Unimodal and Multimodal Sensor Fusion for Wearable Activity Recognition

Hymalai Bello*†
Email: hymalai.bello@dfki.de
Supervised by Paul Lukowicz*†, Bo Zhou *† and Sungho Suh *†
*German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany
†Department of Computer Science, RPTU Kaiserslautern-Landau, 67663 Kaiserslautern, Germany

***Abstract*—Combining different sensing modalities with multiple positions helps form a unified perception and understanding of complex situations such as human behavior. Hence, human activity recognition (HAR) benefits from combining redundant and complementary information (Unimodal/Multimodal). Even so, it is not an easy task. It requires a multidisciplinary approach, including expertise in sensor technologies, signal processing, data fusion algorithms, and domain-specific knowledge. This Ph.D. work employs sensing modalities such as inertial, pressure (audio and atmospheric pressure), and textile capacitive sensing for HAR. The scenarios explored are gesture and hand position tracking, facial and head pattern recognition, and body posture and gesture recognition. The selected wearable devices and sensing modalities are fully integrated with machine learning-based algorithms, some of which are implemented in the embedded device, on the edge, and tested in real-time.**

***Index Terms*—Multimodal Fusion, Multi-positional Fusion, Activity Recognition, Embedded Intelligence, TinyML**

I. Problem Statement

Human activities are highly context-dependent. The same activity can have a different meaning in different contexts. Its complexity is proportional to the combination of parallel cues from human gestures, movements, and spoken/unspoken body language. Different sensor positions and multiple sensing modalities help to form a unified perception and understanding of complex situations. And it takes into account the inherent variability of human behavior. Wearable devices are the most promising option for ubiquitous human activity recognition (HAR). In contrast, vision-based systems need to be deployed in the environment, which limits their ubiquity. Even so, exploiting multi-positional or multimodal information fusion is not straightforward. This is mainly due to the different nature of the sensors. Moreover, wearable-based HAR challenges include hardware-software-based solutions with size and user acceptance constraints. Addressing this challenge requires a multidisciplinary approach, including expertise in sensor technologies, signal processing, data fusion algorithms, and domain-specific knowledge. Using the most common wearable accessories on the market to deploy the HW/SW systems is one way to gain user acceptance. 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 body gestures and facial and head pattern recognition (see Fig. 1). The state-of-the-art has mainly focused on motion capture with IMUs deployed on elastic or tight-fitting clothing [1]. In the facial expression scenario, the authors of EarIo [2], present an earpiece to track facial expressions by sending 16-20 kHz sound waves around the face and tracking the deformation of the face. In food monitoring [3], the authors proposed a chewing detection system attached to a glasses' frame and employed IMU and piezoelectric sensing modalities.

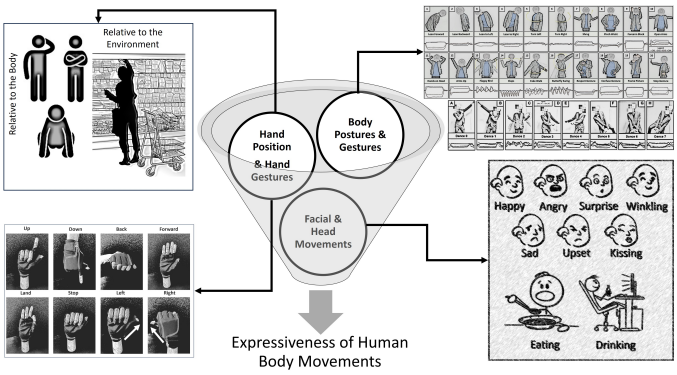

Fig. 1. Simplify Diagram of the Unspoken Expressiveness of Human Body Movements with Specific Example Scenarios Studied in Thesis.

II. Research Conducted and Future Plan

This section presents an overview of the contributions this work has made to the wearable community, as shown in Fig. 1. A comment on future direction is introduced.

Facial Movements: In [4], a novel idea is presented to detect facial muscle movements. Six stethoscope microphones are deployed over the face (positions compatible with smart glasses frames) to measure sound mechanomyography. To our knowledge, this was the first time that passive sound was used to detect subtle facial movements such as those associated with facial expressions. The work focuses on the fusion of unimodal and multipositional sound pressure. As a continuation of this idea, [5], presents a multimodal sensing alternative for monitoring facial muscle movement based on inertial, planar pressure, and acoustic sensors distributed in a minimally obstructive wearable accessory (sports cap). A modular multimodal fusion method is adopted in [5]. The fusion is based on sensor-dependent neural networks using a late fusion approach with a low memory footprint ($\leq$ 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). Facial muscle movement recognition was then combined with the detection of eating/drinking episodes in [6], [7]. The neural network-based models were deployed on-the-edge and evaluated in real-time for both scenarios (facial and food monitoring). In summary, sound mechanomyography and planar-pressure mechanomyography are combined and evaluated. Inertial information is also merged with mechanomyography modalities with the use of wearable devices (a helmet, sports cap, and glasses).

Body Gestures: In [8], differential atmospheric pressure (between two barometer sensors) and radio frequency identification (RFID) synchronization are fused to recognize the vertical position of the user's hand ($\leq$30 cm) with Naive Bayes Classifier. Two scenarios were evaluated. The first scenario is shelf level recognition (6 levels) and the second is up/down position recognition around the body (Head-Chest-Feet). A barometer is placed on the wrist with the RFID tag. A second barometer (as reference) is placed on a table in the first scenario (simulating an order-picking car), and in the pocket in the second scenario (simulating the position of a smartphone). The underlying idea is that each time the user's hand approaches the reference location (order picking car or pocket), the RFID reader detects it and the stationary barometer and wrist barometer readings are synchronized. Wrist elevation is tracked by comparing the signal from the wrist barometer to the reference. A comparison between unimodal, barometer only, and multimodal, barometer plus RFID synchronization, is presented. In [9], a formal jacket (loose-fitting garment) is transformed into a wearable theremin using two off-the-shelf theremin devices and the antennas were textile capacitive antennas. A dictionary of 20 upper body movements was classified with deep neural networks such as 1D-LeNet, DeepConvLSTM, and Conv2D. The idea behind this work is that different distances between body parts can describe different postures; thus, appropriately shaped antennas embedded in garments will result in specific frequency profiles. In [10] the work is extended from unimodal, capacitive sensing only, to capacitive and RFID fusion-based synchronization. The start-end points of the gestures were automatically segmented by a pair of RFID tag-reader, a tag in the pocket, and a reader on the wrist. When the hand moves away from the pocket, the starting point is marked. And, when the hand is back around the pocket, the endpoint is marked. After RFID-based segmentation, gestures were recognized in real-time by the model running on a PC. Textile capacitive antennas were also deployed in sports gloves to recognize real-time and on-the-edge gestures related to drone control in The CaptAinGlove [11]. The works proposed 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 model (linear acceleration). Then, using a capacitive model, the second stage recognizes 8 hand gestures to control drones. In Body Gesture recognition, a foreseen direction is the extension of the CaptAinGlove for a realistic Smart Factory scenario. Pairs of CaptAinGlove are used in a real factory setting to recognize activities such as walking, opening, working, and closing factory modules to monitor the work and safety of the workers. Another idea is to expand the tracking of upper body movements with barometers and IMUs distributed across the body. The idea is to use hardware built using bracelets and glasses to fuse pressure and inertial modalities in a realistic environment and with state-of-the-art fusion strategies.

III. Conclusion

The use of unimodal, multimodal, and multi-positional sensing modalities has shown potential for robust HAR models. However, it is not simple. The goal is to have HAR models gain a deeper understanding of the expressiveness of human body movements and capture when there is a benefit from multimodal or multi-positional information. This Ph.D. work evaluated the fusion of inertial, pressure-based (audio and atmospheric pressure), and textile capacitive sensing modalities for HAR in the context of hand position tracking, facial and head pattern recognition, and body posture and gesture recognition.

References

[1] H. T. Butt, M. Pancholi, M. Musahl, P. Murthy, M. A. Sanchez, and D. Stricker, "Inertial motion capture using adaptive sensor fusion and joint angle drift correction," in *2019 22th International Conference on Information Fusion (FUSION)*. IEEE, 2019, pp. 1–8.

[2] K. Li, R. Zhang, B. Liang, F. Guimbretière, and C. Zhang, "Eario: A low-power acoustic sensing earable for continuously tracking detailed facial movements," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–24, 2022.

[3] J. Shin, S. Lee, T. Gong, H. Yoon, H. Roh, A. Bianchi, and S.-J. Lee, "Mydj: Sensing food intakes with an attachable on your eyeglass frame," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–17.

[4] H. Bello, B. Zhou, and P. Lukowicz, "Facial muscle activity recognition with reconfigurable differential stethoscope-microphones," *Sensors*, vol. 20, no. 17, p. 4904, 2020.

[5] H. Bello, L. A. S. Marin, S. Suh, B. Zhou, and P. Lukowicz, "Inmyface: Inertial and mechanomyography-based sensor fusion for wearable facial activity recognition," *Information Fusion*, p. 101886, 2023.

[6] H. Bello, S. Suh, B. Zhou, and P. 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.

[7] H. Bello, S. Suh, B. Zhou, and L. Paul, "Meciface: Mechanomyography and inertial fusion based glasses for edge real-time recognition of facial and eating activities," *arXiv preprint arXiv:2306.13674*, 2023.

[8] H. Bello, J. Rodriguez, and P. Lukowicz, "Vertical hand position estimation with wearable differential barometery supported by rfid synchronization," in *EAI International Conference on Body Area Networks*. Springer, 2019, pp. 24–33.

[9] H. Bello, B. Zhou, S. Suh, and P. 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.

[10] H. Bello, B. Zhou, S. Suh, L. A. Sanchez Marin, and P. Lukowicz, "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*, vol. 4, p. 915280, 2022.

[11] H. Bello, S. Suh, D. Geißler, L. S. S. Ray, B. Zhou, and P. 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.