was considered as the null class because it is trained with data outside our gesture dictionary.

The same experiment was repeated three times (sessions) per volunteer with a gap/resting period of a few hours or days. The gap is introduced to ensure the participants’ facial muscles are properly rested, as the mechanomyography could be used as a measure of the fatigue of muscle [133, 209]. We are using the definition of fatigue as ”any reduction in the force-generating capacity regardless of the task performed” [12]. We collected 240 samples per activity(10 repetitions per gesture, three sessions for eight volunteers) for a total of 2640 (null class included) samples.

3.1.7. Signal and Data Processing

In this section, we explain our data analysis approach, including feature calculation, feature selection, and classifier selection. We evaluate our approach on an individual basis (user-dependent) and the dataset as a whole for a cross-user evaluation. All the validation is carried out with the leave-session-out scheme. A block diagram to summarize the steps is presented in Fig. 3.10.

Feature Selection

The DMA pairs showed in Fig. 3.1 are the inputs to the feature extractor. To explore various time series features we employed the open-source Python library Tsfresh version 0.16.0 [46]. Tsfresh is a time series feature extraction library based on scalable hypothesis tests. Using the library we extracted 754-time features per input of DMA (11 in total), for a total of 8294 features. For feature selection, Tsfresh provides a feature extractor based on the vector of p-values. Where a smaller p-value means a higher probability of rejecting the null hypothesis. To select the threshold for the p-value, the library uses the Benjamini-Yekutieli (BY) procedure[24]. A summary of the BY procedure would be; (1) organize the p-values from lower to higher (step-up) and (2) select a small group of them, where the boundary between the selected features is set by the condition $P_{(k)} \leq \frac{k}{m \cdot c(m)} \alpha$. Where P_k is the p-value, k is the last p-value to be declared as valid for a given α (rejecting the null hypotheses), m is the total number of hypothesis/features, and $c(m)$ is a constant defined as $c(m) = 1$ when the features are independent or positively correlated, and as $c(m) = \sum_{i=1}^m \frac{1}{i}$ when there is an arbitrary dependency(selected case). This

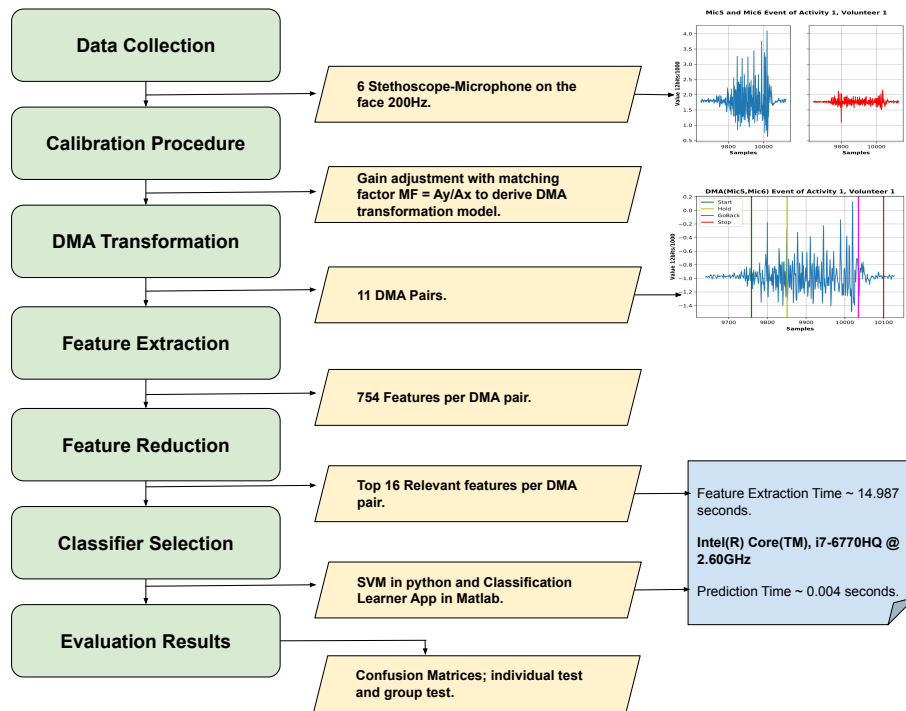


Figure 3.10.: Flow Diagram for the Facial Muscle Movement Recognition with Differential Microphone Stethoscope

relationship is a simple graph of p-values as dependent variable ("y") and independent variable ("x") equal to the range of 1 to k, with slope $= \frac{\alpha}{m \cdot c(m)}$.

The Benjamini-Yekutieli technique is used for feature reduction. The selected features were the sixteen top common features for all the participants. Sixteen features were extracted by each DMA pair in Fig. 3.1 for a total of 176 features. A list of the sixteen features is shown below:

- F1 80% quantile
- F2 10% quantile
- F3 Absolute FFT coefficient #94
- F4 Absolute FFT coefficient #38
- F5 Absolute FFT coefficient #20
- F6 P-Value of Linear Trend
- F7 Standard-Error of Linear Trend
- F8 Energy ratio by chunks(num-segments=10,segment-focus=1)
- F9 Energy ratio by chunks(num-segments=10,segment-focus=8)
- F10 Autocorrelation of lag=2

- F11 $c3 = \{E\}[L^2(X)^2 \cdot L(X) \cdot X]$ lag=3
- F12 Count below mean
- F13 Minimum R-Value of Linear Trend(chunk-length=10)
- F14 Largest fixed point of dynamics(PolyOrder=3,#quantile=30)
- F15 Ratio beyond r-sigma(r=1.5)
- F16 Mean change quantiles with absolute difference(qH=1.0,qL=0.0)

Classifier Selection

The next step is to find the best classifier for the selected features. In [205] there is evidence that Support Vector Machine (SVM) is an option for avoiding overfitting. Others [197, 216] have also achieved excellent results by using SVM with mechanomyography signals. In addition to the SVM option, we also decided to experiment with standard Matlab[®]classifiers.

We retained 33% as a hold-out from the training set for classifier fine-tuning and started by looking at the default setting for KNN(K-nearest neighbors), SVM, and Ensemble-classifiers (Bootstrap Aggregation(Bagging) and Sub-space) were tested. The best-performing candidates were then fine-tuned to obtain the optimal hyperparameters. The automatic performance metric was "accuracy" defined as $\frac{TP+TN}{TP+TN+FP+FN}$ where TN=True-negatives and FP=False-positives.

Grid search is used for hyper-parameters improvement[80]. Grid search is an exhaustive search based on a defined subset of the hyper-parameter space. In the SVM case, there exists a kernel parameter that we can use to estimate if our data is linearly or non-linearly separable. Using grid-search, we tested the kernel to 2 types, one linear and the other as a polynomial. We searched the best fit for a range of values of the regularization parameter (C) equal to [0.001,0.01,1,10], in case of polynomial, C=[7,8,9,10,12,15,20], degrees-options=[1,2,3] and the γ was set to $[1 \cdot 4/F, 1 \cdot 16/F, 1/F, 1/4 \cdot F, 1/16 \cdot F]$. Where F is the top sixteen features by DMA pairs. And, in the cross-user case, the Gaussian kernel was added with C=[3,5,6,7,8,9]. The validation of the grid-search selected was with 10fold cross-validation, and the performance metric was "recall" defined as $\frac{TP}{TP+FN}$; where TP=True-positives and FN=False-negatives. We focused on SVM in Python using the scikit-learn library [155] version 0.23.1 and compared the result with the standard Matlab classifiers as a baseline. The evaluation is performed in user-dependent and cross-user with a leave-one-session-out validation scheme.

3.1.8. Results and Discussion

In this section, we present the results of the user-dependent and cross-user evaluation scheme. Fig. 3.11, presents the confusion matrices for both cases. The best performance for the user-dependent models was obtained by the

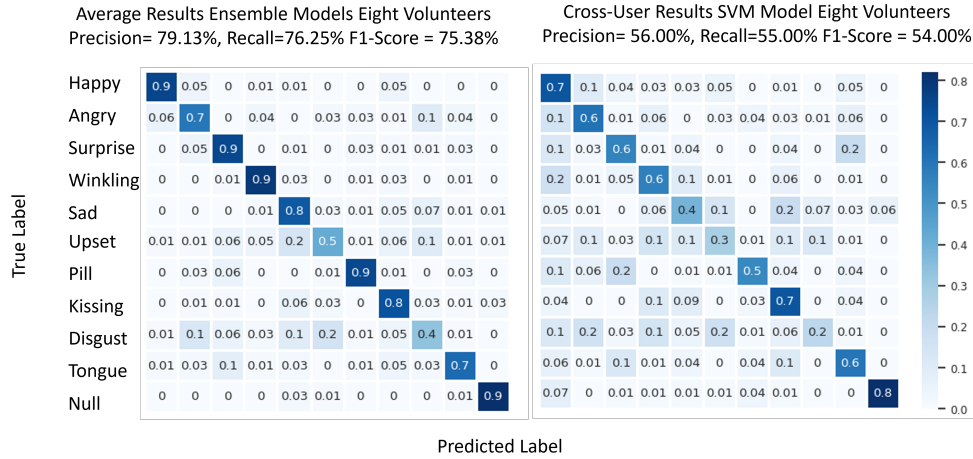


Figure 3.11.: *Left* Average Results of Eight Volunteer User-Dependent Evaluation with Ensemble Model in Matlab. *Right* Cross-User and Leave-One-Session-Out Evaluation Results of Eight Volunteers with SVM in Python.

Matlab ensemble classifier tool (see Fig. 3.11Left). The support vector machine (SVM) model in Python gave the best performance for the cross-user and leave-one-session-out evaluation scheme, as shown in Fig. 3.11Right. The Gaussian kernel is selected for the SVM model. Both results were obtained by substituting the maximum values for the matching factor MF from the Table 3.1. The selection gave the most balanced recall results for the case of the first volunteer’s data. The results in general are highly affected by the participant’s ability to mimic the facial movements in the dictionary Fig. 3.2.

For the user-dependent in Fig. 3.11Left the F1-Score is 75.38% which is above the chance level. And, for the cross-user evaluation, the F1-score is 54.00%, which is also above the chance level (around 9%). This confirms that the approach can extract relevant information about facial muscle activity patterns. In particular, the null class with more than 80.00% recall for both evaluation schemes means that the specificity of our system is reasonably good at picking relevant action from noise. In the cross-user case, the weakest facial movements to recognize are sad, upset, and disgust. On the other hand, for the user-dependent, the average recognition of the sad face improves to 80.00%. And, it can be seen that upset increases to 50.00% with 20.00% of the instances being confused with sad. This is reasonable due to the similarity between the mimicked faces of upset and sad in Fig. 3.2.

The results in Fig. 3.11 prove that the differential sound signal from the chosen locations contains information about relevant facial actions. Moreover, it shows that our processing chain manages to extract much of this information. The fact that the error is not equally distributed but instead some classes are recognized much better than others is an indication that the results are not limited by system noise but by the actual information content. The results must be seen in the context of two things that make achieving good results

difficult and indicate that the approach is suitable for real-life applications. First, given the diversity and complexity of facial gestures, from the point of view of machine learning, the training data set is relatively small. Second, each user recorded three sessions with a long (hours or days) pause between sessions and, most importantly, the sensors being removed and placed again on the user before each session. This means that sensor placement inaccuracies/variations, which are major concerns in many wearable applications, are already factored in the results.

The cross-user results show a minimum F1-score of 60.00% (Volunteer 8) while the score for the best user goes up to 89.00% (Volunteer 1). The differences between users can be attributed to three sources; (1) Physiological differences between users. (2) Different ways users may express specific actions. (3) Related to the last point, the inability of some subjects to mimic specific actions accurately.

A detailed understanding of which of the above accounts for which aspects of the system's performance requires further research, including a more detailed analysis of the correspondence between physiological actions and sound signals. Preliminary indications can be inferred from some qualitative observations. Thus, the most accurate volunteer was the person whose expressions were easier to decode by an observer. In the third volunteer case, we noticed this person was doing exaggerated imitations compared to the rest and was commonly moving the entire face in all the gestures. In the case of volunteers 4 and 8, their movements were more subtle than the rest. The next step must be to assemble a large number of user representatives both in terms of physiology and the type of expressions and investigate how advanced deep learning methods can generalize those for a more robust user-independent recognition.

3.1.9. Conclusion

We have demonstrated the feasibility of using differential sound mechanomyography as an unobtrusive mechanism for sensing facial muscle activity patterns. In particular, we have shown that sensors placed at locations roughly corresponding to the outline of typical smart glasses can provide enough information about muscle activity on the face as a whole to reliably identify meaningful expressions and facial actions. Our approach has an average F1-score of 75.38% for the user-dependent case and an F1-score of 54.00% for the cross-user evaluation. Key specific takeaways are:

1. Using differential signals between suitable pairs of microphones is a key feature of our system. This is probably related to the fact that it captures temporal patterns of muscle activation rather than a precise sound corresponding to the specific type of activation of a particular muscle. It also helps us deal with inter-person variability and noise.
2. The eyebrows-cheeks' positions are the most informative locations for most of the investigated gestures and actions.

3. Using a stethoscope-like sound acquisition setup has significantly improved the signal quality.
4. We have also seen a strong dependency on the person's ability to recognize and mimic the expressions with the best user reaching an F1-score=89.00% and the worst one being 60.00%.

Our solution has some limitations. The size of the stethoscope microphones is not comfortable to wear daily. It is possible to reduce the number of stethoscope pairs to the most relevant positions, such as eyebrow and cheeks or temple and cheek muscles. The reduction of the stethoscope geometry is desirable to embed the solution into a smart glasses frame and increase the ubiquity. In the next section, we will also investigate the fusion of differential sound information with other sensing modalities in particular with inertial measurement units (IMU), and pressure mechanomyography (PMMG).

3.2. InMyFace: Inertial and Mechanomyography-Based Sensor Fusion for Wearable Facial Activity Recognition

3.2.1. Problem Statement

Facial expression recognition is a complex problem for several reasons. First and foremost, its large interpersonal variability [215]. It is influenced by cultural background [87, 173], age, sex [190], race, and other person-specific characteristics, leading to a user-dependent solution in many related works [213]. Recently, in [123], the authors have formally studied the impact of a sex-balanced dataset on the fairness of the results for facial expression recognition (happiness, sadness, surprise, fear, disgust, and anger). They conclude that training with the mixed dataset achieves the best results in all cases. Furthermore, fairness is compromised in training with a highly biased dataset, especially when classifying particular expressions. In addition, different expression categories may have only minor differences (e.g., anger and sadness, as depicted in Fig. 3.12)

Recognizing facial activity is a well-understood (but non-trivial) computer vision problem. However, reliable solutions require a camera with a good view of the face, which is often unavailable in wearable settings. Furthermore, in wearable applications, where systems accompany users throughout their daily activities, a permanently running camera can be problematic for privacy (and legal) reasons. In addition, as the example of GoogleGlass has shown, having a permanently body-worn camera in everyday situations can be socially awkward and even illegal in some countries. Computer vision-based methods also tend to have a larger memory footprint and power consumption than non-visual sensor-based solutions, such as the one presented in [188].

Monitoring facial expressions has been investigated in many pervasive computing applications. Examples include novel human-computer interfaces [65, 213], learning feedback [149], and recognizing a car's driver's fatigue or mood

[83, 90]. Moreover, facial expression monitoring is particularly relevant during multi-user activities for automatically analyzing non-verbal behaviors for detecting group membership [114] or to gain insights into the participants' emotions and engagement levels. For example, a smile on a participant's face could indicate that they are enjoying the activity, while a furrowed brow could suggest confusion or dissatisfaction. By monitoring participants' facial expressions, we can adjust the activity in real-time to maximize engagement and satisfaction, an essential factor in providing immersive experiences while playing console games [70]. This type of behavior is also known as situation-aware wearable computing systems, where wearable devices can sense and understand what is happening in the environment to adapt their behavior and anticipate users' needs[49]. Overall, creating a relationship between facial expressions and multi-user activities can provide valuable insights into human behavior and help optimize the design and delivery of these activities.

Alternatively, multimodal approaches have been studied to exploit the limitations of independent sensing modalities. The idea is to combine multiple data sources with complementary information to reduce ambiguity, add completeness to the situation being studied ("gain in representation"), improve the signal-to-noise ratio (assuming independent error sources), and increase the confidence in the model decision ("gain in robustness"). In [127], the authors proposed a combination of IMU and 16 optical sensors in smart glasses to detect eight temporal facial gestures. Obtaining F1 score results of 91.10 % for the case of one model per person for the recognition of facial action units (AUs: AU12, AU27, LP, AU1+2, AU4, AU43, AU46R, AU61) [55]. Optical sensors for facial muscle movements limit the system to stable light conditions. In [115], another solution based on the fusion of IMU and electrooculography is presented to recognize kissing gestures, obtaining an accuracy of 74.33% in a cross-user scheme. IMU is the common basis in the multimodal scheme, so it is a source of information to be considered.

This work presents an alternative multimodal solution based on the fusion of wearable inertial sensors, planar pressure sensors, and acoustic mechanomyography (muscle sounds). The facial expressions to be evaluated come from the Warsaw Photoset (seven expressions) and [21] (two facial movements); thus, it will be simpler to compare with future solutions, see Fig. 3.12. The sensors were placed unobtrusively in a sports cap to monitor facial muscle activities related to facial expressions. We present our integrated wearable sensor system, describe data fusion and analysis methods, and evaluate the system in an experiment with thirteen subjects from different cultural backgrounds (eight countries) and both sexes (six women and seven men). In addition, our unique set of participants and minimally biased experimental design demonstrate the inclusiveness of the approach, which is beneficial for further generalizability.

3.2.2. Contributions

In summary, our contributions are:

- We present a multimodal sensing alternative for facial muscle motion

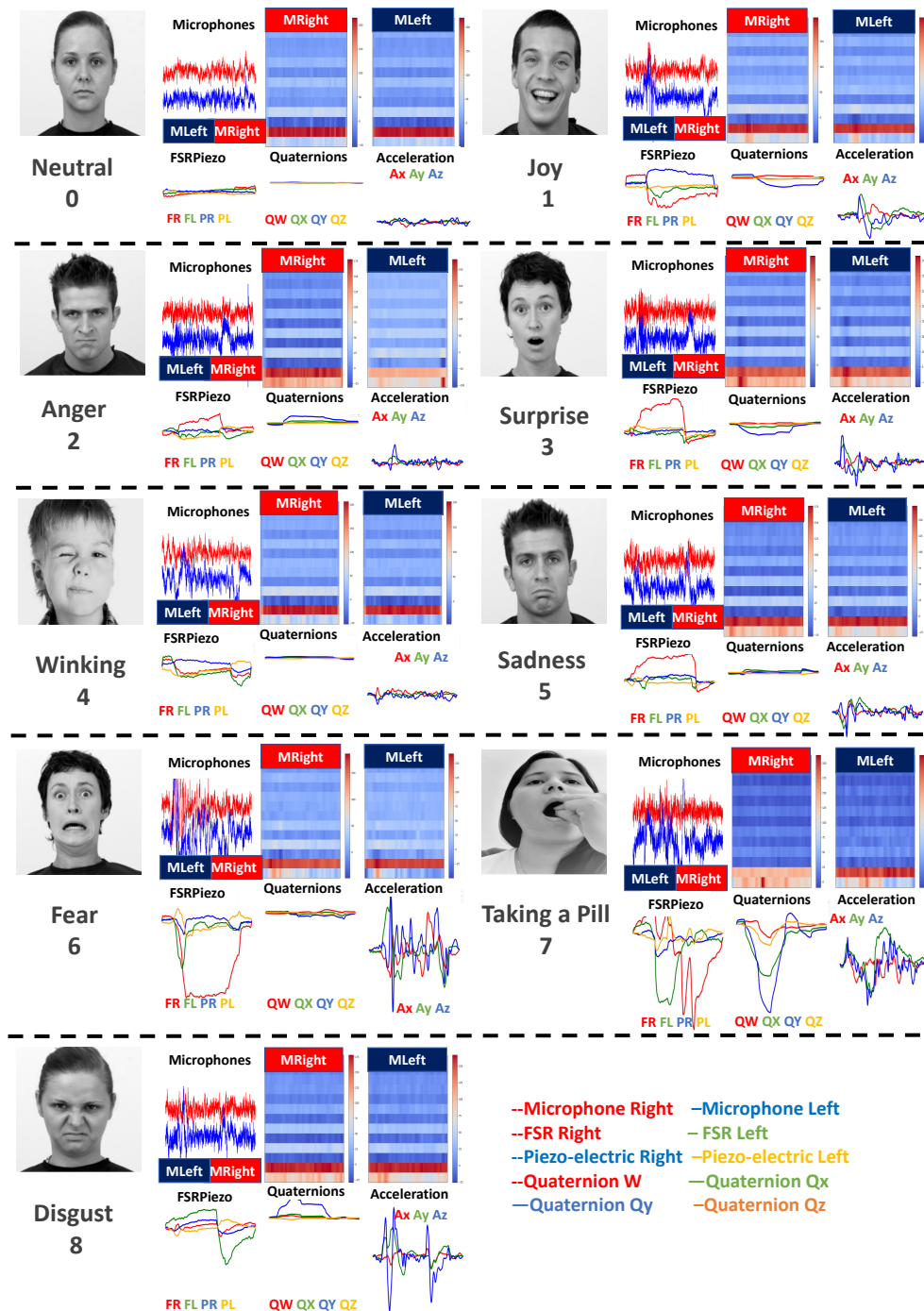


Figure 3.12.: Facial Muscle Activities Dictionary with Sensor Signal Examples; 7 Facial Expressions from Warsaw Photoset [145] and 2 Gestures from [21]. Two Channel Raw Audio Data, Thirteen Mel Frequency Cepstral Coefficients (Two Channel Audio-Monophonic), Force Sensitive Resistor, Piezoelectric Film, Orientation and Acceleration.

monitoring based on inertial, planar pressure, and acoustic sensors dis-

tributed in a minimally obstructive wearable accessory (sports cap). Furthermore, the idea can be adapted and potentially unobtrusively integrated into other head-worn platforms (e.g., glasses and headbands) without compromising the sensing capability.

- We adopt a modular multimodal fusion method based on sensor-dependent neural networks using a late fusion approach with a low memory footprint (≤ 2 MB) to simplify the future deployment of the idea in wearable/embedded devices with tiny dimensions and reduce memory (4 MB to 16 MB Flash).
- We conduct a user study with thirteen participants from diverse cultural backgrounds (eight countries) and both sexes (six women and seven men). Our unique dataset demonstrates the inclusiveness and generalizability of the results. For the evaluation, we use the Warsaw Photoset [145] plus two facial gestures from [21].
- We evaluate the system using a hybrid fusion approach with locally connected inception blocks with dimension reduction per sensing modality for the best eight imitators of the facial expression dictionary in Fig. 3.12, and six classes.

3.2.3. Approach

Our system combines inertial, pressure, and audio sensors to recognize facial muscle activity from an unobtrusive sports cap platform. Fig. 3.12 presents the facial muscle movements dictionary. The IMU has already demonstrated its potential to distinguish various face- and head-related movements [115, 127]. PMMG provides a flexible and comfortable solution for facial gesture recognition with moderate accuracy [221]. AMMG for facial muscle activity recognition was used in [21]. However, the design was bulky and obtrusive to achieve high sensing accuracy. Piezoelectric thin films (PEF) have been used in real-time [188] to detect and classify skin deformation to decode facial movements in patients with amyotrophic lateral sclerosis. Their work is intended to be used in clinical settings for nonverbal communication and neuromuscular monitoring conditions. PEF sensing technology is lightweight, customized, and with mechanical harvesting capability [168]; therefore, we could claim that PEF technology is worthy of research and study in specific applications. Combining the three sensing modalities with an appropriate sensor fusion pipeline allows us to achieve high accuracy with an unobtrusive system. Here, we proposed to fuse passive sensing such as; pressure mechanomyography (PMMG) using a force-sensitive resistor (FSR) and piezoelectric film (PEF), inertial sensing based on orientation and acceleration, and acoustic mechanomyography (AMMG).

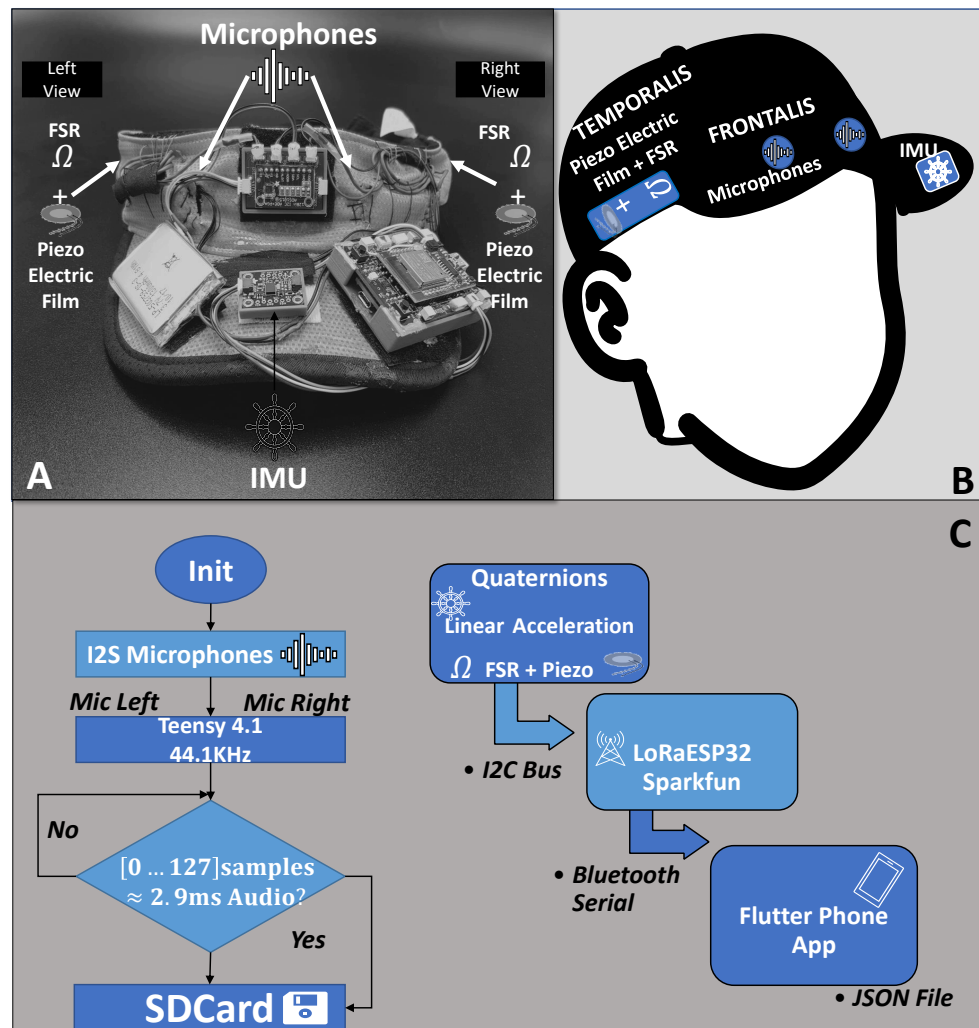


Figure 3.13.: Hardware Prototype and Data Collection Diagram. **A** Sports Cap with Sensors Distribution. **B** Sensor Placement on the Frontalis and Temporalis Muscles of the Participant. **C** Data Acquisition Diagrams with the Custom Printed Circuit Board.

3.2.4. Apparatus

The prototype hardware is shown in Fig. 3.13. The FSR sensors were distributed on the frontalis and temporalis muscles using a sports cap as a wearable accessory. Two Inter-IC Sound (I2S) microphones sampled at 44.1kHz were placed on the left/right side of the cap as shown in Fig. 3.13. A custom printed circuit board (PCB) based on Teensy 4.1¹ and LoRaESP32² was used to sample the sensor data. The integrated SD card of the Teensy 4.1 is used to store the acoustic data. The microphone openings are pointed towards

¹Teensy 4.1, Paul Stoffregen: <https://www.pjrc.com/store/teensy41.html> DLA: January 2, 2025

²LoraESP32, Sparkfun: <https://www.sparkfun.com/products/18074> DLA: January 2, 2025

the frontalis muscle to capture the mechanical sound information (AMMG). The analyzed sound information comes from the muscle movements, so it is a privacy-aware system (no speech or ambient sounds). Pairs of FSR and PEF were placed to the left and right of the temporalis muscle. The FSR and PEF were selected to measure the PMMG generated by facial muscle movements. An IMU was in the sports cap viewer to capture head movements related to facial expressions. FSR, PEF, and IMU data were transferred by Bluetooth Serial (BT) to a cell phone application (Flutter Framework [140]) to save them in a JSON file for further analysis. The sampling rate was about 100 Hz for FSR, PEF, and IMU data. The Bluetooth communication scheme limited the data acquisition but is still fast enough to capture micro and macro expressions with a duration $\leq 200ms$ and duration ≥ 200 ms, respectively [170]. Detailed sensor data acquisition diagrams are in Fig. 3.13.

The sensor distribution (around the head) and the dimensions/weights of the selected sensors, see Table 3.2, make our design suitable for integration into other head-related accessories such as headbands and glasses. In our sports cap, the dimensions and weight are negatively affected by the selected battery (1300 mAh) and the main board of the prototype (6.5 x 4.3 cm); both could be reduced in a future design and improve user comfort.

The power consumption of the prototype is around 0.22A at 4.97V (1.09 Watts continuous mode)³. Several possibilities exist for further power consumption reduction based on the overall concept. On the one hand, FSR-PEF data could trigger IMU and audio data acquisition, reducing power consumption to 0.06A at 4.99V (0.3 Watts) when no pressure is detected in the temporalis muscle. On the other hand, PEF is a mechanical energy source. Mechanical energy is considered ubiquitous ambient energy that can be converted into electric power [168]. Employing piezoelectric as a mechanical energy harvesting mechanism is an active field of research [168, 195, 218]. Harvesting energy from human motion and at the same time classifying such motions has also been demonstrated before [57, 118, 135, 218] which could be employed to reduce the power consumption further.

In addition to mitigating power issues, the FSR has the advantage of being robust against motion artifacts. We thus recommend the use of the slope/-gradient of the FSR signal as a possible trigger for automatic segmentation of the input data and to avoid motion artifacts. For example, the signal's slope of acceleration data was automatically used to segment MMG data in [204]. Finally, although we use off-the-shelf and non-textile FSR and PEF sensors for fast prototyping, it is noteworthy that both technologies are already available in textiles [135, 206, 221].

In Fig. 3.14, the synchronized multimodal signals are displayed. Due to the variability of the sampling rate between the different sensing modalities, the synchronization is based on the time-stamped by the camera, which acts as a global timer. At the beginning of data collection, a button is pressed in front of the camera, and this signal is used to calculate the delay between the

³USB Digital Power Meter: <https://www.az-delivery.de/en/products/charger-doktor>
DLA: January 2, 2025

Table 3.2.: Sensors Characteristics

Sensor	Manufacturer Name	Dimensions (cm)	Weight (grams)	Benefits
FSR ¹	Alpha MF01A-N-221-A01	1.25 diameter	0.26	Ultra-thin and flexible
PEF ²	TE SDT1-028K shielded	4.45 x 1.97 x 0.32	0.30	Low noise, shielded and flexible
Microphones ³	Knowles SPH0645LM4H	0.35 x 0.26 x 0.09	0.40	High SNR of 65dB(A), Flat Frequency Response, Omnidirectional
IMU ⁴	Bosch BNO055	0.38 x 0.52 x 0.11	0.15	Outputs fused sensor data

¹ https://www.mouser.de/datasheet/2/13/MF01A__c3_a2_c2_96_c2_a1_A01-1915118.pdf DLA: January 2, 2025

² <https://www.te.com/usa-en/product-CAT-PFS0010.html> DLA: January 2, 2025

³ <https://media.digikey.com/pdf/Data%20Sheets/Knowles%20Acoustics%20PDFs/SPH0645LM4H-B.pdf> DLA: January 2, 2025

⁴ https://www.mouser.de/datasheet/2/783/BST_BNO055_DS000-1509603.pdf DLA: January 2, 2025

camera and sensor signals. The delay between the sensors and the camera remains constant throughout one experiment session.

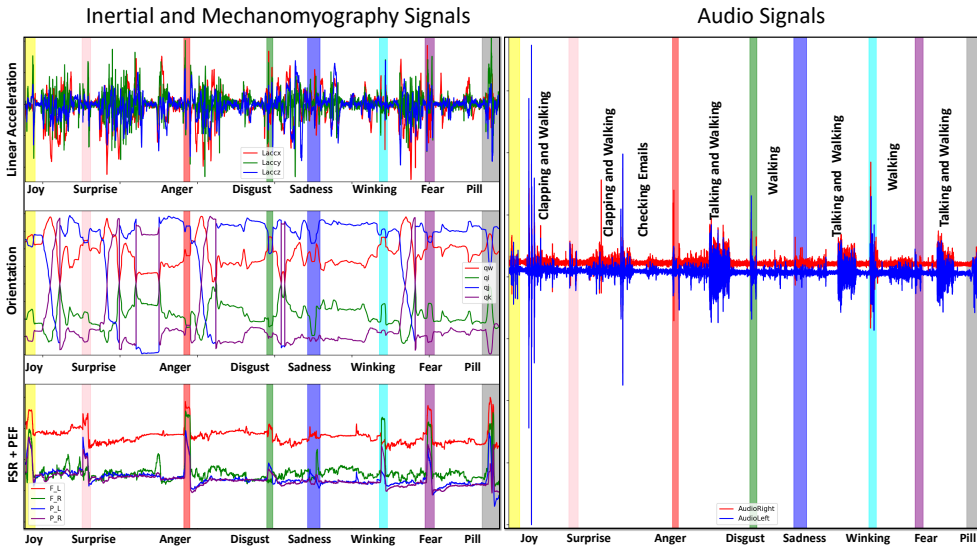


Figure 3.14.: Comparison between Synchronized Sensors Signals for the Dictionary of Facial Movements in Fig. 3.12 (without neutral) Versus Activities, such as; Clapping, Walking, Checking Emails and Talking. Joy is yellow, Surprise is pink, Anger is red, Disgust is green, Sadness is blue, Winking is magenta, Fear is purple, Taking a Pill is grey, and the in-between (white spaces) are the noise factors activities.

Fig. 3.14 compares the facial expression-generated signals to those generated by noise factors, such as clapping, walking, checking emails, and talking. The robustness to environmental noise and not facial-related movements of the multimodal fusion approach can be observed by the distinctive FSR-PEF signals. In the case of facial muscle movement, the signal is remarkable even to the naked eye. While affecting the audio signal, the environment's noise does not influence the PMMG or inertial-based sensors. The movements generated by clapping and walking affect the inertial sensors more than the FSR-PEF

sensors. The inertial information is quite distinctive inside our dictionary without considering the noise factors. Hence, the fusion power of the FSR-PEF to signal muscle movement detection is encouraged as a simple and robust technique to avoid undesirable confusion. The proposed multimodal fusion favors noise reduction and reduces the algorithm’s complexity for detecting facial muscle motion.

3.2.5. Multimodal Sensor Fusion

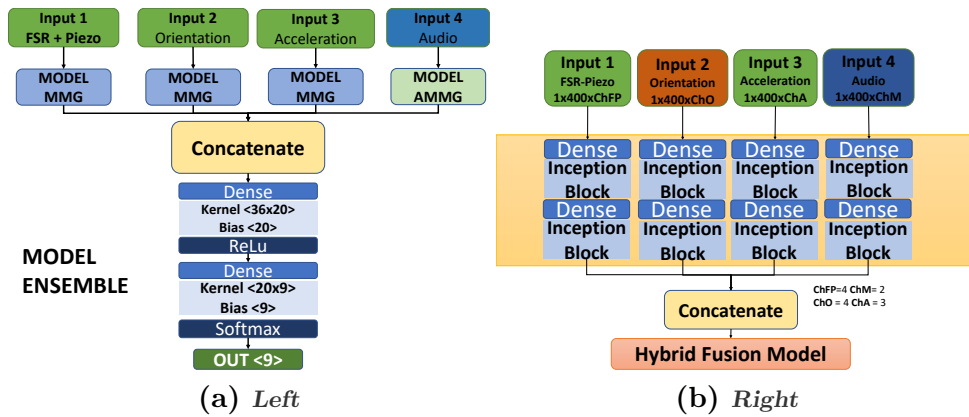


Figure 3.15.: *Left Ensemble Multimodal Sensor Fusion Model Overview; Fusion in the Prediction Phase. Right Hybrid Multimodal Sensor Fusion Model Overview; Fusion Within Hidden Layers, and Before Prediction.*

Multimodal sensor fusion can be broadly classified into early and late fusion, depending on the position of the fusion within the processing chain [61]. Our primary approach is referred to as late fusion. The fusion is performed in the individually trained networks’ decision phase (confidence scores). The late fusion method has the advantage of extracting the specific patterns of each sensor independently and the parallel deployment of each sensor-dependent neural network (NN) on multiple microcontrollers (MCUs), reducing the recognition latency and memory requirement per MCU (≤ 2 MB per network). The main drawback of late fusion is the limited potential to extract cross-correlation between sensing modalities and channels. Additionally, we also explored a hybrid fusion alternative. The fusion performed in the hidden layers of the neural network and before the decision layer is called hybrid fusion. In our work, the hybrid fusion structure and evaluation are made considering the outcomes from the late fusion performance. An overview of the late fusion and hybrid fusion diagrams is depicted in Fig. 3.15. The Fig. 3.15a **Left** shows the concatenation of the sensor-dependent models after the decision phase using an ensemble NN. The specific sensor-dependent models are explained in Section 3.2.5. The Fig. 3.15b **Right** presents the primary blocks of the hybrid fusion method. The blocks consist of sensor inputs, inception blocks, and hybrid fusion NN in Section 3.2.5.

In the case of late fusion, our approach leverages sensor-dependent fusion techniques to combine heterogeneous sensing modalities to recognize facial muscle motion ubiquitously. Furthermore, for the case of hybrid fusion, we propose to employ the inception block with dimension reduction as an early feature extractor per sensing modality to reduce the complexity of the NN and the number of parameters. This method can be easily exported to other types of sensors as a size-reduction technique for heterogeneous NN-based algorithms.

Multimodal Ensemble Late Sensor Fusion

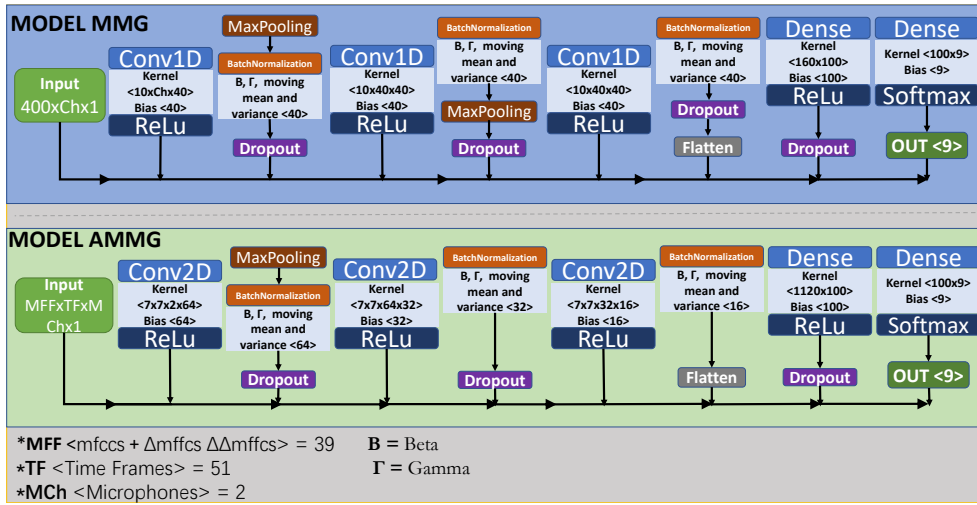


Figure 3.16.: Sensor Dependent Neural Network Models. **Top** Neural Network for FSR-PEF (PMMG) and IMU(Orientation and Acceleration) Information. **Bottom** Neural Network for Mel Frequency Cepstral Coefficients of Audio (AMMG).

We employ sensor-based late fusion [139] as depicted in Fig. 3.15a **Left** and Fig. 3.16. The data was divided by sensor type, and we implemented four sensor-dependent neural networks (NN) models in Fig. 3.16. This approach gives each NN the advantage of learning the unique properties of each modality and facilitates a simple fusion method. In the initial step, the inputs to the sensor-dependent NNs were pre-processed in a sensor-specific manner, as described below. The data of each movement was processed as a whole instance. The NNs were developed using the TensorFlow framework (version 2.9.2). The training included early stopping with the patience equal to 30 and restored weights option equal to true to avoid overfitting and ran for 500 epochs. In the individual models, the learning rate was manually tuned for each participant, and for the case of one ensemble model for all participants (cross-user case), it was set to 0.03. Categorical cross-entropy loss function and Adagrad optimizer were used to optimize the sensor-type NN. The ensemble NN consisted of one fully connected layer of 20, Adam optimizer (0.01), and softmax function with nine probability outputs, see Fig. 3.15a **Left Model Ensemble**. The details

of the sensor-type dependent NNs are explained below.

Pressure Mechanomyography (PMMG) and IMU: We sense PMMG with a combination of FSR and PEF. For the case of inertial sensing, the quaternions were selected for orientation to avoid the gimbal lock problem [29] and to improve stability. The FSR, PEF, and inertial sensors (quaternions and acceleration) data were normalized by subtracting the average of the gesture’s first (starting point) and last values (ending point). Since each facial event is a temporal series with variable lengths, a dynamic resample procedure to 400 samples was applied [22, 23]. Then, the resampled signals were fed to a first-degree Butterworth low pass filter with a cut frequency of 5Hz to remove the ringing peak in the signal’s edges coming from the resampling procedure and to highlight the low-frequency range.

FSR and PEF signals were treated as a pair; hence they have a joint NN. Orientation and acceleration were processed in separate NNs. In total, three networks were trained for PMMG and inertial sensing. The NN structure was based on a modified 1D-LeNet5 network [108] as depicted in Fig. 3.16 **Model MMG**. The network consisted of a convolution (conv)—max pooling (maxpool)-conv-maxpool-conv—fully connected (fc)-fc-softmax layers with batch normalization and dropout on the convolution layers. The convolution layers contain 40 filters, a kernel size of 10, and the activation function ReLu. For max pooling, the pool size was (40, 40) for the first convolution (400, 40) and (4, 40) for the second convolution (40, 40). The third convolution was of size (4, 40) without pooling. A flattening layer of 160 was followed by a fully connected layer of 100. The nine outputs for the different facial muscle activities in Fig. 3.12 are then converted into probabilities by a fully connected layer and softmax function. A detailed view of the sensor-dependent neural network structure for the PMMG and IMU is shown in Fig. 3.16 **Top Model MMG**.

Acoustic Mechanomyography (AMMG): As shown in Fig. 3.13, two I2S microphones sampled at 44.1kHz were positioned on the sports cap to cover the frontalis muscle of the volunteer. The audio information was used as AMMG as proposed in [21] but without the stethoscope to amplify the audio. Two channels of audio were resampled to 52000 (1.17 seconds). Muscle’s audio data was transformed to the mel spectrum to reduce the dimension of the audio and speed up the convergence time of the neural network for a small dataset. The Mel Frequency Cepstral Coefficients (MFCCS) have been used for many applications [39, 176]. Thirteen Mel filters are a common choice in the literature for audio analysis [39, 75]. The short-time Fourier transform (STFT) parameters were defined as; a sampling rate of 44.1kHz, hamming window and size of 4096 (93 ms), and hop length of 1024 (23 ms). Although we are not using speech information, the STFT configuration was selected to match the default configuration defined in [132] (the typical setting for speech analysis) for simplicity. In addition to the MFCCS, we also calculated the first and second derivatives of the MFCCS to boost the recognition accuracy [79]. The input shape to the NN was (39,51,2); 13 MFCCS, 13 MFCCS’ delta, and 13 MFCCS’ second delta, 51-time frames, and two audio channels. The

deep learning model was defined as two-dimensional convolution (conv2D)-max pool-conv2D-conv2D-fc-fc-softmax layers with batch normalization and dropout on the convolution layers. The first, second, and third conv2D consisted of 64,32, and 16 filters, respectively, with a kernel size of 7 and an activation layer of ReLu as depicted in Fig. 3.16 **Bottom Model AMMG**.

In the case of a future real-time implementation, the modular approach offers the flexibility of sharing the computation of each NN model between microcontrollers (MCUs). The NNs developed in this work are less than 2 MB, a practical size for an embedded device (typically between 4 and 16 MB of flash), and with the TensorFlow Lite framework⁴, can be compressed for use on mobile/embedded devices. Compared to a wearable camera solution with 162 MB in [41], our 2 MB distributed NNs are two orders of magnitude smaller. Therefore, we could schedule the FSR-PEF and IMU model (Fig. 3.16 **Model MMG**) into MCU-one and the audio model (Fig. 3.16 **Model AMMG**) into MCU-two, which in the end will be merged by the ensemble model using the Bluetooth-capable phone. As lightweight and individually deployable models in the MCU(s), only the prediction results are sent out of the embedded device, which will maintain the privacy of the user’s data.

Multimodal Hybrid Fusion

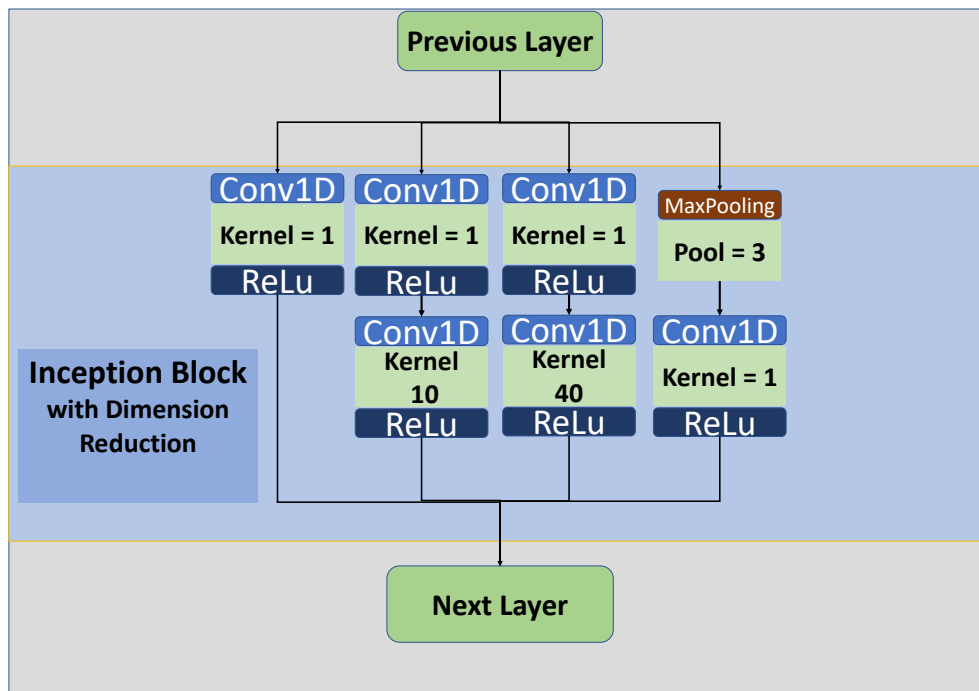


Figure 3.17.: Modified Inception Block with Dimension Reduction Locally Connected per Sensing Modality

⁴Machine Learning for Mobile and Edge Devices - TensorFlow Lite <https://www.tensorflow.org/lite>

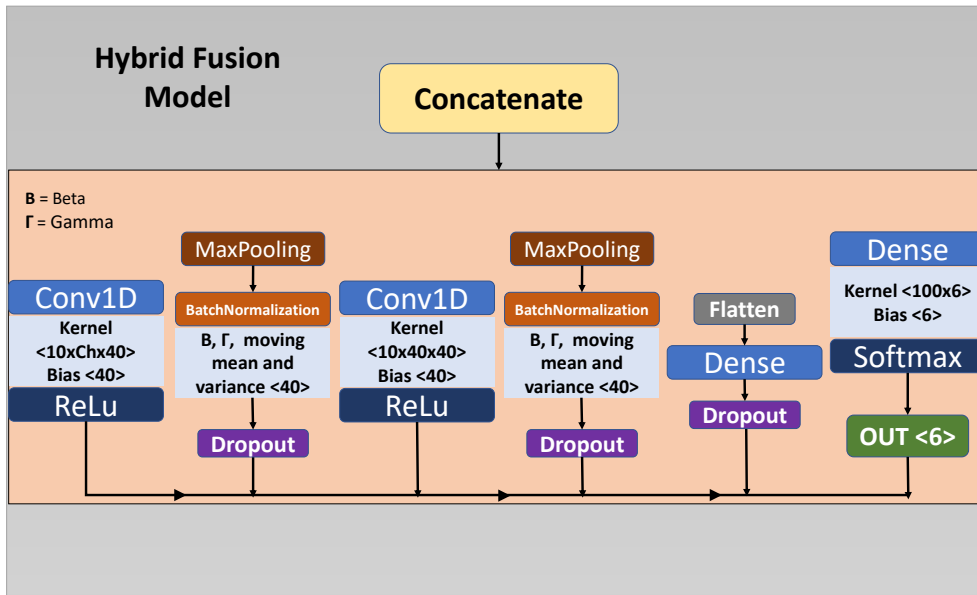


Figure 3.18.: Multimodal Hybrid Fusion Model

An early/hybrid fusion approach learns the contributions to the recognition performance of all sensing modalities as a unit, exploiting the cross-correlation between information sources. On the other hand, the NN structure for early/hybrid multimodal sensor fusion could be complex to give competitive results, consequently increasing computational time [61, 139]. In some cases, fusing low-level features might be irrelevant to the task, thus decreasing the fusion power [117]. Hence, the level at which the merge is made will influence the quality of the information being fused. Additionally, this fusion method requires all the sensor's data to enter the same processing pipeline, reducing the parallelism and increasing deployment complexity in embedded devices. The challenge is maintaining a low-complexity model while achieving performance comparable to the late fusion approach. In this section, we evaluate the performance of such a sensor fusion methodology to test our system's facial muscle movements recognition capability.

Our experiment is based on imitating facial movements related to typical facial expressions. Therefore, the ground truth could be affected negatively by the imitation ability of the volunteers (with no acting experience). In this section, the eight best imitators were selected to test the system's feasibility without outliers due to the imitation inaccuracy of our participants. The selection is based on the visual perception of the similarity of the volunteers' facial dictionary imitation performance, as assessed visually by most participants. The main drawback of reducing the number of volunteers is that it increases the risk of overfitting, which is already high due to the higher complexity of hybrid fusion modeling compared to late fusion for our specific task. Therefore, we employ techniques developed to reduce the risk of overfitting. These techniques include: early stopping with patient equal to 30 and restored weights option enabled, batch normalization, and max pooling, the evaluation

scheme is defined by leaving one session out, and the NN structure is based on a lightweight NN designed to reduce the number of parameters. We focused on using the main block of one of the most popular lightweight NN as the stem network [191], locally connected to each sensing modality, known as the inception block as shown in Fig. 3.17. These main blocks were explicitly designed to reduce the number of parameters while maintaining a balance between latency and accuracy to allow their deployment in mobile/embedded systems. Moreover, the loss function performs soft labeling (also called label smoothing) with 10% distributed equally over the opposite classes. Hard labeling is the typical way of assigning labels to class members, where class membership is binary, i.e., labels are true/false. In soft labeling, class membership is based on a probability score assigned to each member. Thus, a 10 % soft labeling means that, for example, the probability of class 1 is 90 % when class 1 is the truth value, and 10 % is equally distributed over the opposite classes.

To reduce the complexity of the hybrid model structure, we have reduced the number of classes based on the risk of confusion. The joy and surprise classes are merged due to the similarity of visual perception of facial gestures, assessed visually by most participants (survey). In this case, the visual perception excludes (on purpose) the gesture of opening the mouth to focus on the direct sensing areas of a sports cap attachment, such as the frontal and temporalis muscles, leaving the indirect perceptual area of the masseter out of the equation. These two classes differ only in the mouth movement, and the placement of the sensors on the forehead makes their recognition indirect, which could lead to considerable confusion between them. In addition, the following section Section 3.2.7 shows that the confusion in the late fusion approach is mainly dominated by two pairs of classes, "Disgust-Anger" and "Sadness-Anger", when all sensing modalities are combined, leading us to merge these categories. The classes recognized were Neutral, Joy + Surprise, Disgust + Sadness + Disgust, Wink, Fear, and Take a Pill, for a total of six categories. Despite all the parameters and complexity reduction techniques, the hybrid fusion in our case reached 939,706 parameters (2.35x Late Fusion Parameters). However, based on the number of parameters is still a small network compared to typical networks considered smallish, such as MobileNetV2 with 3.5 million parameters [161].

As depicted in Fig. 3.15b, we employed inception blocks locally connected to each sensing modality, followed by a concatenation. The output is then fed to a hybrid fusion model/NN. Before entering the NN, the same signal processing steps performed in Section 3.2.5 are applied per sensor type. Our data sources are not spatiotemporally aligned due to the number of channels/resolutions and different sampling frequencies. Therefore, a certain degree of preprocessing is necessary before concatenating them using a CNN. As a result, the data of all sensors were in the form (Time steps = 400, Channels). The channels were four, four, three, and two; for PEF and FSR, orientation, linear acceleration, and audio, respectively. The locally connected inception block structure is depicted in Fig. 3.17. The inception block is then followed by a hybrid fusion model as shown in Fig. 3.18. The hybrid model consisted of conv1D-

maxpool-conv1D-maxpool-flatten-fully connected(fc)-(fc)-softmax layers with batch normalization and dropout on the convolution layers. The convolution layers contain 40 filters, a kernel size of 10, and the activation function ReLu. For max pooling, the pool size was (40, 40) for the first convolution (400, 40) and (4, 40) for the second convolution (40, 40). A flattening layer of 160 was followed by a fully connected layer of 100. The six outputs for the six categories (Neutral, Joy + Surprise, Anger + Disgust + Sadness, Winking, Fear, and Taking a Pill) are then converted into probabilities by a fully connected layer and softmax function. The training ran for 200 epochs with early stopping enabled with a patient equal to 30 and restoring weights option enabled. Categorical cross-entropy loss function and AdaDelta optimizer with a learning rate of 0.09 were used to optimize the sensor-type NN. AdaDelta optimizer is a method that performs an adaptive learning rate per dimension, and its main advantage is that there is no need to select a global learning rate. Moreover, it can handle the intrinsic continuous decay of learning rates throughout training [212].

3.2.6. Experiment Design

Thirteen participants from diverse backgrounds (Germany, Italy, Peru, India, France, China, Republic of Korea, and Venezuela) mimicked the facial muscle movements defined in Fig. 3.12 in a random sequence per session while wearing our sports cap prototype.

The dictionary of Fig. 3.12 contains seven of the facial expressions proposed in the Warsaw Set of Emotional Facial Expression Pictures (Warsaw Photo-set) [145], which is a database of high-quality photographs of genuine facial expressions. The photographs were taken after appropriate training, and the participants (actors) were inclined to express felt emotions. In addition, the facial movements of taking a pill and winking [21] were added to the dictionary to extend the scope of recognition beyond the imitated expressions. Specifically, taking a pill encompasses a more sophisticated/complex facial gesture, including picking up the pill from the table, going to the mouth and opening the mouth, tilting the head back, and swallowing the imaginary pill. The gesture of taking a pill can be replaced for a future application by the typical behavior of users snacking while playing video games. Snacking while playing includes taking a snack from the table, bringing it to the mouth, and chewing/swallowing. This cycle will generate facial muscle movements, which are not related to tracking the player’s satisfaction with the game, and most of the snacking steps are contained in the gesture of taking a pill.

It is essential to accentuate that the system recognizes consciously simulated facial expressions and not authentic expressions, as in previous hardware-related works, to test the feasibility of fusing the sensing modalities. Therefore, we emphasize that what we recognize are ”facial muscle movements” related to creating features similar to what we expect to see in genuine facial expressions.

Participants’ muscle movement sounds and pressure patterns were recorded without any additional conditions other than that they mimic the dictionary as closely as possible. It should be noted that the volunteers did not receive

any prior training and that the facial muscle movements were executed as they saw fit. Subjects were not forced to make sounds or restricted to a specific time to perform the facial muscle movements. Therefore, the amount of time for each gesture is variable, even for the same participant. All participants performed five sessions. One session consisted of four randomized/shuffled appearances of each face gesture within the dictionary, which is used to avoid muscle fatigue and avoid correlation between instances of the same gesture. A total of 180 instances per volunteer were collected. The neutral face marks the start and end points of a gesture. In total, 2160 valid facial movements were collected.

Mechanomyography provides the timing requirements for the experiment to avoid corrupting the data because of the tiredness of the participant [209]. Therefore, the duration of a session was between five and seven minutes. On average, subjects rested for at least 10 minutes between sessions (without wearing the sports cap). For some volunteers, the experiment was completed in 2 days. The volunteers were six women aged 21 to 30 years and seven men aged 21 to 35 years (mean 27.00 ± 4.11) with head diameters of 52 to 61 cm (mean 55.54 ± 2.56) and with different hairstyles (straight, curly, and no hair). The experiments were carried out in an office, and the participants remained seated during the experiment. All participants signed an agreement following the policies of the university's committee for protecting human subjects and following the Declaration of Helsinki [171]. The experiment was video-recorded for a further confidential analysis. The observer and participant followed an ethical/hygienic protocol following the mandatory public health guidelines at the date of the experiment. The minimum number of valid sessions was 4 (Volunteers 3,4,8, and 10), and the typical case was five valid sessions. The training/testing scheme was defined as a 5-fold stratified cross-validation with leave one session out scheme [25], similar to the cross-session validation in [113]. Leaving one session out of cross-validation reinforces the robustness of the training against re-wearing of the system, as is common with wearable devices.

3.2.7. Results

Fig. 3.19 **A** shows the performance in the case of individual models (per-user) versus the ensemble model. The average improvement of the ensemble model was 14.30% (F1 score) compared to the best sensor-dependent model. The confusion matrix in Fig. 3.19 **B** shows an average F1 score of 85 % for the case of the individual concatenated models. The ensemble model for all (cross-user, see Fig. 3.19 **C**) yielded an F1-score of 79.00%. In addition, the one-ensemble model for thirteen participants achieved a 16.00% increment in the F1-score compared with the best sensor-type dependent NN (see Fig. 3.19 **C**).

Specifically, the F1 values of sensor-dependent NNs for thirteen participants were as follows: FSR+PEF = 51.00%, orientation = 58.00%, acceleration = 63.00%, and microphones = 44.00%. For subjects 2, 3, 6, 9, and 11, the ensemble model increased performance by up to 28.00% of the F1 result. In some cases, the ensemble model results were limited to the best sensor type

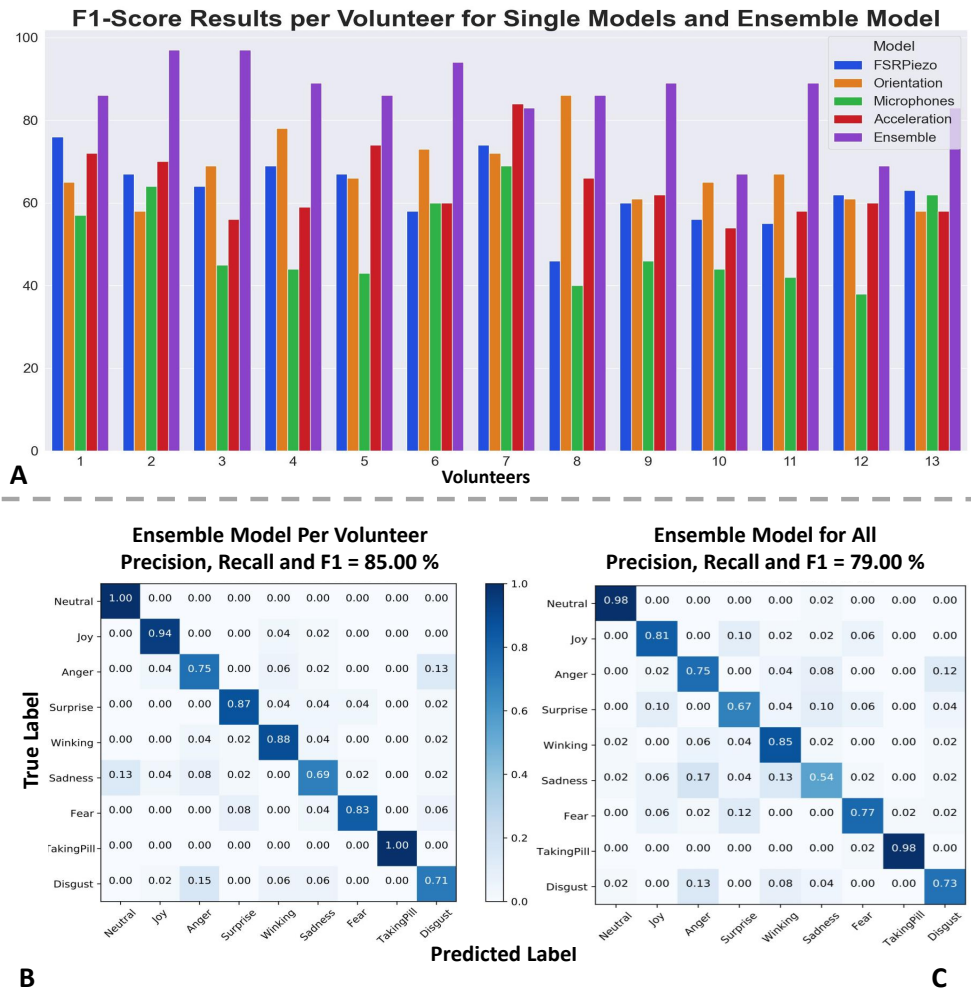


Figure 3.19.: Results Sensor Dependent and Ensemble Neural Networks **A** F1-Score of Sensor Dependent Neural Network per Volunteer Comparison. **B** Recall Results of Ensemble Model per Participant Avg F1=85.00%. **C** Recall Results One Ensemble Model for All Participants F1=79.00%

NN (inertial sensing) performance in participants (7 and 8), indicating that the ensemble model did not improve the performance in these particular cases.

In Fig. 3.19 C, the most evident misclassification is between sadness and anger with up to 17.00% confusion. This could be the consequence of the misinterpreted eyebrow movement while doing sadness; instead of eyebrows up, many volunteers moved their eyebrows down and the intensity of the gesture. A 13.00% confusion happens with the faces of anger and disgust, which are depicted as similar in Fig. 3.12. Another relevant misclassification (10.00%) occurs with surprise and joy gestures, which are faces distinguished by the mouth movement. The recognition value of 98.00% of the neutral face (null class) indicates the high specificity of the design, making null class recognition a suitable candidate for automatic data segmentation in real-time.

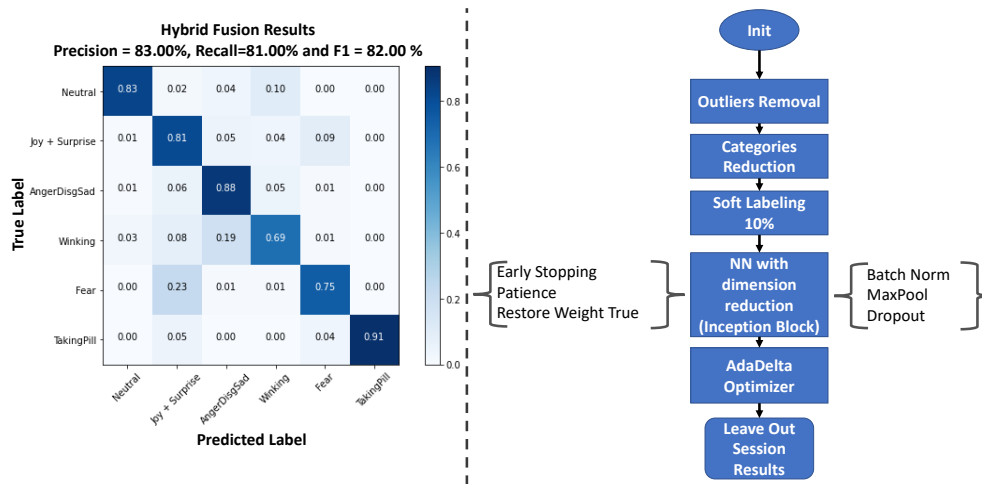


Figure 3.20.: *Left* Recall Results of Hybrid Fusion Model Leave Out Session for the Eight Best Imitators F1 = 82%. *Right* Data Analysis Techniques Applied in the Hybrid Fusion Modeling Pipeline.

The AMMG sensing modality has the weakest performance. Multiple reasons can lead to this result; the facial expression usually involves audio patterns in a natural setting, which might not be present in a mimicked experiment. The audio analysis in this work only considers the MFCCS with 13 filters to speed up the NN converging time but overlooks the spectral information in the entire spectrum. Due to the small dataset, more elaborated image processing techniques were not explored. In the future, it is highly recommended to do data augmentation and transfer learning to exploit the audio information completely.

Fig. 3.20 **Left** shows the performance of the hybrid fusion model using a leave-out session evaluation scheme. The result has an F1-score = 82 %, which is the average of five iterations/sessions. The main confusions are between the pairs; "Fear-Joy + Surprise" and "Winking-Anger + Disgust + Sadness", with 23 % and 19 %, respectively. It is relevant to notice that the confused classes have similar eyebrow movements. The first pair consists of expressions dominated by up-eyebrow movements and down-eyebrow movements mainly dominate the second. However, the confusion is in the categories with fewer instances. Due to the merged categories, the model can learn the distinction in the complementary classes with recall above 80.00 % for the case of "Joy + Surprise" and up to 88.00 % recall for the case of "Anger + Disgust + Sadness". Despite our small dataset, an F1 score over 80 % is an encouraging outcome. The outcome can be improved if more data is fed to the NN.

The result of the hybrid fusion is the output of the data analysis pipeline depicted in Fig. 3.20 **Right**. Thus, a direct comparison between late and hybrid fusion is impossible. Still, the low complexity and parallelism of the modular late fusion, one sensor-dependent NN per MCU, make it our preferable option to be used in embedded devices. Additionally, the hybrid fusion employs 2.35x more parameters than the early fusion to exploit the correla-

tions between the sensing modalities. In the future, the hybrid model results could be improved if more data were supplied to the NN. It is worth exploring NN partitioning techniques to solve the lack of parallelism in the hybrid fusion model. In [72], the authors proposed to leverage the power of multiple embedded/edge devices to run a large CNN. They proposed a framework that automatically partitions a CNN model into sub-models and generates the code for the execution of these sub-models on multiple edge devices (possibly heterogeneous) while supporting the exploitation of parallelism between and within edge devices.

3.2.8. Discussion

The experiment design tests the proposed approach’s feasibility in classifying facial movements. Notably, this study aims to recognize facial muscle pattern movements, and **emotion recognition is beyond our scope**. The selection of the dictionary of facial expressions is based on the Warsaw Set of Emotional Facial Expression Pictures[145], which is a database of high-quality photographs of genuine facial expressions. The photographs were taken after appropriate training, and participants (actors) were inclined to express felt emotions. In addition, the facial movements of taking a pill and winking an eye were added to the dictionary to extend the scope of recognition outside of the mimicked expressions.

The results above are based on a facial expression imitation experiment with participants from different cultures without acting experience. Although each participant interpreted facial activity in their own way, we found that complementary sensing modalities could detect patterns. Our results are minimally biased due to the high intrapersonal and interpersonal variability of imitated facial movements/expressions in a data set containing subjects from eight different cultural backgrounds. Furthermore, our results are inclusive due to a sex-balanced data set. Unfortunately, none of these characteristics were notable in the related works, mainly containing highly biased datasets of one cultural background and sex. A promising idea to improve the results will be to include a procedure for pre-training participants to close the gap between imitation and authentic facial expressions, as employed in [44].

Due to the design of our prototype, our solution has several limitations. One of them is that when using a sports cap as a wearable accessory, cheek and mouth movements are captured indirectly and without a complete map of the facial activity. The size of the custom-built electronics board and the chosen battery make the system heavy for real-life daily use. Downsizing the prototype board is inevitable to test the idea in realistic scenarios with natural facial expressions. The distributed sensing modalities in the frontalis and temporalis muscles may also lead to variations in MMG signal strength. A definitive statement about the best sensor modality could only be made if the placement of the sensors is the same, but this would be at the expense of user comfort, at least with our current hardware. A possible alternative is to stack (sandwich) all the sensors in a miniature circuit prototype; in this way, a quantitative comparison can be made between the sensing modality and the

type of activity.

On the other hand, based on our encouraging offline results, real-time implementation of the models is a reasonable next step. However, at the current stage, the system relies on segmenting the entire facial gesture (start/end point) to do the recognition. In the future, a plan to test the idea of online inference is to use the FSR gradient as a trigger for automatic data segmentation (unrest state recognition) and then proceed to inference.

In general, the sensing approaches and the modular technique can be applied to many different fields, such as psychology, to monitor students' learning process in classrooms [149]. Nevertheless, a clinician/cognitive expert will still need to redesign the experiment to adapt it to the psychological requirements. Overall, our design goal is to explore the system's feasibility in detecting facial muscle movements.

3.2.9. Conclusion

This work presents a privacy-preserving, low-memory, low-power consumption, and unobtrusive alternative system for facial expression detection. Although the design uses microphones, it only uses them to capture the sound of muscles while facial movements are performed, so no voice or ambient sound will affect the privacy requirement. Our system has demonstrated the ability to detect facial expressions using non-camera sensors mounted on a sports cap. In addition, the system works on participants of both sexes and from different cultures, demonstrating the inclusivity and generalizability of the approach. In the individual and cross-user results, the remounting of the sports cap was accountable; hence, our results are robust against everyday re-wearing in wearables. The results indicate that a multimodal approach based on the proposed sensors is well-suited for recognizing facial activities with an unobtrusive wearable sensor system.

Compared to the solutions based on our single sensing models, our ensemble approach provides a performance improvement of 16%. IMU and PMMG were the two sensing methods that contributed the most significantly to our results. An interesting question for future studies is how to extract more information from AMMG while maintaining a low-memory, privacy-friendly design. In particular, data augmentation and transfer learning methods could be used to understand the AMMG better. Also, it would be necessary to consider different sensors' placement around the head for a detailed comparison between sensing methods. One promising approach will be to design a miniature prototype as a stack of sensor units consisting of pressure sensors, piezoelectric sensors, a mechanical microphone, and an inertial measurement unit. The sensor stack would simplify the connection between the sensor to be selected and the types of muscle activities to be monitored.

Our results are from an offline analysis. Therefore, obstacles may appear to the system's deployment, such as environmental sounds, power consumption, mobile scenarios, and user comfort. In the future, a miniaturized version of the system is desirable to perform experiments with more people and in realistic (out-of-office) scenarios. The use of an entire gesture-instance data

processing technique has the advantage of lighter and faster models (due to its simplicity). At the same time, it makes the design dependent on automatic segmentation techniques to determine the start/end point of a facial gesture for the case of real-time gesture recognition. For automatic gesture segmentation, it is possible to select one or a combination of the following paths; the FSR pressure data gradient can be used as a trigger to detect a person's unrest state and then proceed to signal the start/stop of data collection to make the prediction. The FSR gradient option is also suitable for reducing power consumption. The solution can be combined with a weighted belief system and use the PEF gradient with the FSR; it is necessary to consider that the PEF is susceptible to the motion of the worn accessory, a sports cap, in our case. The FSR-PEF disturbance detection system can be the first step in a hierarchical procedure in which the FSR-PEF model is used to detect the null class (83.00 % recall for thirteen volunteers). Then the ensemble/fusion model is activated.

We employ seven validated facial expressions (Warsaw Photoset) in our work to compare our results with future solutions. Although a fair comparison with related work is negatively affected by many research papers using personalized facial/head activities, we believe our results are competitive with the state of the art. In the next section, we embedded the pressure mechanomyography and the inertial modalities into a glasses frame to monitor facial expressions and eating activities in real-time and on-the-Edge.

3.3. MeciFace: Mechanomyography and Inertial Fusion-based Glasses for Edge Real-Time Recognition of Facial and Eating Activities

3.3.1. Problem Statement

Facial expression recognition and eating monitoring technologies have become increasingly essential in understanding stress-related eating behaviors and their impact on overall health [137]. Wearable devices offer a convenient and non-intrusive solution to detect potential health issues, such as binge eating, stress-related overeating, and anorexia [156], and monitor these behaviors, catering to individuals with specific dietary restrictions, such as diabetic patients or those with food sensitivities. In addition, wearables can help individuals develop coping mechanisms to manage stress and maintain a healthy lifestyle.

The use of glasses for human activity recognition (HAR) in wearable technologies is widespread due to their ubiquity and strategic position in front of the user's face. In general, glasses-based wearables offer a comprehensive approach to capturing visual cues, tracking eye movements, integrating sensors, providing real-time information, and enabling hands-free use. These aspects contribute significantly to accurate and efficient activity recognition [127], [10, 105, 115, 128, 130, 208]. Commercial solutions, such as OCOsense (Emteq Labs), have already been introduced, utilizing optical-flow, inertial, pressure,

and microphone sensors to monitor facial expressions and emotions [66].

However, existing state-of-the-art glass-based wearables typically rely on external devices, such as computers, servers, or smartphones/tablets for real-time data processing and inference. The distribution of power consumption, latency, and memory can limit the efficiency of the system. Handling data across multiple devices can become a privacy and security concern. To address these limitations, in this paper, we introduce MeciFace, a real-time solution that performs facial activity recognition and eating/drinking gesture detection on-the-edge. By embedding data acquisition, signal processing, and inference within the MeciFace hardware, the system minimizes reliance on external devices.

The MeciFace system utilizes neural network models (NN) deployed on a microcontroller (MCU) using the TensorFlow Lite for microcontrollers framework. The proposed system fuses information from inertial and mechanomyography (MMG) sensors, ensuring privacy and low power consumption while achieving robust recognition performance [15]. The facial expression dictionary is defined in Fig. 3.21, in addition to the null/else class as static face [15, 21]. In particular, the gesture of taking a pill is included to differentiate eating/drinking episodes from the sporadic gesture of touching the face/mouth. The eating scenario-related classes are eating, drinking, and null to extend the work in [221].

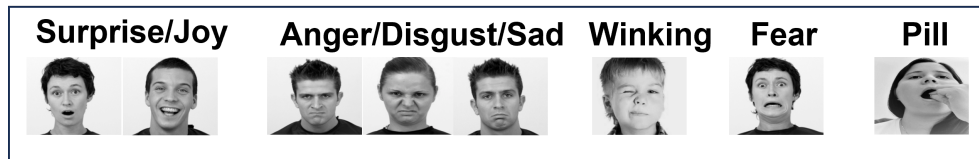


Figure 3.21.: Facial Muscle Activities Dictionary; 6 Facial Expressions from Warsaw Set of Emotional Facial Expression Photoset [145] and 2 Gestures from [15]. Taking a Pill Facial Muscle Movement is Included to Differentiate Eating/Drinking Episode with the Sporadic Gesture of Touching Face/Mouth.

3.3.2. Contributions

The main contributions of our approach can be summarized as follows:

- We present MeciFace, a state-of-the-art real-time solution for facial activity related to facial expressions and eating/drinking gestures that uses a fusion of mechanomyography and inertial sensing, providing flexibility, low power consumption, and cost-effectiveness, with potential future applications such as monitoring sporadic episodes of emotional eating.
- We employ lightweight neural network models to ensure a low memory footprint, providing an embedded and sustainable solution.
- We propose a hierarchical multimodal fusion to reduce energy consumption and increase robustness against the null class, in which the first

stage detects motions and recognizes a non-null facial gesture using an MMG model. Then, using an inertial model, the second stage recognizes the dictionary in Fig. 3.21.

- The hierarchical multimodal fusion is extended for the case of eating/drinking monitoring. The first stage discriminates between null and eating/drinking categories with an MMG model. The second stage employs an inertial model to classify between eating and drinking.
- Our work is a first step towards a ubiquitous system that monitors facial expressions and eating/drinking episodes to add contextual information from both scenarios, relevant to detecting stress-triggered eating episodes.

3.3.3. Apparatus

The MeciFace prototype is shown in Fig. 3.22A. The microcontroller is a QTPy ESP32 from Adarfruit. The MCU is an ESP32-S3-Dual-Core 240MHz Ten-silica with 8MB flash, 512KB SRAM, and Bluetooth low energy (BLE). The prototype includes an SD Card, which is used as a data logger. The sensors are an inertial measurement unit (IMU-BNO085), an atmospheric pressure, environmental and gas sensor (BME688), an analog microphone (SPH8878LR5H), a force-sensitive resistor (FSR), and a piezo-electric film (PEF) for MMG, see Table 3.3.

Table 3.3.: *MeciFace Sensors Characteristics*

Sensor	Vendor	Dimensions(cm)	Grams	Benefits
FSR	Alpha MF01A-N-221-A01	1.25 diameter	0.26	Ultra-thin/flexible
PEF	TE SDT1-028K shielded	4.45 x 1.97 x 0.32	0.30	Low noise/shielded/flexible
IMU	Bosch BNO085	0.38 x 0.52 x 0.11	0.15	Fused data, Auto calibration
Barometer	Bosch BME688	3.0 x 3.0 x 0.9 mm ³	0.3	Pressure and Gas Sensor with AI
Microphone	Knowles SPH8878LR5H	0.35 x 0.27 x 0.13	0.25	Low Noise and Omni-directional

In Fig. 3.22B, the block connections diagram is presented. The IMU sensor connects to the MCU via a serial peripheral interface (SPI) bus. The hardware includes a pressure and environmental sensor (BME688) and analog audio (SPH8878LR5H). The data from the BME688 and SPH8878LR5H is not used in this work. Still, the option of monitoring environmental and audio data makes our design extendable for future analysis. FSR and PEF information is transferred to the MCU using an inter-integrated circuit (I2C) with the intermediate assistance of ADS1015. The ADS1015 is an analog to I2C converter. Converting analog signals to digital makes the system robust to subtle movements/motion artifacts, and reduces the signal’s sensitivity to the distance between the sensor position and the MCU. Besides, it is also easier

to add slaves to an I2C bus compared to adding more analog channels to the system. The sampling rate of the sensors is around 50 Hz. The battery is a LiPo; 3.7V, 500mAh(3.5x3 cm).

The IMU is on the nose bridge of the glasses to mimic the position of the temporal muscle in [15]. The IMU position is suitable to capture head displacement, cheek movements, and symmetrically sense vibrations on the glasses frame. The FSR and PEF are on the temples (right/left) muscles also used in [15]. The temple position is relevant to monitor masseter muscle-related movements. Masseter's movements include chewing, swallowing, and tongue sweeping for the case of eating activities and smiling or getting angry for the facial expression scenario.

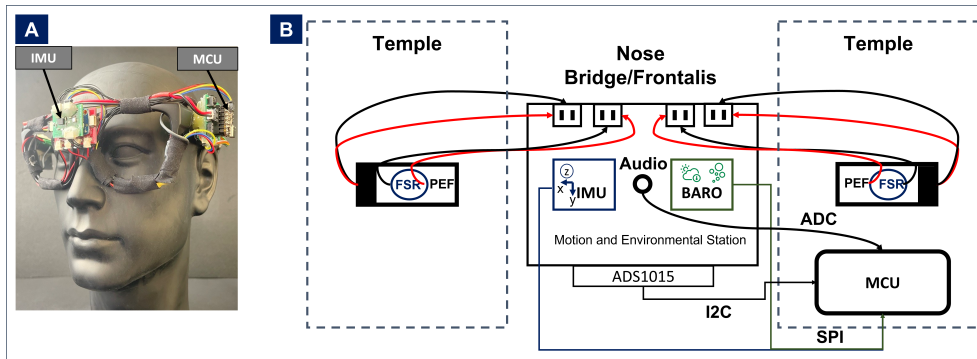


Figure 3.22.: MeciFace Prototype **A.** Hardware Connections Blocks: Motion and Environmental Station on The Glasses' Nose Bridge with BNO085 (IMU), SPH8878LR5H (Microphone) and BME688 (Barometer). On The Temples are The Force Sensitive Resistor (FSR), Piezoelectric Film (PEF), and QtPy ESP32 (MCU) **B.**

3.3.4. Multimodal Sensor Fusion

As shown in Fig. 3.23A and in Fig. 3.23B, two collaborative models were deployed for the facial and eating monitoring applications. The first neural network model (NN) is the FSR-Piezo (MMG-model) with four channels as input. Two FSR and two Piezo channels complete the four inputs of the MMG model. This model is used to distinguish the null class from activity detection. The null class includes activities such as; walking, talking, standing/sitting down, picking cutlery, and working on the PC, among others. The output of the MMG model served as a trigger for the second model, the inertial model. In the event of an activity being classified as non-null, the inertial model is activated. The second model fused inertial information, including acceleration and orientation, as seven input channels. Specifically, the input channels are linear accelerations (x, y, and z axes) and quaternions as orientation. For the case of facial expressions, the second model outputs are the classes in Fig. 3.21. For the case of eating monitoring, the inertial model returns eating/drinking classes. The hierarchical approach reduces the complexity of the models, leveraging the information fusion with lightweight NNs

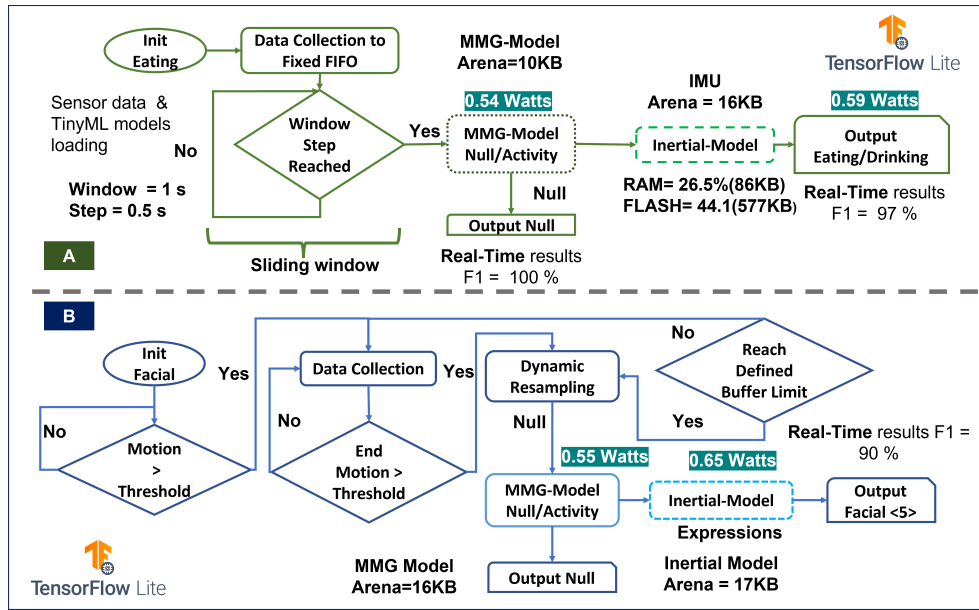


Figure 3.23.: Real-Time and on-the-Edge Flow Diagram Implementation for the Eating/Drinking Scenario with the Two Stages Hierarchical Modeling; First Stage is the Mechanomyography-based Model (MMG-Model) to Detect Null/Activity. The Second Stage is the Inertial-Model to Classify Eating and Drinking Episodes by Window Size of One Second and Window Step of Half a Second **A**. Real-time and on-the-Edge Flow Diagram Implementation for the Facial Expressions Scenario with Motion Threshold Detection and Two Stages Hierarchical Modeling; The First Stage is the MMG-Model to detect Null/Activity. The Second Stage is the Inertial-based Model to Classify the Facial Movements Dictionary in Fig. 3.21 **B**.

(11-19KB) to be deployed in tiny MCUs. Using the MMG-model as a trigger signal, the power consumption remains below or equal to 0.55 Watts.

The NN structure consists of a convolution (filters=3, kernel= 10, ReLu), a layer normalization, and batch normalization layers with max-pooling ((5,1)) and dropout (0.5), followed by a flattening layer, a fully connected (FC) layer of 10 and an FC with softmax. The NN optimizer is AdaDelta, with a learning rate of 0.9 and categorical cross-entropy (label smoothing 30%) as a loss function. The metric to monitor during training is a recall at a precision of 0.9. This NN structure is used in the MMG and inertial models for both applications (Expressions/Eating). The training ran for 200 epochs with early stopping (patience 30 and restoring weights). The number of parameters of our NNs is ≤ 3890 ; thus it is a lightweight design and less susceptible to overfitting. The NN models were trained using the TensorFlow/Keras 2.12.0 framework.

3.3.5. Experiment Design

Ethical Agreement All participants signed an informed consent following the Declaration of Helsinki. The ethical committee of Kaiserslautern Univer-

sity and the German Research Center for Artificial Intelligence have approved the study. Participation was entirely voluntary and could be withdrawn at any time. The participants did not receive any compensation for their participation. The subjects could deny answering questions if they feel uncomfortable in any way. There are no risks associated with this user study. Discomforts or inconveniences will be minor and are not likely to happen. All data provided in this user study will be treated confidentially, will be saved encrypted, and cannot be viewed by anyone outside this research project unless separate permission is signed to allow it. The data in this study will be subject to the General Data Protection Regulation (GDPR) of the European Union (EU) and treated in compliance with the GDPR.

Evaluation Eating/Drinking Scenario Two groups of volunteers were recruited for the evaluation, for a total of ten participants. The training group and the test group have the same size and sex composition, but the participants do not overlap. For the training group, five volunteers (three female and two male) participated in the eating monitoring experiment. The volunteers come from Germany, the Republic of Korea, China, and the United Kingdom, and range in age from 24 to 64 years old (mean 47). The participants consumed their lunch or dinner without any restriction in a natural setting during four separate sessions. It is important to note that participants were not forced to perform special activities or follow a script to ingest their food. Thus, the null activities are acquired in a natural/authentic setting. The four sessions per participant were recorded on different days, ensuring that our device was worn repeatedly. For the eating/drinking case, the offline evaluation scheme was 4-fold cross-validation with a leaving-one-session-out. In addition, another group of five participants (testing group) was recruited for the real-time and on-the-edge evaluation. Therefore, the real-time evaluation was performed with another five participants (three female and two male), whose data were not used during model training. For the RTE assessment, volunteers come from India, Poland, USA, Germany, and Venezuela, and range in age from 25 to 34 years (mean 28,6). With this methodology, our results are user-independent and sex-balanced, with high cultural variability.

Evaluation Facial Scenario For the facial scenario, one person mimicked (randomly-10 sessions) the dictionary in Fig. 3.21 while wearing the MeciFace. A 10-fold cross-validation with a leave session out scheme was used. The ten sessions were on different days. Each session has four random tries per expression. The facial experiment is an extension of the previous work in [15]. In [15], we fused MMG and inertial data to monitor facial expressions with a sports cap design and thirteen participants (offline evaluation). In this work, we focus on the real-time glasses-based idea implementation for a more ubiquitous/embedded solution.

3.3.6. Real Time and On-The-Edge Recognition

The real-time and on-the-edge flow diagram is presented in Fig. 3.23. The flow diagram is split into two specific applications; Eating and Facial Muscle Movement recognition.

Eating Scenario: TensorFlow Lite for MCU was used to generate the embedded version of the NN models. For RTE recognition, two algorithms were used. In Fig. 3.23A is the eating/drinking flow diagram. A sliding window of 2 seconds (100 samples) with a step size of 0.5s is used as an input data frame to the NNs. For the eating monitoring, the PC is 0.5489 Watts (only MMG-model), and when the inertial model is activated, the PC is 0.5988 Watts.

Facial Scenario: The Fig. 3.23B depicts the procedure for the facial expressions' case. The first step consists of movement detection (using acceleration), reducing power consumption by 16% (from 0.55 to 0.46 Watts). The movement detection is based on a threshold condition ruled by $\sum_{n=0}^5 = |a_x|_n + |a_y|_n + |a_z|_n$. The motion detection only applies to the facial scenario as depicted in Fig. 3.23B. Then, the data collection will run until no movement is detected. Therefore, the size of the input window is variable and depends on the duration of the detected movements. The NN input is fixed at 100-time samples, so dynamic resampling of the window size is necessary. After data collection, the data is resampled to 100 samples using the equation: $Y_i = (p*a_{index+1} + (NS-p)*a_{index})/NS-1$. Where NS =new sampling, OS =old sampling, $p = i * OS \% NS$, $index = i * (OS/NS)$ for $i \in (0, NS - 1)$. The resample's output is the input to the MMG model. In the case of *activity* \neq *Null*, the inertial model will output the recognized facial expression. The power consumption (PC) for the facial expression solution is 0.55 Watts and 0.65 Watts when MMG and inertial model are activated. ⁵

3.3.7. Results and Discussion

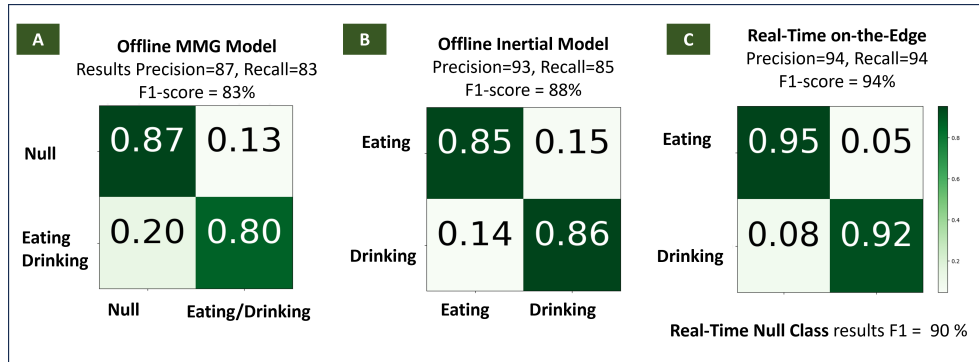


Figure 3.24.: Results of the offline MMG-Model with Five Volunteers (Leave-one-session-out cross-validation) in Lunch/Dinner Scenario; F1-score=83 % (A). Results of the offline Inertial-Model with Five Volunteers (Leave-one-session-out cross-validation) Lunch/Dinner Scenario; F1-score=88 % (B). Real-Time on-the-Edge Recognition Results for Five Unseen Volunteers (User-independent) in Snacking Scenario; F1-score = 94 % (C).

⁵We used the USB Digital Power Meter: available in <https://www.az-delivery.deenproductscharger-doktor> DLA: January 2, 2025

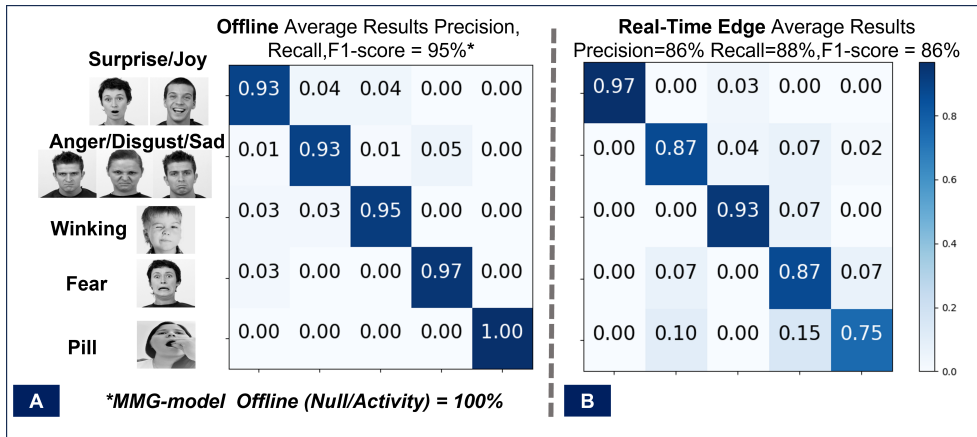


Figure 3.25.: Results of the offline Inertial-Model for Ten Sessions on Different Days with Leave-One-Session Out Cross Validation for the Recognition of the Dictionary in Fig. 3.21; Joy/Surprise(1), Anger/Disgust/Anger(2), Winking(3), Fear(4) and taking a pill(5) and F1-score=95%(A). Real-Time and on-the-Edge Results of the Inertial-Model for Three Sessions on Different Days for the Recognition of the Facial Activities in the Dictionary in Fig. 3.21; F1-score=86%(B).

The test group and the training group have the same size and sex distribution but with no overlap between participants. Fig. 3.24A-B shows the eating/drinking (5 persons, 4-fold cross-validation) offline results for the collaborative approach; MMG model (Null vs Activity) Fig. 3.24A and Inertial model (Eating vs Drinking) Fig. 3.24B. The results in Fig. 3.24A-B are from the training group (three females and two males) in a leave-one-session out cross-validation scheme. In Fig. 3.24C are the results for the real-time and on-the-edge evaluation, performed by the test group. For the RTE, the test group (three women and two men) ate and drank freely without any restrictions or instructions. The RTE for the eating/drinking monitoring F1-score was 94% with five additional subjects (test group), not within the training group. A total of ten participants performed the eating/drinking scenario evaluation. Hence, our approach could be generalized, but more unknown participants are still required.

In Fig. 3.24B, the offline results of the classification between eating and drinking instances with a score $F1 = 88\%$ are displayed. Instances are defined with a window size of 2 seconds and a step size of 0.5 seconds. Fig. 3.24C shows the online results of the NN network embedded in the glasses with a score $F1 = 94\%$. In the online evaluation, a majority voting system was used to classify the eating/drinking episodes. The voting buffer consists of five windows of size 2 seconds and step size 0.5 seconds, for a total inference time of about 4 seconds. Therefore, we can assume that the voting mechanism is the main reason for the improved performance of 6%.

A 10-fold cross-validation with a leave-one session out (10 sessions on different days) was performed to obtain the offline results in Fig. 3.25A. The real-time and on-the-edge results for the facial expression scenario (3 sessions,

different days, one volunteer) are shown in Fig. 3.25B. There was a reduction of 10% in the F1-score between the offline and the embedded solution. We believe this is due to errors in the motion detection algorithm in conjunction with the simplified (linear-based) dynamic resampling technique deployed in the prototype compared to the Fourier-based dynamic resampling of the training set.

The class of taking a pill degrades the most in performance between online and offline results, from 100% to 75%. In the offline results, the start and end of facial movements are manually annotated using the videotaped sessions as ground truth. In contrast, in online recognition, the start and end of facial movements depend on the motion detection algorithm. The motion detection algorithm is based on a threshold to determine the transition from a static to a moving state. Most classes are composed of a strong movement to make the facial expression, then a static period, and return to a neutral face with another strong movement. These steps are easily detected by the motion detection algorithm. But, the gesture of taking a pill is more complex in comparison. The activity of taking a pill involves a strong movement of the hand toward the face, followed by a slow movement of inserting the pill into the mouth, and ending with a strong movement of the hand coming back to rest. Thus, the gesture of taking a pill has a semi-static state (inserting the pill) compared to the other categories with a more defined static state. Therefore, we believe that the motion detection algorithm is the main cause of a reduction in the performance of 25% in the gesture of taking a pill. A future solution could be to deploy two different motion detection algorithms and make a probabilistic voting decision for classification.

The results (both scenarios) have an F1-score $\geq 86\%$, meaning that our approach holds promise for further development. This work has demonstrated the feasibility of using the MeciFace idea in two scenarios. Future work can focus on merging the two scenarios into one to obtain contextual information about users during sporadic eating activities. Besides, it would be meaningful to exploit the additional sensing information with the barometer/gas sensor and the microphone. For example, the gas sensor can detect volatile organic compounds (VOCs), volatile sulfur compounds (VSCs), and carbon monoxide, among other gases. Detecting VSCs is an indicator of bacteria growing. In [104], the authors use the VSCs in exhaled breath as a potential diagnostic method for oral cancer. On the other hand, the audio information can be used to detect sound related to emotions as a contextual source [21].

Integrating IMU sensors into smart glasses is not a challenge as they are already unobtrusively integrated into a commercial smart wearable like in [66]. The FSR and Piezo sensors are also straightforward to integrate due to their flexible design and minimal circuit requirement with only an analog to digital input constraint per sensor. Additionally, the mechanomyography sensing modality is completely passive compared to the IMU-based modality, so the power consumption is kept low.

Limitations Here we present a list of the limitations we have identified as well as future directions for optimizing the system:

Experiment Extension. The evaluation is considered preliminary with only one user for the facial and ten volunteers for the eating. Thus, an extended experiment setting is crucial to demonstrate the generality of the approach. The experiment should include variability in sex and culture to have fair training of the NNs to recognize facial activities as in [15].

Neural Network Tuning. The NNs were deployed using the TensorFlow Lite framework for MCU without any additional optimization technique. It is relevant to explore optimization approaches such as; quantization aware training, pruning, and quantization in the bits level, to improve the power consumption and the performance of the NNs. These optimization techniques are highly dependent on the selected embedded hardware. We leave this exploration for future work.

Miniaturization. The hardware in this work is a fast prototype, but the selected sensors have reduced dimensions, as shown in Table 3.3, which could be fully embedded in the glasses frame to improve comfort.

Human Feedback. After all the above limitations are addressed, it is crucial to do a human study to expose the weaknesses of the design and tune it to include user perception.

3.3.8. Conclusion

In this section, we proposed MecifFace, an innovative energy-efficient wearable system for real-time facial and eating activity recognition. By leveraging a glass frame as the wearable accessory, we strategically deployed sensors and the microcontroller, ensuring minimal intrusion into the user's daily life. The fusion of mechanomyography and inertial information on eyeglass temples and nose bridges allowed for comprehensive monitoring of facial expressions and eating activities. In the experimental results, we demonstrated the performance of MecifFace in real-time and on-the-edge, achieving an F1 score of $\geq 86\%$ for the facial expressions scenario and an F1 score of 94% for the eating scenario with a test group of five volunteers (user-independent case). Two groups of volunteers were recruited for the evaluation. The training group and the test group have the same size and sex composition, but the participants do not overlap. For the training group, five volunteers (three female and two male) participated in the eating monitoring experiment in a natural setting during lunch/dinner. The second group of five volunteers (three women and two men) were recruited for the real-time and on-the-edge evaluation of the eating monitoring case. Hence, with this methodology, our results are user-independent and sex-balanced, with high variability across cultures.

The hierarchical scheme implemented in the system significantly reduced power consumption, maintaining it below 0.55 Watts, thus enhancing the wearability and practicality of the device.

The TensorFlow Lite for Microcontroller framework enabled the seamless deployment of neural network-based models in their embedded versions. Techniques such as quantization and pruning can contribute further to memory reduction and efficient utilization of embedded resources, ensuring the system's

3.3. MeciFace: Mechanomyography and Inertial Fusion-based Glasses for Edge Real-Time Recognition of Facial and Eating Activities

sustainability. MeciFace can be easily extended to include contextual information from the environment, thanks to the incorporation of barometer/gas sensors and a microphone on the glasses' nose bridge. This extension enhances the potential of the system to detect stress-triggered eating episodes and offers a holistic approach to monitoring emotional eating behaviors.

Chapter 4

Body Posture and Gestures Recognition with Multipositional Capacitive Fusion

The author of this thesis has published the content, figures, and tables included in this chapter in the following publications:

Bello, H., Zhou, B., Suh, S., & Lukowicz, P. (2021, September). Mocupaci: Posture and gesture detection in loose garments using textile cables as capacitive antennas. In Proceedings of the 2021 ACM International Symposium on Wearable Computers (pp. 78-83).

Bello, H., Zhou, B., Suh, S., Sanchez Marin, L. A., & Lukowicz, P. (2022). Move with the theremin: Body posture and gesture recognition using the theremin in loose-garment with embedded textile cables as antennas. *Frontiers in Computer Science*, 4, 915280. **Journal**

Contents

4.1. Problem Statement	90
4.2. Contributions	91
4.3. Electronic and garment prototype	91
4.4. Experiment Design	93
4.4.1. General Dictionary Experiment	93
4.4.2. Dance Movements Experiment	94
4.5. Signal and Data Processing	95
4.5.1. General Dictionary Experiment Evaluation	95
4.5.2. Dance Movements Experiment Evaluation	96
4.6. Results	98
4.6.1. General Dictionary Experiment Results	98
4.6.2. Dance Movements Experiment Results	99
4.7. Discussion	101
4.7.1. General Dictionary Experiment Discussion	101

4.7.2. Dance Movements Experiment Discussion	102
4.8. Conclusion	103

4.1. Problem Statement

Human activity recognition (HAR) is an umbrella term that gives shelter to various specific applications to understand human behavior. Body postures and gestures (BPG) recognition are an essential piece of HAR. The popularity of BPG recognition is well earned due to the ability to describe human activities by changing postures or by detecting specific gestures [51]. BPG detection could lead to the generation of emotion and personality profiles [92, 144], to understand implicit social interactions [62, 71], to aid in sign language communication [56], and to predict people’s intentions [162].

Many wearables sensing applications have found their purpose in BPG, delivering highly developed solutions such as commercial motion capture systems [166]. The commercial and research markets for BPG recognition are mainly dominated by inertial measurement units (IMU) wearable-based techniques [32, 74, 164], and on the textile side by stretch or pressure sensors [37]. Most current solutions for BPG recognition have a common baseline requirement: the sensors need to be firmly attached to the body using tight garments or dedicated accessories, such as bracelets and straps. Therefore, we could argue that a reliable method for BPG recognition with loose garments remains a largely open problem. This work proposes a loose garment solution based on non-contact capacitive sensing with off-the-shelf components.

We present a novel intelligent garment design approach for body posture/gesture detection in the form of a loose-fitting blazer prototype, “the MoCa-Blazer”. The main component of our system is a modified electronic musical instrument, the theremin [180] for BPG recognition. The well-known musical instrument usually consists of one or two long metal rod/loop antennas emitting sub-MHz frequencies. As the thereminist moves inside the antennas’ range, volume and pitch can be controlled by his/her hand’s position. The theremin antennas are metallic, but any conductive wire/textile can be used as an antenna due to its intrinsic capacitive sensing. We substituted the metal rod with soft wires and integrated them inside a loose-fitting garment. The use of soft textile antennas as the sensing element allows flexible garment design and seamless tech-garment integration for the specific structure of different clothes. Our novel approach is evaluated through two experiments involving defined movements. First, the system is evaluated in the recognition of 20 arm/torso gestures. Secondly, a set of dance movements is evaluated to demonstrate the potential use case of our design as a sophisticated/elegant game controller. The inspiration came from the Nintendo Wii Rayman Raving Rabbids (®): TV Party- ShakeTV [201]

4.2. Contributions

Distinctive aspects in our design are a discrete gesture dictionary and the antennas move with the wearer’s body motion, consequentially changing the signal. Our **contributions** include:

- Presenting a wearable approach for detecting BPG that does not require sensors to be firmly fixed to the body or integrated into a tight-fitting garment. Instead, sensing is incorporated into a loose-fitting garment.
- Implementing a prototype, ”MoCaBlazer” that adapts the famous theremin musical instrument [64] as a sensor merged into a loose man’s jacket by integrating and modifying off-the-shelf components.
- Evaluating the proposed approach with the MoCaBlazer with 14 diverse participants in an experiment to detect 20 body postures and gestures.
- Applying several deep neural network models from the wearable HAR domain to the collected data, demonstrating accuracy of 86.25% for the leave-person-out (LPO) case and up to 97.18 % for the leave-recording-out (LRO) scenario.
- Fusing multipositional capacitive sensing with Radio Frequency Identification (RFID) synchronization for real-time and wireless recognition for one participant and six classes of a dance movements dictionary, obtaining an f1-score = 82 %. Hence, our ”MoCaBlazer” could be a promising alternative for an elegant/sophisticated game controller.

4.3. Electronic and garment prototype

The principal component in our electronic garment prototype is an off-the-shelf electronic musical instrument, ”The OpenTheremin V3” [64].¹ The theremin produces musical notes based on the frequency fluctuation of its antennas caused by the proximity of a person’s hands. In a theremin, we could find two antennas, one for volume (loop antenna) and another for pitch control (rod antenna) [180]. Capacitive sensing is the physical principle governing the behavior of the theremin. The human body could be modeled as a capacitor plate virtually connected to the earth and, in conjunction with the theremin’s antennas (second plate), completes a capacitor [178]. Thus, human proximity changes the effective capacitance of the Clapp LC oscillator in Figure Fig. 4.1 D, affecting its frequency. Therefore, we could infer that relative differences between body parts and theremin’s antennas could be used to distinguish body postures. In the present work, the pitch and volume antennas were embedded in a tailored garment (men’s blazer); thus, the person’s body moves with the theremin and ”makes music” with different postures and gestures (frequency profiles).

¹The OpenTheremin V3 has been updated to OpenTheremin V4: <https://www.gaudi.ch/OpenTheremin/>

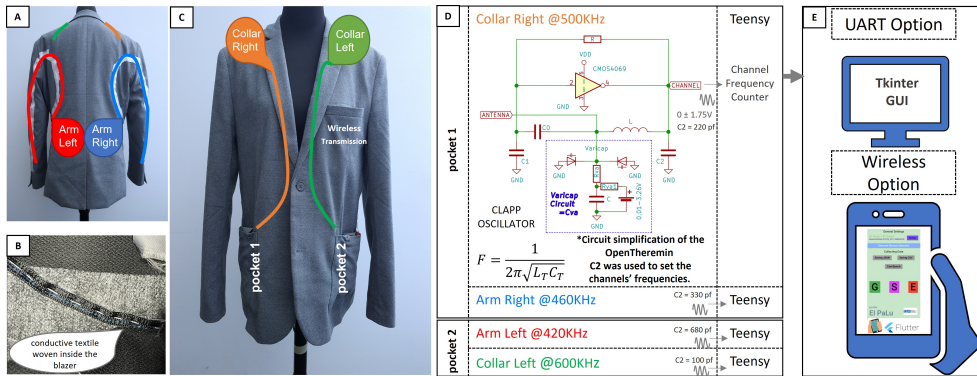


Figure 4.1.: Electronic Garment Design "the MoCaBlazer", **A** Is The Back Part of The Blazer, **B** Are The Textile Cables Sewn Inside The Garment, **C** Is The Front Part of The Blazer, **D** Is The Circuit Simplification Design of The Clapp Oscillator with The Antennas, **E** Are The Two Options for Collecting Data Coming From The Blazer; with a UART-Wired Option, and a Bluetooth-Based Android Application (Flutter framework) as The Wireless Option.

To test our approach we designed a prototype, the "MoCaBlazer", as shown in Figure Fig. 4.1. We employed a Tom Tailor®L/52 size blazer (best suited for 184 cm tall persons). In Figure Fig. 4.1 **A** and **C** the positions and patterns of our four antennas are depicted. The antennas cover the chest, a small part of the shoulders, the arms, and the back, as seen in Figure Fig. 4.1 **A**. This setting was appropriate for detecting upper-body postures and gestures without altering the tailored garment's main structure or hindering the wearer's motion.

The back antennas (standard 28 AWG cables) [8] in 4.1 **A** (Arm-Left, Arm-Right) start from the side pockets and, following a curving pattern (simulating a volume antenna), pass over the latissimus dorsi muscles toward the deltoids; they then turn sharply to go along the outer sleeve lines and terminate before the cuff buttons. The front antennas (TWC24004B textile cables) [199] in 4.1 **C** (Collar-Left, Collar-Right) were sewn inside the lining without modifying the structural design of the blazer (see Figure 4.1 **B**).² The Collar-Left and Collar-Right antennas were arranged to simulate a theremin's pitch antenna as closely as possible. Thus, they begin on the side pockets and go to the front-top button, then turn to align with the inner crease of the lapels and reach the notch; consecutively lead out of the crease and climb around the shoulder to the back, and end at the middle edge of the shoulder pad. The antennas' lengths are 80 cm (front) and 100 cm (back) for this particular blazer size (L/52).

Two "OpenTheremin" boards were inside the side pockets of the "MoCaBlazer" (see Figure Fig. 4.1) to handle four channels. The channels frequencies

²TWC24004B textile cables are deprecated, for an alternative option: Interactive Wear <http://www.interactive-wear.com/>

were modified by changing the capacitor (C2) in the clap-oscillator circuit to minimize cross-talk between them, as depicted in Figure Fig. 4.1 **D**. Then, the channels were sampled (frequency-count [184]) at 100 Hz by the Teensy®4.1 [185] development board.

Two options are available for the data collection: a UART serial (115200 Baud rate) as a wired option and a Bluetooth serial (9600 baud rate) as a wireless option. In the case of the wired alternative, the data is received by the serial port (USB) in a computer. The computer runs a Python script with a graphical user interface (GUI) developed using Tkinter [120], as depicted in Fig. 4.1 **E** upper element. For the wireless option, the data of the four channels is sent using the Huzzah-ESP32 Bluetooth serial protocol [58] (in the upper pocket) to a smartphone. The smartphone runs an android application, developed using Flutter framework [140], as shown in Fig. 4.1 **E** lower element.

4.4. Experiment Design

Two experiments were conducted with our garment prototype, the "MoCaBlazer". The experiments were carried out in an office without user calibration, i.e., without tuning the antennas' base frequencies to reduce the impact of different body capacitances. Inside the office, there were a few metal objects nearby, which are known to affect capacitive sensing [147]. All participants signed an agreement following the policies of the university's committee for the protection of human subjects and following the Declaration of Helsinki. The experiment was video-recorded for a further confidential analysis. The observer and participant followed an ethical/hygienic protocol following the mandatory public health guidelines at the date of the experiment.

The first experiment scenario was based on a general dictionary of posture and gestures in Figure Fig. 4.2. The second one was inspired by dance movements from the Rayman Raving Rabbids: TV Party-Nintendo Wii® as depicted in Figure Fig. 4.3.

4.4.1. General Dictionary Experiment

To study the flexibility of our system to adapt to an abroad type of gestures, a general dictionary of 20 upper-body postures/gestures was defined, see Figure 4.2. Fourteen participants mimicked the postures defined in the dictionary in a random sequence per session while wearing the unbuttoned "MoCaBlazer". The "MoCaBlazer" is based on a size L/52 blazer (Tom Tailor®), a recommended size for 184 cm tall persons. All participants performed five sessions. One session consisted of four random appearances of each gesture inside the dictionary, giving 400 instances per volunteer. The starting and ending point of a gesture was marked by the null position (standing position). On average, the volunteer's resting period was at least 20 minutes (without wearing the blazer) in between sessions. For some volunteers, the experiment was completed in two days. The volunteers were seven women, 24-64 years old and 157-183 cm in height; and seven men, 25-34 years old, and 178-183 cm in

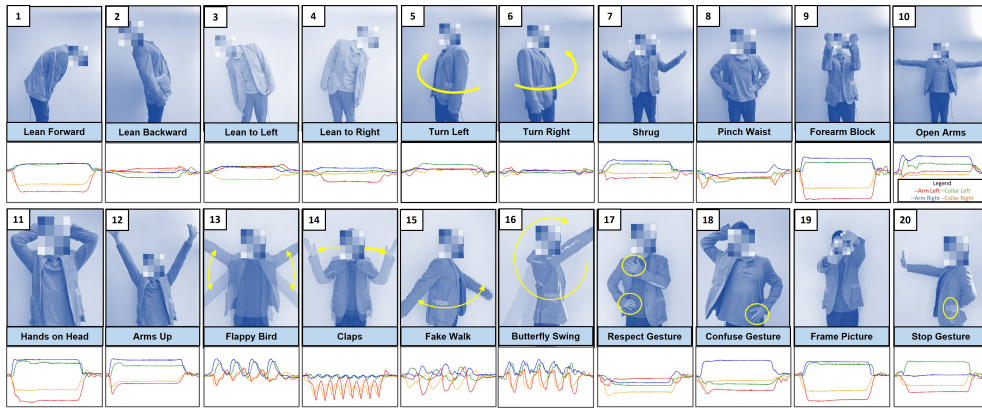


Figure 4.2.: Twenty General Upper-Body Gestures/Postures Dictionary with Example Signals. $x = (0,400)$ Time Steps, y :Norm.

height.

4.4.2. Dance Movements Experiment

As an application-specific experiment, a dance movements dictionary containing the eight postures depicted in Figure 4.3 was defined. It is essential to highlight that the data transmission from the "MoCaBlazer" for this experiment was wireless. Therefore the capacitive channels were floating (not connected to the ground). The dance movements were selected from the game Rayman Raving Rabbids: TV Party-Nintendo Wii® to test the feasibility of using the system as a sophisticated game controller. Three volunteers were asked to imitate the eight movements using the buttoned "MoCaBlazer". Three sessions were recorded per volunteer; each session contained five random appearances per gesture inside the dictionary for a total of 120 instances per participant. The volunteers were asked to rest (without wearing the blazer) for at least 10 minutes in between sessions. The participants were two men and one woman, 26-30 years old and 160-183 cm in height.

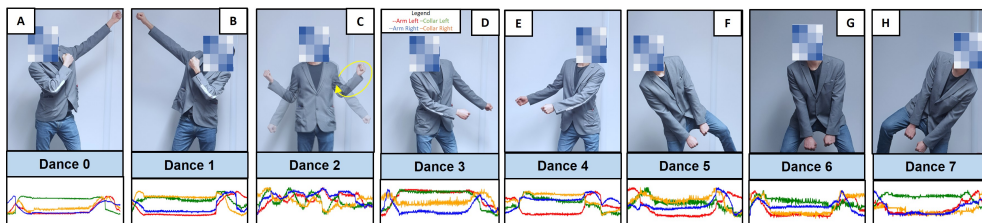


Figure 4.3.: Eight Dance Movements Dictionary with Example Signals. $x = (0,400)$ Time Steps, y :Norm(0,1).

4.5. Signal and Data Processing

As shown in Figure 4.1, the Clapp oscillators generated four data channels. The wearer’s movements alter the channels’ fundamental frequency. The channels’ data is processed as a time sequence. The granularity of the evaluation was a complete gesture/instance. An instance was completed when it included a change from the standing position (starting point) and a return to the standing position (ending point). Furthermore, the impact of common and subtle disturbances on the four capacitive channels was reduced by normalizing the gesture/posture. The digital signal processing was slightly different for the two types of experiments. The videos of both experiments were used as ground truth in a manual labeling procedure.

4.5.1. General Dictionary Experiment Evaluation

The fundamental frequencies of the channels could be seen as a bias difference between the four channels. A normalization procedure was performed to remove these biases and reduce the capacitive sensing modality reliance on the ground. The normalization consisted of subtracting the average of the gesture’s first (starting point) and last values (ending point). Then, the normalized four channels’ time sequences of each posture/gesture were fed to a fourth-order Butterworth band-pass filter with pass frequencies between 1 Hz to 10 Hz. The duration of gestures performed was not constant, which led to variations in the number of samples per instance. The average duration of a gesture was around 2 seconds (200 samples at 100Hz). A window of 4 seconds (400 samples at 100Hz) was selected to guarantee the activity’s capture. The signals were dilated or contracted depending on whether the gesture contained less or more than 400 samples. Due to the dynamic nature of the applied resampling procedure (dilation or contraction), this is called time-warping [69]. The signals dynamically resampled (upsampled or downsampled) to 400-time steps provided a fixed-size input for the neural network. The time-warping process was based on the Fourier method [106] implemented in the SciPy library [193]. The normalization procedure forced the gesture to start and end circa the same value. Hence, the Fourier method was employed without a window function, which is a method customarily used to avoid ringing artifacts.

A total data of 5600 gestures/instances of the dictionary in Figure Figure 4.2 (fourteen participants) were processed.

Deep Learning Model

Deep learning models such as 1D-LeNet5 [109, 182], DeepConvLSTM [146], and Conv2D [94, 95, 174] were evaluated. The best trade-off between performance, parameters, and training time was obtained from a modified 1D-LeNet5 model (see Table 4.1). The modified 1D-LeNet5 was defined as a convolution (conv) - max pooling (maxpool)-conv-maxpool-conv- fully connected (fc)-fc-softmax layers with batch normalization [86] and dropout [183] on the convolution layers.

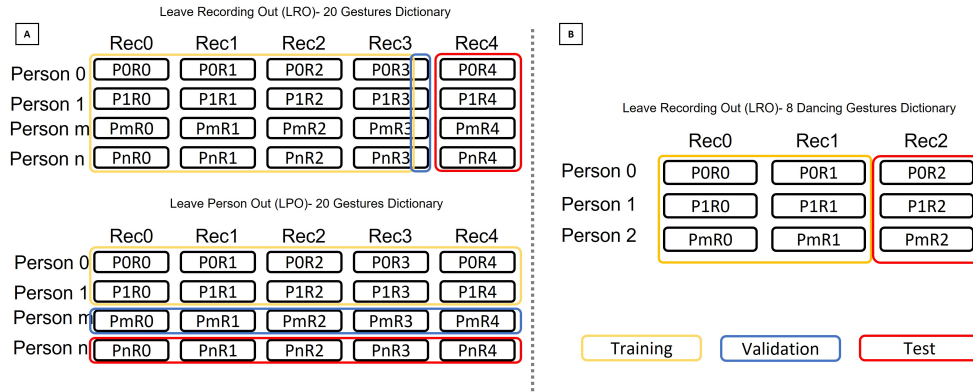


Figure 4.4.: Data Partition Scheme to Train and Test the Deep Learning Models. **A** Shows the Leave-Recording Out (LRO) and Leave-Person Out (LPO) Paradigms Used for the Data of the Twenty General Postures. **B** Shows the Leave-Recording Out (LRO) Scheme Employed for the Data of the Eight Dance Movements.

Leave-recording out (LRO) and Leave-person out (LPO) schemes were used as depicted in Figure 4.4 **A**. The LRO paradigm studies the method’s performance for a known group of people, while LPO evaluates the model’s performance in the case of unknown persons. We ran all the person’s permutations or recording combinations within each run and summarized the confusion matrix together. That means a complete run of LRO has 5 and LPO has 14×13 train-valid-test cycles. The number of epochs used was 500, stopping when there were signs of overfitting. The three convolution layers are used with a kernel size of 41 and the activation function of ReLU. For max pooling, the pool size was (40, 40) for the first convolution (400, 40) and (4, 40) for the second convolution (40, 40). The third convolution was of size (4, 40) without pooling. A flattening layer of 160 was followed by a fully connected layer of 100. The twenty outputs for the different activities in Figure 4.2 are then converted into probabilities by a fully connected layer and softmax function. The categorical cross-entropy loss function and Adam optimizer [97] were used in the optimization of the neural network.

4.5.2. Dance Movements Experiment Evaluation

In this experiment, the time sequences of the four channels were resampled/time-warped to 400-time steps using the same methodology as above. The signals were normalized between 0 and 1, as $x_{norm} = \frac{x - \min(X)}{\max(X) - \min(X)}$. Where x is a one-time step, X is a sequence of 400-time steps, and x_{norm} is the normalized time step.

In total four deep learning models were generated.³ Three individual models per volunteer were trained; two sessions from the same person were used as

³The deep learning framework was TensorFlow version 2.8.0 [2, 124] and Keras version 2.8.0 [45] in Google Colab environment [27].

training, and the third session was for testing. Moreover, a fourth model was developed using two sessions from each participant (three in total) as training and the third session for testing as shown in Figure Fig. 4.4 **B**. A total data of 360 gestures of the dictionary in Figure Fig. 4.3 (three participants) were fed into a one-dimension convolutional neural network as shown in Figure 4.5. The neural network's input layer was a time series of 400 samples per four channels/antennas (400,4,1). Two convolutional layers followed this with a max-pooling of 10, batch normalization, and dropout of 20 %. A third convolutional layer was added but without max pooling. Next, a flattening layer of 160 was followed by a fully connected layer of 100. The eight outputs from the different activities in Figure 4.3 were converted to probabilities using a fully connected layer and a softmax function. The training consisted of 500 epochs for all the models. The optimization of the neural network used the categorical cross-entropy loss function and stochastic gradient descent (SGD) [160] optimizer with learning rate=0.005 and momentum=0.001.

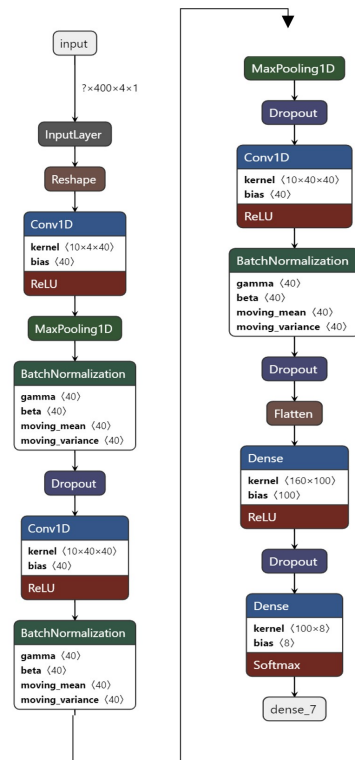


Figure 4.5.: Structure of the 1DConv Neural Network Model Used for The Data of The Eight Dance Movements. Input Shape(time-Steps,Channels,1) = (400,4,1) and Output Shape = 8 Classes.

Real-Time recognition with RFID Synchronization

Following the training and testing paradigms in Fig. 4.4 **B** a group model was built for the three participants in the dance experiment. The resulting model considered an entire gesture when the person follows the sequence; standing-gesture-standing. So, this sequence needs to be matched to do a real-time evaluation. We proposed to use Radio Frequency Identification (RFID) as a synchronization technique to signal the starting and ending points of the gesture. RFID synchronization was employed in the calibration of atmospheric pressure sensors to estimate the vertical position of the hand in [16]. The RFID system comprehends two parts; the reader and the tag. The most commonly used extension of RFID is near-field communication (NFC), which is available in most smartphones to make over-the-air payments. In [16] the reader was on the wrist and the tag was around the pocket to simulate the NFC systems.

It should be noted that there is already an NFC system in our smartphones and that the pocket is a common position to carry our phones. In addition, RFID stickers are nowadays a commonly used solution for tracking merchandise in stores in a ubiquitous and unobstructed manner. Hence, we propose a setting for the real-time evaluation as the one shown in Figure Fig. 4.6 **A**. The wrist was the selected position for the reader, and the side pocket of the "MoCaBlazer" was the position for the RFID tag (Mifare Classic 13.56 MHz). Figure Fig. 4.6 **B** shows a volunteer wearing the synchronization system. The RFID signal and the "MoCaBlazer" four-channel outputs were sent using Bluetooth serial (wireless) to a Python script running the TensorFlow model. The Python script follows the flow diagram in Figure Fig. 4.6 **C**. The real-time evaluation was performed with participant number two of the three participants pool. The participant was asked to do five repetitions per dance gesture (40 motions).

It is worth mentioning that the real-time recognition with RFID synchronization did not include any pre-training stage with the RFID signal. The model used here was generated from the offline data without RFID. The input data to the offline model was manually labeled with a granularity of 50 fps (recorded video).

4.6. Results

4.6.1. General Dictionary Experiment Results

In Table 4.1 the results for the three models; 1D-LeNet5, DeepConvLSTM, and Conv2D are compared. There is not a remarkable variation across the models. The confusion matrices using Conv2D for the Leave-recording out (LRO) and for Leave-person out (LPO) are depicted in Fig. 4.7. The results confirmed a robust recognition of the 20 postures/gestures dictionary. The LRO or user-dependent case gave an average accuracy of 95%, see Figure Fig. 4.7 **A**. There was a decrease of around 10% for the LPO or user-independent case, shown in Figure Fig. 4.7 **B**. Furthermore, we achieved an average accuracy of 86.25 %, with nine classes out of the 20 returning above 95 % accuracy. Hence, we could



Figure 4.6.: Real-Time Recognition System. **A** The "MoCaBlazer" with the RFID Reader-Tag Pair Positions. **B** Volunteer Wearing the "MoCaBlazer" with the RFID Synchronization System. **C** Flow Diagram of the Real-Time Recognition Python Script.

Table 4.1.: Comparison Results for the General 20 Body Postures and Gestures Dictionary (in %) with Various Models

Method	Accuracy (LRO)	Accuracy (LPO)	Parameters	Training Time
1D-LeNet5	96.86±0.46	85.34±7.83	152,880	1.00x
DeepConvLSTM	94.11±0.82	85.42±5.84	440,852	2.32x
Conv2D	97.18±0.70	86.25±8.09	584,800	0.86x

^aLRO: leave recording out, LPO: leave person out.

^bThe accuracy numbers are represented as *mean ± std*, the standard deviation is from within each complete cross-validation.

^c1.00x Training time of 50 minutes as the baseline of complete LRO on NVidia RTX A6000 with the TensorFlow framework.

conclude that these results are good enough to consider that our model will perform well for the stranger case; people not included in its training phase.

4.6.2. Dance Movements Experiment Results

Four models were generated using the neural network structure in Figure Fig. 4.5. The results for the three individual models are shown in the confusion matrices in Figure Fig. 4.8. Fig. 4.8 **A** presents the recognition for the first model trained (2 sessions) and tested (1 session) with the data from volunteer number one. The first participant obtained the lowest performance, f1-score

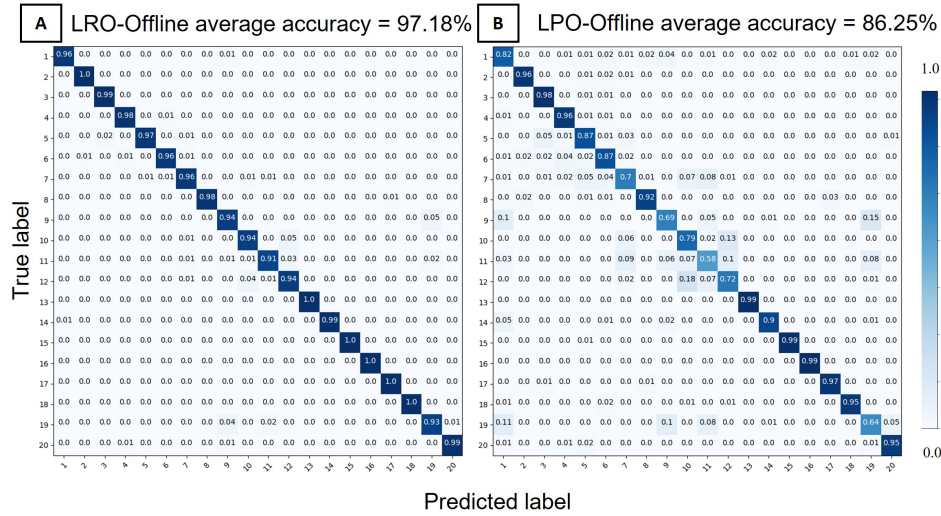


Figure 4.7.: Confusion Matrices for the Data of the Twenty General Gesture Dictionary. **A** Results for the Leave-Recording Out (LRO) Scheme. **B** Results for the Leave-Person Out (LPO) Scheme.

= 93 %. Fig. 4.8 **B** depicts the result for Leave-one recording out (LRO) of the second volunteer, showing an f1-score = 100 %. For the third participant, the results are only 5% less than the perfect f1-score. With this performance, our design successfully recognized the gesture dictionary in Figure Fig. 4.3.

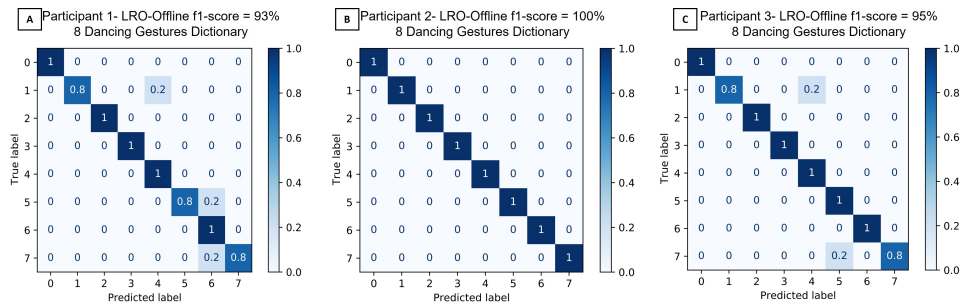


Figure 4.8.: Individual Models Confusion Matrices for the Data of the Eight Dance Movements Dictionary. **A** Results for the Leave-Recording Out (LRO) Scheme for Participant One. **B** Results for the Leave-Recording Out (LRO) Scheme for Participant Two. **C** Results for the Leave-Recording Out (LRO) Scheme for Participant Three.

The data partition (train and test) of the fourth model is in 4.4 **B**, and the result is illustrated in 4.9 **A** with a f1-score = 92 %. The fourth model was tested in real-time in conjunction with RFID synchronization and gave an f1-score = 82 % as shown in 4.9 **B** for 6 classes. In the confusion matrix in 4.9 **B**, the classes 4-5 and the classes 6-7 were merged, which gives a total of 6 classes. In the case of merged classes 4-5, the fourth class was completely

confused, with half of its instances being recognized in class number 1 and the other half in class number 5. Moreover, in the case of the merged classes 6-7, the seventh class was recognized consistently as class number 6. The above indicates that the dance movements 4 and 7 in Fig. 4.3 could not be recognized correctly with the combination of the fourth deep learning model (offline) and the RFID synchronization (online). Despite the negative cases of classes 4 and 7, the real-time recognition with RFID synchronization shows decent performance for the merged classes (6 classes in total).

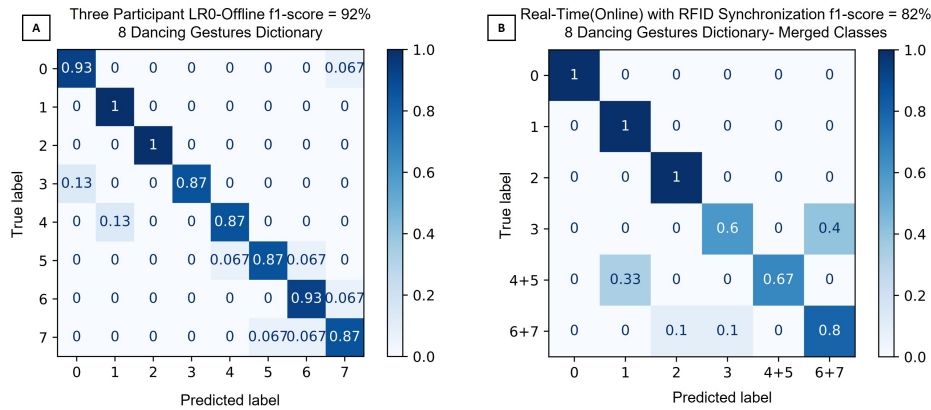


Figure 4.9.: Group Model Evaluation Results. **A** Confusion Matrix for the Offline Test Results with Leave-Recording Out (LRO) for Three Participants. **B** Confusion Matrix for the Online Results Using the RFID Synchronization Method for One Volunteer.

4.7. Discussion

4.7.1. General Dictionary Experiment Discussion

To discuss our results the confusion matrices in Fig. 4.7 and the 20 gesture/-posture dictionary in Figure 4.2 will be referenced as a duo. In the case of the Leave-recording out results in Fig. 4.7 **A**, the accuracy was above 90 % for the 20 classes. On the other hand, in Fig. 4.7 **B** the result for the Leave-person out scheme is depicted, and we could observe several pairs of false recognition. For the pairs of arms-up (Gesture 12) / open-arms (10) and forearms-block (9) / frame-picture (19), the arm motions and directions are physically similar. For the case of lean-forward (1) / frame-picture (19), the similarity is seen in the signals in Figure Fig. 4.2; we believe it is a negative effect of participants of different body shapes wearing the same size blazer L/52, which leads to misclassification of 11 %. Nonetheless, for forearms-block (9) / hands-on-head(11) pair with similar signal patterns and elbow flexion, the misclassification is only 5 %. It is worth noticing that the activities with shoulder motion, such as shrug (7), forearms-block (9), hands-on head (11), arms-up(12), and frame-picture(19), have a reduction in accuracy in the Leave-person out (LPO)

result compared to the Leave-recording out (LRO) case. The confusion could be due to the lack of antennas to cover the shoulders of the "MoCaBlazer" and that all fourteen volunteers (of different body shapes) were wearing the same one-size blazer.

4.7.2. Dance Movements Experiment Discussion

The result of the individual model of participant number one shows some misclassification (see Fig. 4.8 A). For the classes/dance movements 1 and 4, 20 % of the gestures are confused; these two gestures have in common that the arms move to the same side of the body trunk but at a different height. The same happens to participant number three as seen in Fig. 4.8 C. The similarity between these two participants is that they are both men and have a difference in height of 8 cm. In the triplet consisting of dance movements 5, 6, and 7, the seventh and fifth gestures were falsely identified as number six for the case of participant number one. In the case of the third participant, movement number seven has 20 % of its instances confused with the fifth movement. Such gestures include moving both arms in between the legs. A significant difference in the activities is how the legs move; left/right leg in the air or both feet on the ground with the knee bent, and how the shoulders move. The lack of antennas on the shoulder blades and not antennas on the lower part of the body could be the sources of the misclassification. For the second participant, an f1-score = 100 % was achieved. This volunteer is a woman with a height of 160 cm. The "MoCaBlazer" was looser for the second participant, which indicates the blazer has more flexibility and could be interpreted as more wrinkles on the garment while doing the movements.

The fourth model was developed using the LRO scheme depicted in Fig. 4.4 B. With this model two tests were performed; LRO-Offline with the three participants and confusion matrix in Figure Fig. 4.9 A, and the second test was a real-time (online) with RFID synchronization which performance is in Fig. 4.9 B.

For the first test of the fourth model (offline), the highest recognition error was observed for two pairs of classes, 4/1 and 3/0, with 13 % of the instances being wrongly recognized. These two pairs of classes consisted of both arms moving from the standing position (starting point) to the right/left, with the main difference in how much height the arms reach, including a visually distinctive shoulder movement. As seen in the individual models in Fig. 4.8, the classes number 5, 6, and 7 are confused with each other, which also occurs in the group-model/fourth model, so it was a foreseen situation. An f1-score = 92 % for the recognition of the gestures in the dance movements dictionary makes our system a good solution for a sophisticated and elegant dance game controller.


The second test result, the real-time with RFID synchronization in Fig. 4.9 B, shows perfect recognition for the dance movements 0,1 and 2. This is not the case for movement number 3, with 40 % of its instances being confused with the merged class 6-7. The merged class 6-7 could be considered as activity number 6 in Fig. 4.3 G, due to the consistent recognition of dance movement

number 7 as dance movement number 6. Therefore, the comparison between dance gesture number 3 in Fig. 4.3 **D** and gesture number 6 in Fig. 4.3 **G** applies. We suspect two reasons for the 40 % wrongly recognized instances; the first could be the slight height difference in the arms' positions and the non-presence of antennas on the shoulders or around the legs. Secondly, it is essential to remark that this confusion is not present in the offline results, which concludes that our solution depends highly on excellent labeling to mark the gesture's starting point and ending point.

The offline results were obtained using labeling/marketing the starting point and ending point with high accuracy in 50 fps/camera. The RFID labeling or marking of the starting point and ending point has an intrinsic error of a slight hand movement (location of RFID reader) to get close enough and detect the RFID tag (on the side pocket). In addition, the RFID solution has a granularity of seconds instead of milliseconds (video-based labeling /offline case)

The merged class 4-5 has a 33 % misclassification with class number 1, and this confusion can also be observed in the offline result of the three volunteers model. Despite the far from perfect RFID synchronization to signal a gesture sequence "standing-gesture-standing" in comparison with the offline version (in the order of milliseconds), we could consider it a promising technique for real-time recognition. A solution to improve the RFID fusion results could be to train the model with data synchronized through the RFID in-situ labeling.

4.8. Conclusion

This work has explored a method for posture and body gesture recognition based on a commercially available electronic theremin, the "OpenTheremin", which, together with conductive textile antennas, was embedded in a loose-fitting garment, the "MoCaBlazer". Our solution can be deployed and integrated in a fashion and fast manner into loose garments. The "MoCaBlazer" was evaluated with fourteen participants (sex-balanced) mimicking a general dictionary of 20 upper-body movements. Additionally, as an application-specific evaluation, a pool of three volunteers participated in mimicking an eight dance movements dictionary inspired by the Rayman Raving Rabbids: TV Party-Nintendo Wii game.

For the 20 gestures dictionary, different deep learning models were selected, such as 1D-LeNet5, DeepConvLSTM, and Conv2D. For the case of the eight dance movements dictionary, a one-dimension convolutional neural network was selected. In both evaluations, the system has offered competitive performance compared to state-of-the-art in loose garments for BPG detection. In the experiment design, repeated wearing of the "MoCaBlazer" was enforced (per session) to make the results robust against disturbances of re-wearing.

With our chosen sensing modality, the non-contact capacitive method, we use the advantages of being independent of muscular strength/pressure and, therefore, no need for tight or elastic garments. In addition, it is relatively not sensitive to sweat or skin dryness [219]. A limitation of the capacitive sensing

modality is that it is sensitive to conductors, which include persons/objects in close range with different dielectric properties compared to the antennas [147]. To avoid the effect of environmental disturbances as much as possible, we normalized our data per gesture window, removing the dependency on absolute values, and built our system upon the relative differences between capacitive channels.

The "MoCaBlazer" data collection for the dance gesture experiment was wirelessly transmitted to an Android phone application. With an f1-score = 92% for eight classes with wirelessly collected data, our design demonstrated robustness against capacitive channel drifting values due to floating ground conditions (typical case in wearables). Moreover, a real-time test with RFID synchronization was done (wireless-online) for one volunteer with f1-score = 82 % for six classes.

Our "MoCaBlazer" evaluation has shown promising results in loose garments as a body posture detection method. Hence, we would continue developing elaborated garment integration; with miniaturized sensing modules, more channels, stretchable antennas, and different antenna pattern designs. In the future, the fusion with other sensors such as IMU for continuous posture detection will be an exciting field to explore, in addition to real-time system deployment/evaluation at the edge (embedded devices).

Chapter 5

Conclusion

This dissertation has brought to the wearable community a set of measurement tools that can be used to expand the knowledge about the expressiveness of body movements. This work focused on wearable-based design solutions to be ubiquitous and include the situational context of the body's actions. The aim was the exploration of experimental scenarios and gestures that happen when the body expresses itself. It is relevant to highlight that it is understood that the evaluations presented in this thesis are hardware-based capability evaluations. This means that it was never intended to stimulate real emotions in the participants. Overall, the designs have been tested in experimental settings with evaluation based on mimicked gestures.

The exploration includes scenarios such as vertical hand positioning in cases such as a position relative to the environment in an order and picking scenarios and positions relative to body parts (head, chest, and feet). A fusion between differential atmospheric pressure and radio-frequency identification technology is employed for those specific scenarios. Hand gesture recognition is explored with a multimodal fusion, employing textile-based capacitive sensing channels and inertial information to recognize a gesture dictionary used for drone control. Even though a specific set of gestures was recognized, the system can be extended to applications such as signal language and game control. The hand vertical positioning and hand gesture recognition application are the topics of Chapter 2.

In Chapter 3, the goal is to monitor facial and head muscle movements. The modalities explored were passive-based such as mechanomyography (MMG), including sound and pressure-based mechnomyography and inertial sensing information. The chapter presented the evolution of the hardware-software co-design. The first design is a bulky but wearable helmet design with stethoscope microphones distributed around the face. With this bulky design, this work put forward the idea of using differential sound MMG information for facial muscle movement recognition. The second design demonstrates that the privacy-aware audio information for facial muscle pattern recognition is relevant but requires higher complexity than pressure MMG and inertial information. Besides, the size factor of the stethoscope-microphone negatively

affects the participant’s comfort. Hence, the last design of the chapter is a glasses-based design that includes pressure MMG and inertial sensing modalities for the real-time and on-the-edge evaluation of the idea. The glasses-based design recognizes the facial muscle movement related to facial expression and eating/drinking episodes. The goal is to expand further capabilities to quantify facial movement caused by stress-related eating/drinking behaviors.

Finally, in Chapter 4, body posture gesture (BPG) recognition is studied. We introduce an approach for detecting BPG that does not require sensors to be firmly fixed to the body or integrated into a tight-fitting garment. Instead, sensing is incorporated into a loose-fitting garment. Evaluating the proposed approach with the MoCaBlazer with 14 diverse participants in an experiment to detect 20 body postures and gestures. The work is expanded by fusing multipositional capacitive sensing with Radio Frequency Identification (RFID) synchronization for real-time and wireless recognition for one participant and six classes of a dance movements dictionary. Thus, our ”MoCaBlazer” could be a promising alternative for an elegant/sophisticated game controller.

In summary, our hardware and software co-designs focused on the challenges in wearable and ubiquitous computing. Such as restrictions in memory, power, and reduced computing capabilities compared to server/cloud-based solutions. Furthermore, our sensing modalities are tailored to face specific tasks which reduces the complexity of the algorithm while increasing its performance in embedded devices. The sensing modalities employed are privacy-aware and passive. The experiment designs followed the Declaration of Helsinki[171] for the protection of human subjects and the Ethical Principles for Research with Human Subjects, the Belmont Agreement[43]. This means that the participant’s comfort, data, and safety were the priority in developing human-centric artificial intelligence-based solutions.

5.1. Limitations and Future Work

In general, our set of measurement tools for the “Expressiveness of Human Body Movements” are not end-to-end designs. This means that our systems, although they have proven their relevance in controlled experimental environments, are still far from being commercial solutions. Experimental settings in the wild and on-the-edge evaluations are important future steps. A summary list of relevant limitations is as follows:

- **Size of the prototypes:** Our designs have the potential to be greatly reduced in size with proper electronic printed board designs. This will facilitate experimental settings in the wild and the evaluation of the user experience on an everyday basis.
- **Latency:** The algorithms are not optimized to the capabilities of the hardware. Techniques such as multicore running, sleep modes, and memory sharing can positively influence the inference time of the models.
- **Model Size:** For the case of neural network-based solutions, we have

not employed size optimization techniques such as quantization-aware training, pruning, or compression using knowledge distillation.

- **Stimulate real expressions:** Our experiments are based on mimicking. They focus on testing the potential of the hardware/software designs. Still, monitoring of real expressions is missing. To induce real expression in humans is another field of research and it is outside of the scope of this work.

The two most relevant future aspects are the reduction of the size of the systems and their evaluation in the wild with realistic body movements. All of the designs have the potential of being deployed on the edge. With tuned hardware designs it is possible to reduce the latency and deploy the neural network-based algorithm into the embedded device. Complete data processing in embedded devices will improve data security and privacy protection and, more importantly, will give systems the ubiquity needed for real and continuous monitoring of user expressiveness.

To finish this dissertation here are the thoughts of Nikola Tesla. *It is paradoxical, yet true, to say, that the more we know, the more ignorant we become in the absolute sense, for it is only through enlightenment that we become conscious of our limitations. Precisely one of the most gratifying results of intellectual evolution is the continuous opening up of new and greater prospects.*

Bibliography

- [1] *25 Celebrities Sticking Out Their Tongues — Brad pitt, Stick it out, George clooney*. <https://www.pinterest.de/pin/243757398561743241/>. (Accessed on 30/01/2024)).
- [2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. “Tensorflow: A system for large-scale machine learning”. In: *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016, pp. 265–283.
- [3] T. D. Abhayapala and A. Gupta. “Alternatives to spherical microphone arrays: Hybrid geometries”. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Apr. 2009, pp. 81–84. DOI: 10.1109/ICASSP.2009.4959525.
- [4] Zamir Ahmed Abro, Yi-Fan Zhang, Cheng-Yu Hong, Rafique Ahmed Lakho, and Nan-Liang Chen. “Development of a smart garment for monitoring body postures based on FBG and flex sensing technologies”. In: *Sensors and Actuators A: Physical* 272 (2018), pp. 153–160.
- [5] Talha Agcayazi, Jordan Tabor, Michael McKnight, Isaac Martin, Tushar K Ghosh, and Alper Bozkurt. “Fully-textile seam-line sensors for facile textile integration and tunable multi-modal sensing of pressure, humidity, and wetness”. In: *Advanced materials technologies* 5.8 (2020), p. 2000155.
- [6] C. Aguilera-Astudillo, M. Chavez-Campos, A. Gonzalez-Suarez, and J. L. Garcia-Cordero. “A low-cost 3-D printed stethoscope connected to a smartphone”. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Aug. 2016, pp. 4365–4368. DOI: 10.1109/EMBC.2016.7591694.
- [7] Takashi Amesaka, Hiroki Watanabe, and Masanori Sugimoto. “Facial expression recognition using ear canal transfer function”. In: *Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 2019, pp. 1–9.
- [8] Amphenol. *Flat Ribbon Cable*. AMPHENOL SPECTRA-STRIP. 2015. URL: <https://at.farnell.com/amphenol-spectra-strip/191-2801-150/kabel-flachb-grau-rast1-27mm-50adr/dp/1170229>.

- [9] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. “Canalsense: Face-related movement recognition system based on sensing air pressure in ear canals”. In: *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. 2017, pp. 679–689.
- [10] Hiroaki Aoki, Ayumi Ohnishi, Naoya Isoyama, Tsutomu Terada, and Masahiko Tsukamoto. “FaceRecGlasses: A Wearable System for Recognizing Self Facial Expressions Using Compact Wearable Cameras”. In: *Proceedings of the Augmented Humans International Conference 2021*. 2021, pp. 55–65.
- [11] D. Aspandi, O. Martinez, F. Sukno, and X. Binefa. “Fully End-to-End Composite Recurrent Convolution Network for Deformable Facial Tracking In The Wild”. In: *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*. 2019, pp. 1–8.
- [12] Bigland-Ritchie B and J Woods J. “Changes in muscle contractile properties and neural control during human muscular fatigue”. In: *Muscle Nerve* 7.9 (1984), pp. 691–699. ISSN: 1097-4598. DOI: 10.1002/mus.880070902.
- [13] Q. Bao, F. Luan, and J. Yang. “Improving the accuracy of beamforming method for moving acoustic source localization in far-field”. In: *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. Oct. 2017, pp. 1–6. DOI: 10.1109/CISP-BMEI.2017.8302132.
- [14] Hymalai Bello, Daniel Geißler, Sungho Suh, Bo Zhou, and Paul Lukowicz. “TSAK: Two-Stage Semantic-Aware Knowledge Distillation for Efficient Wearable Modality and Model Optimization in Manufacturing Lines”. In: *Pattern Recognition*. Ed. by Apostolos Antonacopoulos, Subhasis Chaudhuri, Rama Chellappa, Cheng-Lin Liu, Saumik Bhattacharya, and Umпада Pal. Cham: Springer Nature Switzerland, 2025, pp. 201–216. ISBN: 978-3-031-78389-0.
- [15] Hymalai Bello, Luis Alfredo Sanchez Marin, Sungho Suh, Bo Zhou, and Paul Lukowicz. “InMyFace: Inertial and mechanomyography-based sensor fusion for wearable facial activity recognition”. In: *Information Fusion* (2023), p. 101886.
- [16] Hymalai Bello, Jhonny Rodriguez, and Paul Lukowicz. “Vertical hand position estimation with wearable differential barometry supported by RFID synchronization”. In: *EAI International Conference on Body Area Networks*. Springer. 2019, pp. 24–33.
- [17] Hymalai Bello, Sungho Suh, Daniel Geißler, Lala Ray, Bo Zhou, and Paul Lukowicz. “Real-Time and on-the-Edge Multiple Channel Capacitive and Inertial Fusion-Based Glove”. In: *Body Area Networks. Smart IoT and Big Data for Intelligent Health Management*. Ed. by Marouan Mizmizi, Maurizio Magarini, Prabhat Kumar Upadhyay, and Massimil-

- iano Pierobon. Cham: Springer Nature Switzerland, 2024, pp. 166–176. ISBN: 978-3-031-72524-1.
- [18] Hymalai Bello, Sungho Suh, Daniel Geißler, Lala Shakti Swarup Ray, Bo Zhou, and Paul Lukowicz. “CaptAinGlove: Capacitive and inertial fusion-based glove for real-time on edge hand gesture recognition for drone control”. In: *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*. 2023, pp. 165–169.
- [19] Hymalai Bello, Sungho Suh, Bo Zhou, and Paul Lukowicz. “FaceEat: Facial and Eating Activities Recognition with Inertial and Mechanomyography Fusion using a Glasses-Based Design for Real-Time and on-the-Edge Inference”. In: *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*. 2023, pp. 199–199.
- [20] Hymalai Bello, Sungho Suh, Bo Zhou, and Paul Lukowicz. “MeciFace: Mechanomyography and Inertial Fusion based Glasses for Edge Real-Time Recognition of Facial and Eating Activities”. In: *arXiv preprint arXiv:2306.13674* (2023).
- [21] Hymalai Bello, Bo Zhou, and Paul Lukowicz. “Facial muscle activity recognition with reconfigurable differential stethoscope microphones”. In: *Sensors* 20.17 (2020), p. 4904.
- [22] Hymalai Bello, Bo Zhou, Sungho Suh, and Paul Lukowicz. “Mocapaci: Posture and gesture detection in loose garments using textile cables as capacitive antennas”. In: *Proceedings of the 2021 ACM International Symposium on Wearable Computers*. 2021, pp. 78–83.
- [23] Hymalai Bello, Bo Zhou, Sungho Suh, Luis Alfredo Sanchez Marin, and Paul Lukowicz. “Move with the theremin: Body posture and gesture recognition using the theremin in loose-garment with embedded textile cables as antennas”. In: *Frontiers in Computer Science* 4 (2022), p. 915280.
- [24] Yoav Benjamini and Daniel Yekutieli. “The control of the false discovery rate in multiple testing under dependency”. In: *Ann. Statist.* 29.4 (Aug. 2001), pp. 1165–1188. DOI: 10.1214/aos/1013699998. URL: <https://doi.org/10.1214/aos/1013699998>.
- [25] Romain Bey, Romain Goussault, François Grolleau, Mehdi Benchoufi, and Raphaël Porcher. “Fold-stratified cross-validation for unbiased and privacy-preserving federated learning”. In: *Journal of the American Medical Informatics Association* 27.8 (2020), pp. 1244–1251.
- [26] Chongguang Bi, Jun Huang, Guoliang Xing, Landu Jiang, Xue Liu, and Minghua Chen. “Safewatch: A wearable hand motion tracking system for improving driving safety”. In: *ACM Transactions on Cyber-Physical Systems* 4.1 (2019), pp. 1–21.

- [27] Ekaba Bisong. “Google Colaboratory”. In: *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*. Berkeley, CA: Apress, 2019, pp. 59–64. ISBN: 978-1-4842-4470-8. DOI: 10.1007/978-1-4842-4470-8_7. URL: https://doi.org/10.1007/978-1-4842-4470-8_7.
- [28] Ali Boyali, Manolya Kavakli, et al. “A robust and fast gesture recognition method for wearable sensing garments”. In: *Proc. Int. Conf. Adv. Multimedia*. Chamonix / Mont Blanc, France: International Academy, Research, and Industry Association (IARIA), 2012, pp. 142–147.
- [29] Danail S Brezov, Clementina D Mladenova, and Iva Mladenov. “New perspective on the gimbal lock problem”. In: *AIP Conference Proceedings*. Vol. 1570. 1. American Institute of Physics. 2013, pp. 367–374.
- [30] Y. Buchris, I. Cohen, and J. Benesty. “Asymmetric Supercardioid Beamforming Using Circular Microphone Arrays”. In: *2018 26th European Signal Processing Conference (EUSIPCO)*. Sept. 2018, pp. 627–631. DOI: 10.23919/EUSIPCO.2018.8553582.
- [31] Y. Buchris, I. Cohen, and J. Benesty. “First-order differential microphone arrays from a time-domain broadband perspective”. In: *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. Sept. 2016, pp. 1–5. DOI: 10.1109/IWAENC.2016.7602886.
- [32] Hammad Tanveer Butt, Manthan Pancholi, Mathias Musahl, Pramod Murthy, Maria Alejandra Sanchez, and Didier Stricker. “Inertial Motion Capture Using Adaptive Sensor Fusion and Joint Angle Drift Correction”. In: *2019 22th International Conference on Information Fusion (FUSION)*. Ottawa, ON, Canada: IEEE, 2019, pp. 1–8.
- [33] Joon Byun, Young-cheol Park, and Sung Wook Park. “Continuously steerable second-order differential microphone arrays”. In: *Acoustical Society of America Journal* 143.3 (Mar. 2018), EL225–EL230. DOI: 10.1121/1.5027500.
- [34] *Canadian Kiss Stock-Illustration - Getty Images*. <https://www.gettyimages.de/detail/illustration/canadian-kiss-lizenfreie-illustration/472283539?adppopup=true>. (Accessed on 30/01/2024).
- [35] Youngsu Cha, Hojoon Kim, and Doik Kim. “Flexible piezoelectric sensor-based gait recognition”. In: *Sensors* 18.2 (2018), p. 468.
- [36] Youngsu Cha, Kihyuk Nam, and Doik Kim. “Patient posture monitoring system based on flexible sensors”. In: *Sensors* 17.3 (2017), p. 584.
- [37] Harish Chander, Reuben F Burch, Purva Talegaonkar, David Saucier, Tony Luczak, John E Ball, Alana Turner, Sachini NK Kodithuwakku Arachchige, Will Carroll, Brian K Smith, et al. “Wearable stretch sensors for human movement monitoring and fall detection in ergonomics”. In: *International journal of environmental research and public health* 17.10 (2020), p. 3554.

-
- [38] P. Charlier, C. Herman, N. Rochedreux, R. Logier, C. Garabedian, V. Debarge, and J. D. Jonckheere. “AcCorps: A low-cost 3D printed stethoscope for fetal phonocardiography*”. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. July 2019, pp. 52–55. DOI: 10.1109/EMBC.2019.8856575.
- [39] Paresh M Chauhan and Nikita P Desai. “Mel frequency cepstral coefficients (MFCC) based speaker identification in noisy environment using wiener filter”. In: *2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)*. IEEE, 2014, pp. 1–5.
- [40] T. Chen, Q. Huang, L. Zhang, and Y. Fang. “Direction of Arrival Estimation Using Distributed Circular Microphone Arrays”. In: *2018 14th IEEE International Conference on Signal Processing (ICSP)*. Aug. 2018, pp. 182–185. DOI: 10.1109/ICSP.2018.8652374.
- [41] Tuochao Chen, Yaxuan Li, Songyun Tao, Hyunchul Lim, Mose Sakashita, Ruidong Zhang, Francois Guimbretiere, and Cheng Zhang. “NeckFace: Continuously Tracking Full Facial Expressions on Neck-mounted Wearables”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.2 (2021), pp. 1–31.
- [42] Tuochao Chen, Benjamin Steeper, Kinan Alsheikh, Songyun Tao, François Guimbretière, and Cheng Zhang. “C-face: Continuously reconstructing facial expressions by deep learning contours of the face with ear-mounted miniature cameras”. In: *Proceedings of the 33rd annual ACM symposium on user interface software and technology*. 2020, pp. 112–125.
- [43] James F Childress, Eric Mark Meslin, and Harold T Shapiro. *Belmont revisited: Ethical principles for research with human subjects*. Georgetown University Press, 2005.
- [44] Seokmin Choi, Yang Gao, Yincheng Jin, Se jun Kim, Jiyang Li, Wenyao Xu, and Zhanpeng Jin. “PPGface: Like What You Are Watching? Ear-phones Can” Feel” Your Facial Expressions”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.2 (2022), pp. 1–32.
- [45] Francois Chollet et al. *Keras*. <https://github.com/fchollet/keras>. Last accessed: April 05, 2022. 2015.
- [46] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. “Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)”. In: *Neurocomputing* 307 (2018), pp. 72–77. ISSN: 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2018.03.067>. URL: <http://www.sciencedirect.com/science/article/pii/S0925231218304843>.

- [47] Jan Pieter Clarys and Jan Cabri. “Electromyography and the study of sports movements: a review”. In: *Journal of sports sciences* 11.5 (1993), pp. 379–448.
- [48] Juan Antonio Corrales, Francisco A Candelas, and Fernando Torres. “Hybrid tracking of human operators using IMU/UWB data fusion by a Kalman filter”. In: *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. 2008, pp. 193–200.
- [49] Giuseppe D’Aniello, Raffaele Gravina, Matteo Gaeta, and Giancarlo Fortino. “Situation-aware sensor-based wearable computing systems: A reference architecture-driven review”. In: *IEEE Sensors Journal* (2022).
- [50] Joseph DelPreto, Josie Hughes, Matteo D’Aria, Marco de Fazio, and Daniela Rus. “A Wearable Smart Glove and Its Application of Pose and Gesture Detection to Sign Language Classification”. In: *IEEE Robotics and Automation Letters* 7.4 (2022), pp. 10589–10596.
- [51] Han Ding, Lei Guo, Cui Zhao, Fei Wang, Ge Wang, Zhiping Jiang, Wei Xi, and Jizhong Zhao. “RFnet: Automatic gesture recognition and human identification using time series RFID signals”. In: *Mobile Networks and Applications* 25.6 (2020), pp. 2240–2253.
- [52] *Doc2Us - Your Personal Pocket Doctor*. <https://www.doc2us.com/8-whys-your-toddler-blinking-hard-complete-list>. (Accessed on 30/01/2024).
- [53] G. Drzewiecki, H. Katta, A. Pfahnl, D. Bello, and D. Dicken. “Active and passive stethoscope frequency transfer functions: Electronic stethoscope frequency response”. In: *2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. Dec. 2014, pp. 1–4. DOI: 10.1109/SPMB.2014.7002962.
- [54] Catia Real Ehrlich and Jörg Blankenbach. “Pedestrian localisation inside buildings based on multi-sensor smartphones”. In: *2018 Ubiquitous Positioning, Indoor Navigation and Location-Based Services (UPINLBS)*. IEEE. 2018, pp. 1–10.
- [55] Paul Ekman and Wallace V Friesen. “Facial action coding system”. In: *Environmental Psychology & Nonverbal Behavior* (1978).
- [56] Daniyar Enikeev and Svetlana Mustafina. “Recognition of Sign Language Using Leap Motion Controller Data”. In: *2020 2nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*. Lipetsk, Russia: IEEE, 2020, pp. 393–397.
- [57] Sara V Fernandez, Fiona Cai, Sophia Chen, Emma Suh, Jan Tiepelt, Rachel McIntosh, Colin Marcus, Daniel Acosta, David Mejorado, and Canan Dagdeviren. “On-Body Piezoelectric Energy Harvesters through Innovative Designs and Conformable Structures”. In: *ACS Biomaterials Science & Engineering* (2021).

-
- [58] Limor Fried. *HuzzahESP32*. Adafruit. 2022. URL: <https://learn.adafruit.com/adafruit-huzzah32-esp32-feather>.
- [59] E. Fujiwara, Y. T. Wu, C. K. Suzuki, D. T. G. de Andrade, A. R. Neto, and E. Rohmer. “Optical fiber force myography sensor for applications in prosthetic hand control”. In: *2018 IEEE 15th International Workshop on Advanced Motion Control (AMC)*. Mar. 2018, pp. 342–347. DOI: 10.1109/AMC.2019.8371115.
- [60] Kyosuke Futami, Kohei Oyama, and Kazuya Murao. “A Method to Recognize Facial Gesture Using Infrared Distance Sensor Array on Ear Accessories”. In: *The 23rd International Conference on Information Integration and Web Intelligence*. 2021, pp. 650–654.
- [61] Konrad Gadzicki, Raziieh Khamsehashari, and Christoph Zetzsche. “Early vs late fusion in multimodal convolutional neural networks”. In: *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. IEEE. 2020, pp. 1–6.
- [62] Andre Gaschler, Sören Jentzsch, Manuel Giuliani, Kerstin Huth, Jan de Ruiter, and Alois Knoll. “Social behavior recognition using body posture and head pose for human-robot interaction”. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2012, pp. 2128–2133.
- [63] Shkurta Gashi, Aaqib Saeed, Alessandra Vicini, Elena Di Lascio, and Silvia Santini. “Hierarchical classification and transfer learning to recognize head gestures and facial expressions using earbuds”. In: *Proceedings of the 2021 International Conference on Multimodal Interaction*. 2021, pp. 168–176.
- [64] Urs Gaudenz. *OpenTheremin-GaudiLabs*. <https://gaudishop.ch/index.php/product-category/opentheremin/>. Last accessed: February 06, 2024. GaudiLabs LLC, 2016.
- [65] Ali Ghorbandaei Pour, Alireza Taheri, Minoo Alemi, and Ali Meghdari. “Human-robot facial expression reciprocal interaction platform: case studies on children with autism”. In: *International Journal of Social Robotics* 10.2 (2018), pp. 179–198.
- [66] Hristijan Gjoreski, Ifigeneia Mavridou, James Archer William Archer, Andrew Cleal, Simon Stankoski, Ivana Kiprijanovska, Mohsen Fatoorechi, Piotr Walas, John Broulidakis, Martin Gjoreski, et al. “OCOSense Glasses—Monitoring Facial Gestures and Expressions for Augmented Human-Computer Interaction: OCOSense Glasses for Monitoring Facial Gestures and Expressions”. In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–4.
- [67] Martin Gjoreski, Ivana Kiprijanovska, Simon Stankoski, Ifigeneia Mavridou, M John Broulidakis, Hristijan Gjoreski, and Charles Nduka. “Facial EMG sensing for monitoring affect using a wearable device”. In: *Scientific Reports* 12.1 (2022), p. 16876.
-

- [68] A. K. Godiyal, M. Mondal, S. D. Joshi, and D. Joshi. “Force Myography Based Novel Strategy for Locomotion Classification”. In: *IEEE Transactions on Human-Machine Systems* 48.6 (Dec. 2018), pp. 648–657. ISSN: 2168-2305. DOI: 10.1109/THMS.2018.2860598.
- [69] Siome Goldenstein and Jonas Gomes. “Time warping of audio signals”. In: *Computer Graphics International Conference*. IEEE Computer Society. 1999, pp. 52–52.
- [70] Raffaele Gravina and Giancarlo Fortino. “Wearable body sensor networks: state-of-the-art and research directions”. In: *IEEE Sensors Journal* 21.11 (2020), pp. 12511–12522.
- [71] Hakim Guedjou, Sofiane Boucenna, and Mohamed Chetouani. “Posture recognition analysis during human-robot imitation learning”. In: *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE. 2016, pp. 193–194.
- [72] Xiaotian Guo, Andy D Pimentel, and Todor Stefanov. “Automated Exploration and Implementation of Distributed CNN Inference at the Edge”. In: *IEEE Internet of Things Journal* (2023).
- [73] Marian Haescher, Denys JC Matthies, Gerald Bieber, and Bodo Urban. “Capwalk: A capacitive recognition of walking-based activities as a wearable assistive technology”. In: *Proceedings of the 8th ACM International Conference on Pervasive Technologies Related to Assistive Environments*. Corfu, Greece: Association for Computing Machinery, 2015, pp. 1–8.
- [74] Holger Harms, Oliver Amft, Gerhard Tröster, and Daniel Roggen. “Smash: A distributed sensing and processing garment for the classification of upper body postures”. In: *Proceedings of the ICST 3rd international conference on Body area networks*. Tempe, Arizona: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008, pp. 1–8.
- [75] Md Rashidul Hasan, Mustafa Jamil, MGRMS Rahman, et al. “Speaker identification using mel frequency cepstral coefficients”. In: *variations* 1.4 (2004), pp. 565–568.
- [76] Changli He and Rickard Sandberg. “Dickey–Fuller type of tests against nonlinear dynamic models”. In: *Oxford Bulletin of Economics and Statistics* 68 (2006), pp. 835–861.
- [77] H. He, X. Qiu, and T. Yang. “On directivity of a circular array with directional microphones”. In: *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. Sept. 2016, pp. 1–5. DOI: 10.1109/IWAENC.2016.7602924.
- [78] Abdelfetah Hentout, Mustapha Aouache, Abderraouf Maoudj, and Isma Akli. “Human–robot interaction in industrial collaborative robotics: a literature review of the decade 2008–2017”. In: *Advanced Robotics* 33.15-16 (2019), pp. 764–799.

-
- [79] Md Afzal Hossan, Sheeraz Memon, and Mark A Gregory. “A novel approach for MFCC feature extraction”. In: *2010 4th International Conference on Signal Processing and Communication Systems*. IEEE, 2010, pp. 1–5.
- [80] Chih-wei Hsu, Chih-chung Chang, and Chih-Jen Lin. “A Practical Guide to Support Vector Classification Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin”. In: (Nov. 2003).
- [81] H. Huang, D. Yang, X. Yang, Y. Lei, and Y. Chen. “Portable multifunctional electronic stethoscope”. In: *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (IT-NEC)*. Mar. 2019, pp. 691–694. DOI: 10.1109/ITNEC.2019.8729172.
- [82] L. K. Huang, L. N. Huang, Y. M. Gao, Ž. Lučev Vasić, M. Cifrek, and M. Du. “Electrical Impedance Myography Applied to Monitoring of Muscle Fatigue During Dynamic Contractions”. In: *IEEE Access* 8 (2020), pp. 13056–13065. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2965982.
- [83] Zhentao Huang, Rongze Li, Wangkai Jin, Zilin Song, Yu Zhang, Xiangjun Peng, and Xu Sun. “Face2Multi-modal: In-vehicle multi-modal predictors via facial expressions”. In: *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 2020, pp. 30–33.
- [84] *Hugh Jackman — Wolverine hugh jackman, Hugh jackman, Jackman*. <https://www.pinterest.de/pin/361765782554181392/>. (Accessed on 30/01/2024).
- [85] “InvenSense Inc”. *Microphone Array Beamforming*. <https://invensense.tdk.com/wp-content/uploads/2015/02/Microphone-Array-Beamforming.pdf>. Application Note number AN-1140, Rev 1.0, InvenSense Inc, September 2013. Dec. 2013.
- [86] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International conference on machine learning*. Lille, France: PMLR, 2015, pp. 448–456.
- [87] Kenichi Ito, Takahiko Masuda, and Koichi Hioki. “Affective information in context and judgment of facial expression: Cultural similarities and variations in context effects between North Americans and East Asians”. In: *Journal of Cross-Cultural Psychology* 43.3 (2012), pp. 429–445.
- [88] N. Jatupaiboon, S. Pan-ngum, and P. Israsena. “Electronic stethoscope prototype with adaptive noise cancellation”. In: *2010 Eighth International Conference on ICT and Knowledge Engineering*. Nov. 2010, pp. 32–36. DOI: 10.1109/ICTKE.2010.5692909.

- [89] Shuo Jiang, Bo Lv, Weichao Guo, Chao Zhang, Haitao Wang, Xinjun Sheng, and Peter B Shull. “Feasibility of wrist-worn, real-time hand, and surface gesture recognition via sEMG and IMU sensing”. In: *IEEE Transactions on Industrial Informatics* 14.8 (2017), pp. 3376–3385.
- [90] Husam Khalaf Salih Juboori and Lalit Kulkarni. “Fatigue detection system for the drivers using video analysis of facial expressions”. In: *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*. IEEE. 2017, pp. 1–9.
- [91] Pyeong-Gook Jung, Gukchan Lim, Seonghyok Kim, and Kyoungchul Kong. “A wearable gesture recognition device for detecting muscular activities based on air-pressure sensors”. In: *IEEE Transactions on Industrial Informatics* 11.2 (2015), pp. 485–494.
- [92] Julio Cezar Silveira Jacques Junior, Yağmur Güçlütürk, Marc Pérez, Umut Güçlü, Carlos Andujar, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Marcel AJ Van Gerven, Rob Van Lier, et al. “First impressions: A survey on vision-based apparent personality trait analysis”. In: *IEEE Transactions on Affective Computing* 1.1 (2019), pp. 1–1.
- [93] Vemema Kangunde, Rodrigo S Jamisola, and Emmanuel K Theophilus. “A review on drones controlled in real-time”. In: *International journal of dynamics and control* (2021), pp. 1–15.
- [94] Muhammad US Khan, Assad Abbas, Mazhar Ali, Muhammad Jawad, and Samee U Khan. “Convolutional neural networks as means to identify apposite sensor combination for human activity recognition”. In: *2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE. 2018, pp. 45–50.
- [95] Muhammad US Khan, Assad Abbas, Mazhar Ali, Muhammad Jawad, and Samee U Khan. “Convolutional neural networks as means to identify apposite sensor combination for human activity recognition”. In: *2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE. 2018, pp. 45–50.
- [96] Ali Kiaghadi, Morgan Baima, Jeremy Gummeson, Trisha Andrew, and Deepak Ganesan. “Fabric as a sensor: Towards unobtrusive sensing of human behavior with triboelectric textiles”. In: *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. Shenzhen, China: Association for Computing Machinery, 2018, pp. 199–210.
- [97] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: (2017). arXiv: 1412.6980 [cs.LG].
- [98] Nikita Kiselov. *Drone control via gestures using Mediapipe hands*. Sept. 2021. URL: <https://developers.googleblog.com/2021/09/drone-control-via-gestures-using-mediapipe-hands.html>.

- [99] Byoungchul Ko. “A Brief Review of Facial Emotion Recognition Based on Visual Information”. In: *Sensors* 18 (Jan. 2018), p. 401. DOI: 10.3390/s18020401.
- [100] Ron Koepp, Terry Allen, Jay Fassett, and Annette Teng. “Achieving high speed RFID die pick and place operation”. In: *2008 33rd IEEE/CPMT International Electronics Manufacturing Technology Conference (IEMT)*. IEEE. 2008, pp. 1–8.
- [101] Yuya Koyama, Michiko Nishiyama, and Kazuhiro Watanabe. “Gait monitoring for human activity recognition using perceptive shoe based on hetero-core fiber optics”. In: *2016 IEEE 5th Global Conference on Consumer Electronics*. Kyoto, Japan: IEEE, 2016, pp. 1–2.
- [102] Yuya Koyama, Michiko Nishiyama, and Kazuhiro Watanabe. “Physical activity recognition using hetero-core optical fiber sensors embedded in a smart clothing”. In: *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*. Nara, Japan: IEEE, 2018, pp. 71–72.
- [103] R. K. Kusainov and V. K. Makukha. “Evaluation of the applicability of MEMS microphone for auscultation”. In: *2015 16th International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices*. June 2015, pp. 595–597. DOI: 10.1109/EDM.2015.7184613.
- [104] Ik-Jae Kwon, Tae-Young Jung, Youjeong Son, Bongju Kim, Soung-Min Kim, and Jong-Ho Lee. “Detection of volatile sulfur compounds (VSCs) in exhaled breath as a potential diagnostic method for oral squamous cell carcinoma”. In: *BMC Oral Health* 22.1 (2022), pp. 1–8.
- [105] Jangho Kwon, Jihyeon Ha, Da-Hye Kim, Jun Won Choi, and Laehyun Kim. “Emotion recognition using a glasses-type wearable device via multi-channel facial responses”. In: *IEEE Access* 9 (2021), pp. 146392–146403.
- [106] Angela R Laird, Baxter P Rogers, and M Elizabeth Meyerand. “Comparison of Fourier and wavelet resampling methods”. In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 51.2 (2004), pp. 418–422.
- [107] Arnaldo G Leal-Junior, Diana Ribeiro, Leticia M Avellar, Mariana Silveira, Camilo A Rodriguez Díaz, Anselmo Frizzera-Neto, Wilfried Blanc, Eduardo Rocon, and Carlos Marques. “Wearable and Fully-Portable Smart Garment for Mechanical Perturbation Detection With Nanoparticles Optical Fibers”. In: *IEEE Sensors Journal* 21.3 (2020), pp. 2995–3003.
- [108] Yann LeCun et al. “LeNet-5, convolutional neural networks”. In: *URL: <http://yann.lecun.com/exdb/lenet>* 20.5 (2015), p. 14.
- [109] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

- [110] DongSun Lee, Kwi Woo Park, Chansik Park, and In-Suk Kang. “An efficient heave estimation using Time-Differenced GPS carrier phase measurements and compensated barometer measurement applying error model”. In: *2015 International Association of Institutes of Navigation World Congress (IAIN)*. IEEE. 2015, pp. 1–6.
- [111] Jaehong Lee, Hyukho Kwon, Jungmok Seo, Sera Shin, Ja Hoon Koo, Changhyun Pang, Seungbae Son, Jae Hyung Kim, Yong Hoon Jang, Dae Eun Kim, et al. “Conductive fiber-based ultrasensitive textile pressure sensor for wearable electronics”. In: *Advanced materials* 27.15 (2015), pp. 2433–2439.
- [112] Junchan Li, Yu Wang, Pengfei Wang, Qing Bai, Yan Gao, Hongjuan Zhang, and Baoquan Jin. “Pattern Recognition for Distributed Optical Fiber Vibration Sensing: A Review”. In: *IEEE Sensors Journal* 21.10 (2021), pp. 11983–11998.
- [113] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. “EarIO: A Low-power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.2 (2022), pp. 1–24.
- [114] Qimeng Li, Raffaele Gravina, Ye Li, Saeed H Alsamhi, Fangmin Sun, and Giancarlo Fortino. “Multi-user activity recognition: Challenges and opportunities”. In: *Information Fusion* 63 (2020), pp. 121–135.
- [115] Richard Li, Juyoung Lee, Woontack Woo, and Thad Starner. “Kiss-glass: Greeting gesture recognition using smart glasses”. In: *Proceedings of the Augmented Humans International Conference*. 2020, pp. 1–5.
- [116] Qi Lin, Shuhua Peng, Yuezhong Wu, Jun Liu, Wen Hu, Mahbub Hassan, Aruna Seneviratne, and Chun H Wang. “E-Jacket: Posture Detection with Loose-Fitting Garment using a Novel Strain Sensor”. In: *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. Sydney, NSW, Australia: IEEE, 2020, pp. 49–60.
- [117] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas. “Multispectral deep neural networks for pedestrian detection”. In: *arXiv preprint arXiv:1611.02644* (2016).
- [118] Yuchi Liu, Hamideh Khanbareh, Miah Abdul Halim, Andrew Feeney, Xiaosheng Zhang, Hadi Heidari, and Rami Ghannam. “Piezoelectric energy harvesting for self-powered wearable upper limb applications”. In: *Nano Select* 2.8 (2021), pp. 1459–1479.
- [119] Gabriel Loke, Tural Khudiyev, Brian Wang, Stephanie Fu, Syamantak Payra, Yorai Shaoul, Johnny Fung, Ioannis Chatziveroglou, Pin-Wen Chou, Itamar Chinn, et al. “Digital electronics in fibres enable fabric-based machine-learning inference”. In: *Nature communications* 12.1 (2021), pp. 1–9.

-
- [120] Fredrik Lundh. “An introduction to tkinter”. In: *URL: www.pythonware.com/library/tkinter/introduction/index.htm* (1999).
- [121] Yuntao Ma, Yuxuan Liu, Ruiyang Jin, Xingyang Yuan, Raza Sekha, Samuel Wilson, and Ravi Vaidyanathan. “Hand gesture recognition with convolutional neural networks for the multimodal UAV control”. In: *2017 Workshop on Research, Education and Development of Unmanned Aerial Systems (RED-UAS)*. IEEE. 2017, pp. 198–203.
- [122] B. Malik, N. Eya, H. Migdadi, M. J. Ngala, R. A. Abd-Alhameed, and J. M. Noras. “Design and development of an electronic stethoscope”. In: *2017 Internet Technologies and Applications (ITA)*. Sept. 2017, pp. 324–328. DOI: 10.1109/ITECHA.2017.8101963.
- [123] Cristina Manresa-Yee, Silvia Ramis Guarinos, and Jose Maria Buades Rubio. “Facial Expression Recognition: Impact of Gender on Fairness and Expressions”. In: *XXII International Conference on Human Computer Interaction*. 2022, pp. 1–8.
- [124] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: <https://www.tensorflow.org/>.
- [125] M. Martins, P. Gomes, C. Oliveira, M. Coimbra, and H. P. da Silva. “Design and Evaluation of a Diaphragm for Electrocardiography in Electronic Stethoscopes”. In: *IEEE Transactions on Biomedical Engineering* 67.2 (Feb. 2020), pp. 391–398. ISSN: 1558-2531. DOI: 10.1109/TBME.2019.2913913.
- [126] Tania Marur, Yakup Tuna, and Selman Demirci. “Facial anatomy”. In: *Clinics in Dermatology* 32.1 (2014). Red Face Revisited: I, pp. 14–23. ISSN: 0738-081X. DOI: <https://doi.org/10.1016/j.clindermatol.2013.05.022>. URL: <http://www.sciencedirect.com/science/article/pii/S0738081X13000898>.
- [127] Katsutoshi Masai, Kai Kunze, Daisuke Sakamoto, Yuta Sugiura, and Maki Sugimoto. “Face commands-user-defined facial gestures for smart glasses”. In: *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE. 2020, pp. 374–386.
- [128] Katsutoshi Masai, Kai Kunze, Yuta Sugiura, Masa Ogata, Masahiko Inami, and Maki Sugimoto. “Evaluation of facial expression recognition by a smart eyewear for facial direction changes, repeatability, and positional drift”. In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7.4 (2017), pp. 1–23.
- [129] Katsutoshi Masai, Yuta Sugiura, Masa Ogata, Kai Kunze, Masahiko Inami, and Maki Sugimoto. “Facial expression recognition in daily life by embedded photo reflective sensors on smart eyewear”. In: *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 2016, pp. 317–326.

- [130] Denys JC Matthies, Chamod Weerasinghe, Bodo Urban, and Suranga Nanayakkara. “Capglasses: Untethered capacitive sensing with smart glasses”. In: *Augmented Humans Conference 2021*. 2021, pp. 121–130.
- [131] Iain McCowan. “Microphone Arrays : A Tutorial”. In: 2001.
- [132] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. “librosa: Audio and music signal analysis in python”. In: *Proceedings of the 14th python in science conference*. Vol. 8. 2015, pp. 18–25.
- [133] S. Milanese, D. Marino, F. Stradolini, P. M. Ros, F. Pleitavino, D. Demarchi, and S. Carrara. “Wearable System for Spinal Cord Injury Rehabilitation with Muscle Fatigue Feedback”. In: *2018 IEEE SENSORS*. Oct. 2018, pp. 1–4. DOI: 10.1109/ICSENS.2018.8589763.
- [134] Mohammad Iman Mokhlespour Esfahani and Maury A Nussbaum. “Classifying diverse physical activities using “Smart Garments””. In: *Sensors* 19.14 (2019), p. 3133.
- [135] Fatemeh Mokhtari, Geoffrey M Spinks, Cormac Fay, Zhenxiang Cheng, Raad Raad, Jiangtao Xi, and Javad Foroughi. “Wearable electronic textiles from nanostructured piezoelectric fibers”. In: *Advanced Materials Technologies* 5.4 (2020), p. 1900900.
- [136] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. “IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–12.
- [137] Mehrab Bin Morshed, Samruddhi Shreeram Kulkarni, Koustuv Saha, Richard Li, Leah G Roper, Lama Nachman, Hong Lu, Lucia Mirabella, Sanjeev Srivastava, Kaya de Barbaro, et al. “Food, mood, context: Examining college students’ eating context and mental well-being”. In: *ACM Transactions on Computing for Healthcare* 3.4 (2022), pp. 1–26.
- [138] David Baeza Moyano, Daniel Arranz Paraiso, and Roberto Alonso González-Lezcano. “Possible effects on health of ultrasound exposure, risk factors in the work environment and occupational safety review”. In: *Healthcare*. Vol. 10. 3. MDPI. 2022, p. 423.
- [139] Sebastian Münzner, Philip Schmidt, Attila Reiss, Michael Hanselmann, Rainer Stiefelhagen, and Robert Dürichen. “CNN-based sensor fusion techniques for multimodal human activity recognition”. In: *Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 2017, pp. 158–165.
- [140] Marco Napoli. “Introducing Flutter and Getting Started”. In: Sept. 2019, pp. 1–23. ISBN: 9781119550860. DOI: 10.1002/9781119550860.ch1.

-
- [141] Pradyumna Narayana, J Ross Beveridge, and Bruce A Draper. “Continuous gesture recognition through selective temporal fusion”. In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2019, pp. 1–8.
- [142] Farhat Naseer, Ghufran Ullah, Muhammad Aleem Siddiqui, Muhammad Jawad Khan, Keum-Shik Hong, and Noman Naseer. “Deep Learning-Based Unmanned Aerial Vehicle Control with Hand Gesture and Computer Vision”. In: *2022 13th Asian Control Conference (ASCC)*. IEEE. 2022, pp. 1–6.
- [143] Kathiravan Natarajan, Truong-Huy D Nguyen, and Mutlu Mete. “Hand gesture controlled drones: An open source library”. In: *2018 1st International Conference on Data Intelligence and Security (ICDIS)*. IEEE. 2018, pp. 168–175.
- [144] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. “Survey on emotional body gesture recognition”. In: *IEEE transactions on affective computing* 12.2 (2018), pp. 505–523.
- [145] Michal Olszanowski, Grzegorz Pochwatko, Krzysztof Kuklinski, Michal Scibor-Rylski, Peter Lewinski, and Rafal Ohme. “Warsaw Set of Emotional Facial Expression Pictures: A validation study of facial display photographs”. In: *Frontiers in Psychology* 5 (Dec. 2014). DOI: 10.3389/fpsyg.2014.01516.
- [146] Francisco Javier Ordóñez and Daniel Roggen. “Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition”. In: *Sensors* 16.1 (2016), p. 115.
- [147] Bryce Osoinach. “Proximity capacitive sensor technology for touch sensing applications”. In: *Freescale White Paper* 12 (2007).
- [148] D. Ou, L. OuYang, Z. Tan, H. Mo, X. Tian, and X. Xu. “An electronic stethoscope for heart diseases based on micro-electro-mechanical-system microphone”. In: *2016 IEEE 14th International Conference on Industrial Informatics (INDIN)*. July 2016, pp. 882–885. DOI: 10.1109/INDIN.2016.7819285.
- [149] Chakradhar Pabba and Praveen Kumar. “An intelligent system for monitoring students’ engagement in large classroom teaching through facial expression recognition”. In: *Expert Systems* 39.1 (2022), e12839.
- [150] Sotirios Panagou, W Patrick Neumann, and Fabio Fruggiero. “A scoping review of human robot interaction research towards Industry 5.0 human-centric workplaces”. In: *International Journal of Production Research* (2023), pp. 1–17.
- [151] Jussi Parviainen, Jussi Kantola, and J Collin. “Differential barometry in personal navigation”. In: *2008 IEEE/ION Position, Location and Navigation Symposium*. IEEE. 2008, pp. 148–152.

- [152] Fotini Patrona, Ioannis Mademlis, and Ioannis Pitas. “An overview of hand gesture languages for autonomous UAV handling”. In: *2021 Aerial Robotic Systems Physically Interacting with the Environment (AIRPHARO)* (2021), pp. 1–7.
- [153] Jhonghe City Paul Yang. *Electronic stethoscope with piezo-electrical film contact microphone*. Patent No. US 2005/0157888A1, Filed Jan. 16, 2004, Issued Jul. 21, 2005. July 2005. URL: <https://patents.google.com/patent/US20050157888A1/en>.
- [154] Alexander Pavlosky, Jennifer Glauche, Spencer Chambers, Mahmoud Al-Alawi, Kliment Yanev, and Tarek Loubani. “Validation of an effective, low cost, Free/open access 3D-printed stethoscope”. In: *PLOS ONE* 13.3 (Mar. 2018). Ed. by David T. Eddington, e0193087. DOI: 10.1371/journal.pone.0193087. URL: <https://doi.org/10.1371/2Fjournal.pone.0193087>.
- [155] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [156] Laavanya Rachakonda, Arham Kothari, Saraju P Mohanty, Elias Kougianos, and Madhavi Ganapathiraju. “Stress-Log: An IoT-based smart system to monitor stress-eating”. In: *2019 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE. 2019, pp. 1–6.
- [157] B. Rafaely. “Spatial Sampling and Beamforming for Spherical Microphone Arrays”. In: *2008 Hands-Free Speech Communication and Microphone Arrays*. May 2008, pp. 5–8. DOI: 10.1109/HSCMA.2008.4538673.
- [158] Rajarajan Ramalingame, Rim Barioul, Xupeng Li, Giuseppe Sanseverino, Dominik Krumm, Stephan Odenwald, and Olfa Kanoun. “Wearable Smart Band for American Sign Language Recognition with Polymer Carbon Nanocomposite based Pressure Sensors”. In: *IEEE Sensors Letters* 5.6 (2021), p. 6001204.
- [159] Munawar A Riyadi, Norman Sudira, MH Hanif, and Aris Triwiyatno. “Design of pick and place robot with identification and classification object based on RFID using stm32vldiscovery”. In: *2017 International Conference on Electrical Engineering and Computer Science (ICECOS)*. IEEE. 2017, pp. 171–176.
- [160] Sebastian Ruder. “An overview of gradient descent optimization algorithms”. In: *arXiv preprint arXiv:1609.04747* (2016).
- [161] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. “Mobilenetv2: Inverted residuals and linear bottlenecks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

-
- [162] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva. “Automatic analysis of affective postures and body motion to detect engagement with a game companion”. In: *Proceedings of the 6th international conference on Human-robot interaction*. 2011, pp. 305–312.
- [163] J. Sanz-Robinson, L. Huang, T. Moy, W. Rieutort-Louis, Y. Hu, S. Wagner, J. C. Sturm, and N. Verma. “Large-Area Microphone Array for Audio Source Separation Based on a Hybrid Architecture Exploiting Thin-Film Electronics and CMOS”. In: *IEEE Journal of Solid-State Circuits* 51.4 (Apr. 2016), pp. 979–991. ISSN: 1558-173X. DOI: 10.1109/JSSC.2015.2501426.
- [164] Sanat Sarangi, Somya Sharma, and Bhushan Jagyasi. “Agricultural activity recognition with smart-shirt and crop protocol”. In: *2015 IEEE Global Humanitarian Technology Conference (GHTC)*. Seattle, WA, USA: IEEE, 2015, pp. 298–305.
- [165] Wataru Sato, Sylwia Hyniewska, Kazusa Minemoto, and Sakiko Yoshikawa. “Facial Expressions of Basic Emotions in Japanese Laypeople”. In: *Frontiers in Psychology* 10 (2019), p. 259. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.00259. URL: <https://www.frontiersin.org/article/10.3389/fpsyg.2019.00259>.
- [166] Martin Schepers, Matteo Giuberti, Giovanni Bellusci, et al. “Xsens mvn: Consistent tracking of human motion using inertial sensing”. In: *Xsens Technol* 1.8 (2018).
- [167] K. T. Selvan and R. Janaswamy. “Fraunhofer and Fresnel Distances : Unified derivation for aperture antennas.” In: *IEEE Antennas and Propagation Magazine* 59.4 (Aug. 2017), pp. 12–15. ISSN: 1558-4143. DOI: 10.1109/MAP.2017.2706648.
- [168] Nurettin Sezer and Muammer Koç. “A comprehensive review on the state-of-the-art of piezoelectric energy harvesting”. In: *Nano Energy* 80 (2021), p. 105567.
- [169] Sophia Shen, Xiao Xiao, Junyi Yin, Xiao Xiao, and Jun Chen. “Self-Powered Smart Gloves Based on Triboelectric Nanogenerators”. In: *Small Methods* 6.10 (2022), p. 2200830.
- [170] Xunbing Shen, Qi Wu, Ke Zhao, and Xiaolan Fu. “Electrophysiological evidence reveals differences between the recognition of microexpressions and macroexpressions”. In: *Frontiers in psychology* 7 (2016), p. 1346.
- [171] David A Shephard. “The 1975 Declaration of Helsinki and consent.” In: *Canadian Medical Association Journal* 115.12 (1976), p. 1191.
- [172] Sungtae Shin, Han Ul Yoon, and Byungseok Yoo. “Hand Gesture Recognition Using EGaIn-Silicone Soft Sensors”. In: *Sensors* 21.9 (2021), p. 3204.

- [173] Toshiki Shioiri, Toshiyuki Someya, Daiga Helmeste, and Siu Wa Tang. “Misinterpretation of facial expression: A cross-cultural study”. In: *Psychiatry and clinical neurosciences* 53.1 (1999), pp. 45–50.
- [174] Chamani Shiranthika, Nilantha Premakumara, Huei-Ling Chiu, Hooman Samani, Chathurangi Shyalika, and Chan-Yun Yang. “Human Activity Recognition Using CNN & LSTM”. In: *2020 5th International Conference on Information Technology Research (ICITR)*. IEEE. 2020, pp. 1–6.
- [175] Gil Shomron, Freddy Gabbay, Samer Kurzum, and Uri Weiser. “Post-training sparsity-aware quantization”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 17737–17748.
- [176] Ali I Siam, Atef Abou Elazm, Nirmeen A El-Bahnasawy, Ghada M El Banby, Abd El-Samie, and E Fathi. “PPG-based human identification using Mel-frequency cepstral coefficients and neural networks”. In: *Multimedia Tools and Applications* 80.17 (2021), pp. 26001–26019.
- [177] Nabeel Siddiqui and Rosa HM Chan. “Multimodal hand gesture recognition using single IMU and acoustic measurements at wrist”. In: *Plos one* 15.1 (2020), e0227039.
- [178] Gurashish Singh, Alexander Nelson, Ryan Robucci, Chintan Patel, and Nilanjan Banerjee. “Inviz: Low-power personalized gesture recognition using wearable textile capacitive sensor arrays”. In: *2015 IEEE international conference on pervasive computing and communications (PerCom)*. St. Louis, MO, USA: IEEE, 2015, pp. 198–206.
- [179] Sophie Skach, Rebecca Stewart, and Patrick GT Healey. “Smart arse: posture classification with textile sensors in trousers”. In: *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. Boulder, CO, USA: Association for Computing Machinery, 2018, pp. 116–124.
- [180] Kenneth D Skeldon, Lindsay M Reid, Vivienne McNally, Brendan Dougan, and Craig Fulton. “Physics of the Theremin”. In: *American Journal of Physics* 66.11 (1998), pp. 945–955.
- [181] Xingzhe Song, Kai Huang, and Wei Gao. “Facelistener: Recognizing human facial expressions via acoustic sensing on commodity headphones”. In: *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE. 2022, pp. 145–157.
- [182] M Sornam, Kavitha Muthusubash, and V Vanitha. “A survey on image classification and activity recognition using deep convolutional neural network architecture”. In: *2017 Ninth International Conference on Advanced Computing (ICoAC)*. IEEE. 2017, pp. 121–126.
- [183] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. “Dropout: a simple way to prevent neural networks from overfitting”. In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

-
- [184] Paul Stoffregen. *FreqCount*. https://www.pjrc.com/teensy/td_libs_FreqCount.html. Last accessed: February 06, 2024. Teensy, 2014.
- [185] Paul Stoffregen. *Teensy*. <https://www.pjrc.com/store/teensy41.html>. Last accessed: February 06, 2024. Teensy, 2020.
- [186] Min Su, Pei Li, Xueqin Liu, Dapeng Wei, and Jun Yang. “Textile-based flexible capacitive pressure sensors: A review”. In: *Nanomaterials* 12.9 (2022), p. 1495.
- [187] Sumarna, J. Astono, A. Purwanto, and D. K. Agustika. “The improvement of phonocardiograph signal (PCG) representation through the electronic stethoscope”. In: *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. Sept. 2017, pp. 1–5. DOI: 10.1109/EECSI.2017.8239099.
- [188] Tao Sun, Farita Tasnim, Rachel T McIntosh, Nikta Amiri, Dana Solav, Mostafa Tavakkoli Anbarani, David Sadat, Lin Zhang, Yuandong Gu, M Amin Karami, et al. “Decoding of facial strains via conformable piezoelectric interfaces”. In: *Nature biomedical engineering* 4.10 (2020), pp. 954–972.
- [189] Saiganesh Swaminathan, Jonathon Fagert, Michael Rivera, Andrew Cao, Gierad Laput, Hae Young Noh, and Scott E Hudson. “OptiStructures: Fabrication of Room-Scale Interactive Structures with Embedded Fiber Bragg Grating Optical Sensors and Displays”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.2 (2020), pp. 1–21.
- [190] Joy Swenson and Fred L Casmir. “The impact of culture-sameness, gender, foreign travel, and academic background on the ability to interpret facial expression of emotion in others”. In: *Communication Quarterly* 46.2 (1998), pp. 214–230.
- [191] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.
- [192] Dhruv Verma, Sejal Bhalla, Dhruv Sahnan, Jainendra Shukla, and Aman Parnami. “Expressear: Sensing fine-grained facial expressions with earables”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.3 (2021), pp. 1–28.
- [193] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature methods* 17.3 (2020), pp. 261–272.
- [194] Andrea Vitali. *Microphone array beamforming in the PCM and PDM domain*. https://www.st.com/resource/en/design_tip/dm00528068.pdf. DT0117, STMicroelectronics. Sept. 2018.
-

- [195] Biao Wang, Zhihe Long, Ying Hong, Qiqi Pan, Weikang Lin, and Zhengbao Yang. “Woodpecker-mimic two-layer band energy harvester with a piezoelectric array for powering wrist-worn wearables”. In: *Nano Energy* 89 (2021), p. 106385.
- [196] Chuyu Wang, Jian Liu, Yingying Chen, Lei Xie, Hong Bo Liu, and Sanclu Lu. “RF-kinect: A wearable RFID-based approach towards 3D body movement tracking”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.1 (2018), pp. 1–28.
- [197] H. Wang, L. Wang, Y. Xiang, N. Zhao, X. Li, S. Chen, C. Lin, and G. Li. “Assessment of elbow spasticity with surface electromyography and mechanomyography based on support vector machine”. In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. July 2017, pp. 3860–3863. DOI: 10.1109/EMBC.2017.8037699.
- [198] Liang Wang, Tao Gu, Xianping Tao, and Jian Lu. “Toward a wearable RFID system for real-time activity recognition using radio patterns”. In: *IEEE Transactions on Mobile Computing* 16.1 (2016), pp. 228–242.
- [199] Interactive Wear. *TWC24004B*. <https://www.interactive-wear.com/>. Last accessed: February 06, 2024. Interactive Wear, 2021.
- [200] Sheng Wei, Gu Dan, and Hu Chen. “Altitude data fusion utilising differential measurement and complementary filter”. In: *IET Science, Measurement & Technology* 10.8 (2016), pp. 874–879.
- [201] Nintendo Wii. *Nintendo Wii’s Rayman Raving Rabbids: TV Party - ShakeTV*. Nintendo. 2009. URL: <https://www.youtube.com/watch?v=eoxjA6E1mDs>.
- [202] Michael Woelfle and Willibald A Guenther. “Wearable RFID in order picking systems”. In: *RFID SysTech 2011 7th European Workshop on Smart Objects: Systems, Technologies and Applications*. VDE. 2011, pp. 1–6.
- [203] R. B. Woodward, S. J. Shefelbine, and R. Vaidyanathan. “Pervasive Monitoring of Motion and Muscle Activation: Inertial and Mechanomyography Fusion”. In: *IEEE/ASME Transactions on Mechatronics* 22.5 (Oct. 2017), pp. 2022–2033. ISSN: 1941-014X. DOI: 10.1109/TMECH.2017.2715163.
- [204] Richard B Woodward, Maria J Stokes, Sandra J Shefelbine, and Ravi Vaidyanathan. “Segmenting mechanomyography measures of muscle activity phases using inertial data”. In: *Scientific reports* 9.1 (2019), pp. 1–10.
- [205] H. Wu, Q. Huang, D. Wang, and L. Gao. “A CNN-SVM Combined Regression Model for Continuous Knee Angle Estimation Using Mechanomyography Signals”. In: *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. Mar. 2019, pp. 124–131. DOI: 10.1109/ITNEC.2019.8729426.

-
- [206] Yongling Wu, Yulin Ma, Hongyu Zheng, and Seeram Ramakrishna. “Piezoelectric materials for flexible and wearable electronics: A review”. In: *Materials & Design* 211 (2021), p. 110164.
- [207] Hao Xia, Xiaogang Wang, Yanyou Qiao, Jun Jian, and Yuanfei Chang. “Using multiple barometers to detect the floor location of smart phones with built-in barometric sensors for indoor positioning”. In: *Sensors* 15.4 (2015), pp. 7857–7877.
- [208] Wentao Xie, Qian Zhang, and Jin Zhang. “Acoustic-based Upper Facial Action Recognition for Smart Eyewear”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.2 (2021), pp. 1–28.
- [209] Z. F. Yang, D. K. Kumar, and S. P. Arjunan. “Mechanomyogram for identifying muscle activity and fatigue”. In: *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Sept. 2009, pp. 408–411. DOI: 10.1109/IEMBS.2009.5333666.
- [210] Minjeong Yoo, Yuseung Na, Hamin Song, Gamin Kim, Junseong Yun, Sangho Kim, Changjoo Moon, and Kichun Jo. “Motion estimation and hand gesture recognition-based human–UAV interaction approach in real time”. In: *Sensors* 22.7 (2022), p. 2513.
- [211] Yangguang Yu, Xiangke Wang, Zhiwei Zhong, and Yongwei Zhang. “ROS-based UAV control using hand gesture recognition”. In: *2017 29th Chinese Control And Decision Conference (CCDC)*. IEEE, 2017, pp. 6795–6799.
- [212] Matthew D Zeiler. “Adadelata: an adaptive learning rate method”. In: *arXiv preprint arXiv:1212.5701* (2012).
- [213] Gloria Zen, Lorenzo Porzi, Enver Sangineto, Elisa Ricci, and Nicu Sebe. “Learning personalized models for facial expression analysis and gesture recognition”. In: *IEEE Transactions on Multimedia* 18.4 (2016), pp. 775–788.
- [214] Qihang Zeng, Wei Xu, Changyuan Yu, Na Zhang, and Cheungchuen Yu. “Fiber-optic Activity Monitoring with Machine Learning”. In: *Conference on Lasers and Electro-Optics/Pacific Rim*. Hong Kong, China: Optical Society of America, 2018, W4K–5.
- [215] Haifeng Zhang, Wen Su, and Zengfu Wang. “Weakly supervised local-global attention network for facial expression recognition”. In: *IEEE Access* 8 (2020), pp. 37976–37987.
- [216] Yue Zhang, Jing Yu, Chunming Xia, Ke Yang, Heng Cao, and Qing Wu. “Research on GA-SVM Based Head-Motion Classification via Mechanomyography Feature Analysis”. In: *Sensors* 19.9 (2019). ISSN: 1424-8220. DOI: 10.3390/s19091986. URL: <https://www.mdpi.com/1424-8220/19/9/1986>.
-

- [217] Zhen Zhang, Kuo Yang, Jinwu Qian, and Lunwei Zhang. “Real-time surface EMG pattern recognition for hand gestures based on an artificial neural network”. In: *Sensors* 19.14 (2019), p. 3170.
- [218] Jingjing Zhao and Zheng You. “A shoe-embedded piezoelectric energy harvester for wearable sensors”. In: *Sensors* 14.7 (2014), pp. 12497–12510.
- [219] Enhao Zheng and Qining Wang. “Noncontact capacitive sensing-based locomotion transition recognition for amputees with robotic transtibial prostheses”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.2 (2016), pp. 161–170.
- [220] Bo Zhou, Daniel Geissler, Marc Faulhaber, Clara Elisabeth Gleiss, Esther Friederike Zahn, Lala Shakti Swarup Ray, David Gamarra, Victor Fortes Rey, Sungho Suh, Sizhen Bian, et al. “MoCaPose: Motion Capturing with Textile-integrated Capacitive Sensors in Loose-fitting Smart Garments”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7.1 (2023), pp. 1–40.
- [221] Bo Zhou, Tandra Ghose, and Paul Lukowicz. “Expressure: detect expressions related to emotional and cognitive activities using forehead textile pressure mechanomyography”. In: *Sensors* 20.3 (2020), p. 730.

Curriculum Vitae

I received a master's degree in electrical engineering and computer science from the University of Kaiserslautern in 2017. I am working as a researcher in the Department of Embedded Intelligence at DFKI with Prof. Paul Lukowicz. My main research interests are hardware and software co-designing multimodal sensor-based systems for human activity recognition.

Experience

- 2018–2024** **Researcher** DFKI
Embedded Intelligence Group
- 2018–2018** **Embedded Software Engineer** Nokia
High-Speed Communication Firmware Group
- 2016–2018** **Research Assistant** DFKI
Embedded Intelligence Group
- 2016–2016** **Research Assistant** University of Kaiserslautern
Department of Electrical Engineering
- 2015–2016** **Research Assistant** University of Kaiserslautern
Microelectronic Systems Design Research Group

Education

- 2019–2024** **PhD Candidate** German Research Center for Artificial Intelligence
- 2015–2017** **Master in Embedded System** University of Kaiserslautern
Masterarbeit: *Design, Implementation, and Evaluation of a Wearable Sensing System for Order Picking.*
- 2008–2015** **Diploma in Electronic Engineering** Simon Bolivar University
Bachelorarbeit: *Memory Access Measurements in Multi-core Architecture for WCET Estimation. University of Kaiserslautern*

Kaiserslautern, January 2, 2025

Hymalai Bello