WiFi-based Person Identification Through Motion Analysis

Óscar Martins
CRACS/INESCTEC,
Instituto de Telecomunicações,
and Department of Computer Science
Faculty of Sciences, University of Porto
Porto, Portugal
oscar.g.martins@inesctec.pt

João P. Vilela
CRACS/INESCTEC, CISUC,
and Department of Computer Science
Faculty of Sciences, University of Porto
Porto, Portugal
jvilela@fc.up.pt

Marco Gomes
Instituto de Telecomunicações and
Department of Electrical
and Computer Engineering
University of Coimbra
Coimbra, Portugal
marco@co.it.pt

***Abstract*—By leveraging the advances in wireless communications networks and their ubiquitous nature, sensing through communication technologies has flourished in recent years. In particular, Human-to-Machine Interfaces have been exploiting WiFi IEEE 802.11 networks to obtain information that allows Human Activity Recognition. In this paper, we propose a classification model to perform Person Identification (PI) through Body Velocity Profile time series, obtained by combining Channel State Information containing gesture knowledge from multiple Access Points. Through this model, we investigate the impact of different gestures on PI classification performance and explore how informing the model about the input gesture can enhance classification accuracy. This information may enable the network to adjust to the absence of features capable of adequately characterizing the desired classes in certain gestures. A simplified stacking model is also presented, capable of combining the softmax outputs of K previously proposed individual models. By having the individual models' evaluations of a gesture and the gesture information relating to it, the number of gestures considered was shown to significantly improve the performance of the PI classification task. This enhancement increased 17% of the average F1 scores when compared to the individual model tested on the same data.**

***Index Terms*—Joint-Communication and Sensing, Human Activity Recognition, Person Identification**

I. Introduction

As wireless communications technologies progress and become more widespread new challenges emerge, including the persistent physical limitations of the electromagnetic spectrum. However, some opportunities also arise. A prominent technique in this regard is the Joint-Communication and Sensing (JCAS), also known as Integrated Sensing and Communications (ISAC), where sensing performance is included in the normal operation of wireless systems to enable multiple benefits [1], [2].

While there isn't a specific form of JCAS, various authors have proposed different modes for this technique, with each mode giving different weights to the communication and sensing segments. A common proposition has been the application of the IEEE 802.11 networks, i.e., WiFi, to sense the surroundings using Machine Learning (ML) models trained from data extracted from transmission metrics, such as the Received Signal Strength (RSS), Angle-of-Arrival (AoA), Channel State Information (CSI) or even beamforming feedback information [3] present in network packet data.

Through these metrics, a diverse array of environmental information can be estimated. Authors in [4]–[7] have shown that it is possible to count the number of people inside a room using communication-based sensing techniques with high accuracy. In other examples, works perform localization estimation [8], [9], recognition of performed activities [10], [11] and small gestures [12], [13], and even achieving people identification [14].

Human Activity Recognition (HAR) and Gesture Recognition (GR) have garnered particular interest due to their impact on Human-to-Machine Interface (HMI) systems which have seen significant progress in recent years. Person Identification (PI) also plays a crucial role in these systems, primarily by introducing soft access control layers. This involves identifying the person performing the action and gestures to block or permit the resulting action, allowing for multi-user input in parallel tasks. For example, in a scenario where two users are in the same room, a similar gesture may be used by one person to control the blinds while the other controls the television volume.

Most WiFi-based PI techniques have been focused on gait patterns, due to these being easily obtainable and recognizable metrics that are variable from person to person. However, such methods often fail to perform adequately in scenarios where individuals stand still while performing an HMI input gesture or perform a different activity other than walking.

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. This article/publication is based upon work from COST Action 6GPHYSEC (CA22168), supported by COST (European Cooperation in Science and Technology) and funded by The Science and Technology Development Fund, Macau SAR (File no. 0044/2022/A1). The authors also would like to acknowledge the FCT/MEC through national funds and when applicable co-funded by the European Regional Development Fund (FEDER), the Competitiveness and Internationalization Operational Programme (COMPETE 2020) of the Portugal 2020 framework, Regional OP Centro (POCI-01-0145-FEDER-030588) and Regional Operational Program of Lisbon (Lisboa-01-0145-FEDER-030588) and Financial Support National Public (FCT)(OE), under the projects UIDB/50008/2020, UIDP/50008/2020.

To this end, authors in [15] have proposed a framework capable of being trained over Body-coordinate Velocity Profiles (BVPs) [12], generated by combining CSI data obtained over multiple Access Points (APs). These enable simultaneous recognition of gestures and users while also enabling the recognition of new classes using transfer learning with a new set. The work proposed in [16] shows that WiFi-based Gesture Recognition (GR) and PI can be used simultaneously and made to work in real-time systems while taking inputs that require less computationally intensive processing as inputs to their proposed network.

An important characteristic to consider regarding JCAS is that, while it is described as a less intrusive sensing mechanism compared to a camera-based monitoring system, the ubiquitous nature of communication devices capable of implementing these capabilities can raise concerns regarding potential privacy and security issues. The concerns may stem from both unregulated and malicious usage of this technology. The ability to perform GR, PI, and passive user localization with the accuracy described in various works [10], [15], [16] can arguably compromise the privacy of users in the environment users as the presence of video feed. Moreover, due to the ubiquitous of wireless communication and regulation against these techniques, not only is a legitimate user of HMI affected but so are bystanders that happen to enter the sensed environment without prior knowledge.

This work analyzes the privacy impact of gesture recognition in person identification. In particular, the effect that apriori knowledge of the gesture performed has in identifying its performer and establishing an understanding of privacy and security concerns that can be raised or eased based on the usage of these sensing techniques.

This paper is organized as follows: Section II briefly presents the Methods. Section III explains the proposed model and methodology. Section IV presents the performance results. Section V concludes this paper.

II. Methods for Person Identification Through Motion Analyzes

One of the biggest challenges in JCAS-aided HAR and GR is the creation of domain-independent techniques. The same characteristic that makes most JCAS sensing methods rely on CSI also imposes a heavy restriction on these techniques. As the communication channel is heavily affected by the environment any changes, e.g., moving background objects, different persons being sensed, clothes, and even humidity in the air, can make it difficult to train generalist models. Thus, models using data from a given domain (person, environment, day) suffer a major performance loss when applied to other domains.

Authors in [10], [12] focus on this challenge by providing processing mechanisms capable of removing background effects in the collected CSI samples and creating a better-defining input for a given activity or gesture. While these techniques imply the loss of some user-specific information in trade for more generic and easily classifiable data, the analyses of the performance of a gesture classification network based on training data from different users made [12], show that while some users perform a gesture in a very standard way, others make distinctive enough movements such that the network is incapable of generalizing. Thus, it is expected that there still exists a high enough degree of correlation between a given gesture performed by the same user that can be used to identify them.

The goal of this work is to understand how the knowledge of the performed activity or gesture can improve the performance of a PI task and analyze how each gesture performs at this task. Thus, we evaluate the impact of allowing the end device to consider only the differences in each activity. Note that this assumption of knowledge can be deemed realistic and achieved under the application of a gesture classification framework, such as the one presented in [12], [13], where the activity is identified before the end, or in scenarios where a specific movement is expected from the user.

A. Gesture-based Person Identification

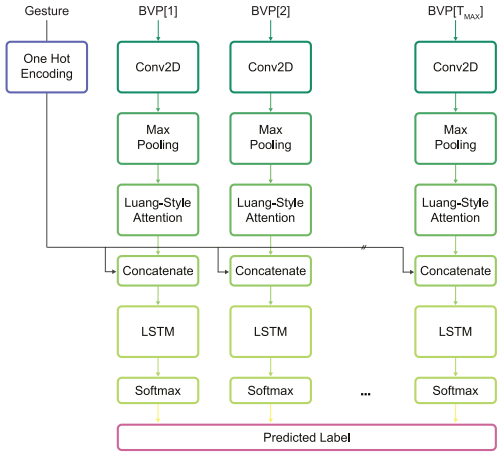
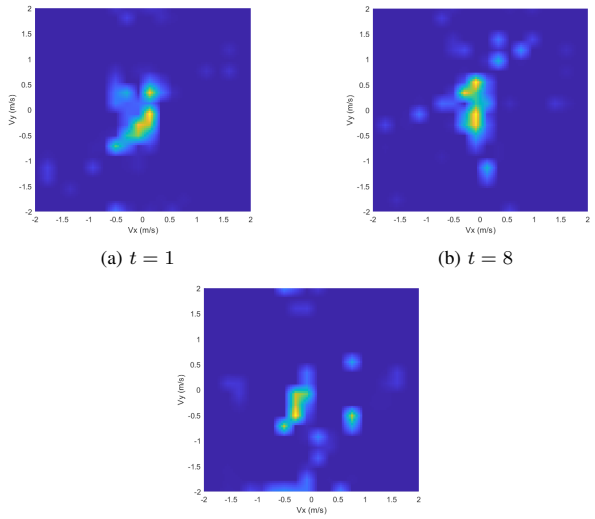


Fig. 1: Structure of the gesture performer identification model.

We establish the Neural Network (NN) defined in Fig 1 based on the BVP dataset to classify the user performing the gestures captured. The BVP is a two-dimensional vectorial representation of user movements with a gesture being characterized by a temporal series of BVPs, as seen in Fig 2. We use a CNN composed of a simple Conv2D layer and a Max Pooling Layer together with a bidirectional Long Short-Term Memory (LSTM) to deal with the spatial and temporal dimensions respectively. The output layer of the model is a fully connected layer with N output values, each representing one of the N possible users, given by the following softmax, $\sigma(x_i)$, activation function

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} \quad \text{for } i = 1, 2, \ldots, N \tag{1}$$



(a) $t = 1$ (b) $t = 8$

(c) $t = 15$

Fig. 2: A Body-coordinate Velocity Profile time series of a person performing a gesture of drawing a triangle at a frequency resolution of about 10 Hz. Based on [12].

to obtain a probability distribution for all the classes. The loss calculation is then given by the categorical cross-entropy function, such that

$$CE = -\log(\sigma(x_i)) = -\log\left(\frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}\right), \qquad (2)$$

where e^{x_i} represents standard exponent function of x_i vector.

As the input and objective of the network share some parallels with video classification tasks, we improve our model performance by adding a Luong-style attention [17] module to our model. The idea behind this addition is that some part of the movements and how they change timewise has more impact than others in defining the user performing a gesture. Thus, we use the attention module to stand out these features that better define the user and make the model rely more on them during the PI task.

We aim to analyze the capability PI performance through the performance of a gesture, with the usage of cross-domain techniques. Since we assume that the gesture being performed is known, we add this information as an input to the ML model expecting it to learn the different characteristics that differentiate the users based on the activity that is being used as input. To do this, the activities present in the training set are each attributed a value ranging from 1 to the number of activities considered, which in this case was six, and posteriorly converted into a vector through one-hot encoding. This vector is then added before the temporal modeling section by concatenating it with the result from the spatial feature extraction, as seen in the model diagram in Fig. 1.

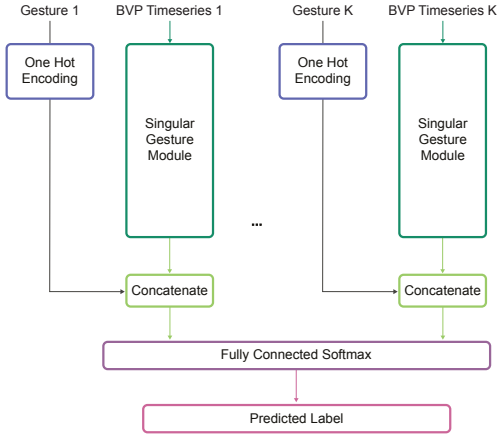


Fig. 3: Structure of the Person Identification stacking ensemble for K gestures.

B. Number of Gestures for Person Identification

Besides analyzing how the knowledge of the gesture affects the performance of the PI task, we also aim to understand the impact of considering multiple gestures in the performance of PI.

Our proposed analysis is based on two factors. Firstly, it is intuitive that some gestures hold more information significant to the PI task than others. Secondly, gestures for HMI are rarely done in an isolated form. Thus, to understand the impact of considering a higher number of gestures while training a network and using a model, we decided to analyze the improvement occurring when considering two cascaded gestures.

The model designed to study the impact of the number of gestures, *K*, is a stacking ensemble using the previously proposed model as individual models, herein called sub-models. Thus, it is simply a singular gesture model repeated *K* times with a fully connected *softmax* layer connecting the outputs of each singular gesture module and outputting a final result. Due to the first factor presented, the gesture information is once more concatenated to the output of each sub-model, allowing the final layer to adapt its weights to this information. The diagram for the final ensemble can be seen in Fig. 3.

III. Implementation

To evaluate the proposed model and methodology presented, we used a dataset composed of 6 activities and 4 users in up to 2 different environments [12] with each environment containing 5 different possible positions and orientations. In total, there were used 1375 samples of BVP time series (6 users × 5 positions × 5 orientations × 6 gestures × 5 to 20 instances). While the dataset created is balanced in relation to the gestures, it is imbalanced concerning the samples per user. A train-to-test ratio of 0.9 was used and the results presented herein were obtained through K-fold cross-validation with $K = 10$. Additionally, the length of the time series is defined

as $T_{MAX} = 30$ for all tests, as the BVP sequences are zero-padded to it. The PI classification model hyperparameters used are present in Table I.

TABLE I: Hyperparameters values for the PI classification model.

Loss Function	Categorical cross-entropy
Learning Rate	10^{-3}
Batch Size	32
Optimization Function	Adam [18]
Number of Epochs	30
Dropout Ratio	0.5
Conv2 Hidden Layers	16
Number of LSTM Hidden Layers	128

A. Datasets and Assumptions

This work is based on a public dataset that includes CSI data for both small motions, i.e. gestures, in various domains, i.e., multiple people, spaces, and days.

The dataset [19] contains CSI data from routers working with IEEE 802.11n WiFi standard and describing hand gestures obtained from 75 different domains through Linux CSI Tools, as well as, processed samples in the form of Doppler Frequency Shifts (DFS) and BVP. The authors apply 6 receivers to obtain data for people in 5 different positions and orientations to jointly apply the obtained data to recognize the performed gesture. All positions are in a 4 square meter area with the transmitter and receivers placed nearby outside this area (0.5 meters away from the borders) and in Line-of-Sight (LOS) to the user performing the gestures.

The computation of BVPs [12], produced through the mentioned work of the same users and used in this work, and based on this dataset share some limitations with it. First, since the gestures are significantly small-scale movements, LOS is required between the sensing target and the receivers to gather enough information to characterize the gesture. Similarly, it also has a strong requirement of knowing the position of transmitters, receivers, and even the torso to estimate the velocity profile associated with the small-scale movement properly. Regarding the number of receivers, while not all are simultaneously used when considering the positions of the targets and receivers it's safe to assume that for any given combination of position and orientation, the data obtained from at least four of these receivers is needed and being used. Additionally, it's relevant to notice that while BVPs can characterize gestures quite accurately, they require a considerable processing time, which makes them less adequate for real-time WiFI-based HAR than simpler inputs such as DFS spectrograms.

Our work is focused on the performance of people identification techniques associated with HAR and gesture recognition implementations. This is done under the assumption that, in the initial stage, the extracted CSI information from multiple APs is segmented into samples containing gestures being fully done and then joined together and processed into BVPs which are then used to predict the gesture being performed and the person doing it. Since the method we propose to study is based on the correlation between a gesture and its performer, we assume that gesture recognition is completed before the beginning of the PI pipeline, at both the training and testing stages, through methods such as the ones presented in [12], [13].

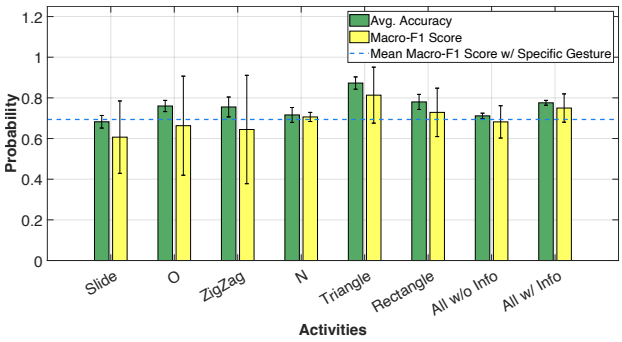


Fig. 4: Bar graphs with accuracy and macro-F1 scores for models trained and tested using: a specific drawing gesture, all gestures without gesture information, and all gestures with gesture information. The dashed horizontal line is the mean Macro-F1 score value for the models using a specific gesture. The error bars in the Macro-F1 scores represent the standard deviation obtained from the different classes evaluated.

The models we presented in this paper for PI are based on supervised learning, therefore it's important to notice that it cannot inherently deal with the presence of users that are not present in the training dataset.

IV. Evaluation

In this section, we analyze the results of PI based on single and multiple gestures. The results shown are presented using both accuracy and macro-F1 scores. The accuracy is the measure of all the correctly identified classifications calculated through

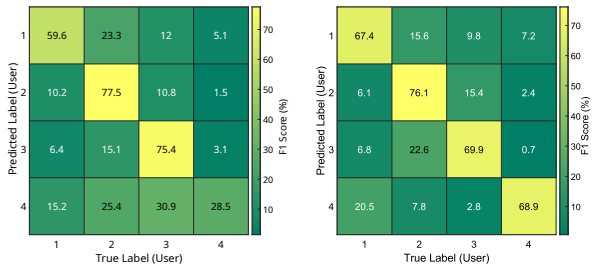
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

where the TP are true positives, TN are true negatives, FP are false positives and FN are false negatives. F1 scores represent the harmonic mean of Precision and Recall, thus presenting a better metric of both correct and incorrect cases, especially with imbalanced datasets. Their calculation goes as follows

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}. \tag{4}$$

A. PI Based on Single Gesture

The proposed single gesture model architecture in Fig. 1 was trained and tested on all 6 gestures together both with the gesture information, being passed to the model through one-hot vector encoding, and without it, aiming to study the performance improvement of this knowledge. Similarly, the model was also tested over every gesture separately and trained and tested solely on that single gesture, to observe the



(a) (b)

Fig. 5: Confusion matrices with F1 scores for the person identification model trained and tested over a specific gesture. (a) Trained with the drawing "Slide" gesture. (b) Trained with the drawing "N" gesture.

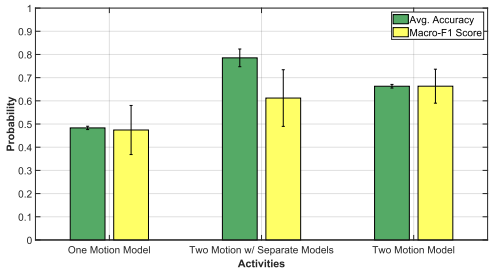


Fig. 6: Bar graphs with average accuracy and F1 scores for meta-models with different numbers of gestures. The two leftmost bars refer to a one-motion sub-model, the middle bars refer to a two-motion model where each 1-gesture sub-model was trained only in a specific gesture, and the two rightmost bars refer to a two-motion model with sub-model trained with all gestures. The standard deviation between multiple classes gives the error bars in the F1 Scores.

variability in performance between the gestures considered. These results can all be seen in Figures 4 and 5, respectively.

Fig. 4 presents the results for the various model configurations tested, with each two bars representing one of those, with the vertical axis indicating the probability associated with the accuracy and macro-F1 scores. It is noticeable that providing the gesture information to the network ("All Gestures w/ Info") slightly improves the PI task performance. Observing the results obtained from testing over the different gestures separately allows us to understand the reason for this improvement. In fact, from the confusion matrices in Fig. 5, one can see how some gestures provide better F1 scores over all classes, i.e. all users considered, while others generally fail to classify some users, i.e. presenting a higher classification bias. Notice that, while some of this behavior could be attributed to the imbalance in the dataset, since all gestures have the same number of samples for every user, this difference should not be substantially smaller or bigger in a given gesture, if not for the inherent characteristics of this gesture. Thus, by knowing the gesture being performed the network's weights can be better trained to deal with such cases, providing a boost in performance.

Regarding the overall effectiveness of the network trained over a specific gesture, it can be seen that some of the gestures allow for the model to achieve a much better performance than the one achieved by relying on all gestures, both with and without gesture information. However, by accounting that any given gesture has the same probability of being performed, the overall accuracy of the PI classification is best when a single model is trained with the added gesture information.

B. Number of Gestures for Person Identification

To evaluate the accuracy performance of increasing the number of gestures considered, we trained a model, i.e. a single fully connected output layer with a *softmax* activation function, fed with the output from two 1-gesture classification sub-models. The sub-models studied were both trained with a single specific gesture and trained with all gestures simultaneously. Similarly to stacking ensembles, the sub-models here are only being tested on new data and the new layers are being trained on the outputs of these. The results of this experiment can be seen in Fig. 6.

As can be observed by comparing the one-motion model to the two-motion model with sub-models trained by all gestures, increasing the number of gestures taken into account appears to result in a relevant performance improvement over the sub-models classification, when the proposed multiple-gesture model is used. In this case, with a simple base model, the accuracy and average F1 score show improvements close to 17%.

When the two-motion model uses sub-models trained for a specific gesture, the accuracy increases even higher, however, it also leads to a lower F1 score and relatively high variance between the different considered classes due to the higher disparity in F1 score in classes on the sub-models used, as shown in the previous subsection. Additionally, since the two sub-models used are trained on different datasets, the output combining layer needs to be trained for every combination of motions, which brings a much higher computational burden.

V. Future Work

Although the proposed models allowed for an improvement in the PI classification task, the presented final results could be further enhanced. A limitation of the proposed work is its limitation to using only BVPs as model inputs. Thus we lack the understanding of how this knowledge would improve considering other inputs, such as DFS or denoised CSI images as proposed in [20]. The performance of the number of gestures considered should also be further evaluated adding the computational cost of training a stacking ensemble over this number, especially on outputs of different complexities.

VI. Conclusion

This article proposed a Machine Learning model for Person Identification which was used to explore the impact of knowing the performed gestures and the number of gestures in the privacy of the user under a JCAS sensing environment.

It was shown that a system could benefit from having different models trained for each single gesture. However, a singular model trained for all gestures and fed with information of the gesture being inputted is expected to perform better under equiprobable gestures. Using a simplified stacking ensemble, we showed that the accuracy of the PI result can be improved by increasing the number of gestures being evaluated simultaneously.

The proposed design was shown to have some limitations concerning it, namely regarding the dataset used, due to the extensive computational power and receivers needed. Furthermore, the proposed work evaluated the impact of knowing the gesture being done, thus for a real-world application, it requires the usage of a gesture recognition module beforehand. Nevertheless, this could be improved by considering a dual-task classification model where gesture and user are classified in parallel with a combined loss.

References

[1] Z. Behdad, Ö. T. Demir, K. W. Sung, E. Björnson, and C. Cavdar, "Power allocation for joint communication and sensing in cell-free massive mimo," in *IEEE Global Communications Conference*, 2022.

[2] Z. Zhou, Z. Yang, C. Wu, W. Sun, and Y. Liu, "Lifi: Line-of-sight identification with wifi," in *Proceedings of IEEE Int. Conf. on Computer Communications*, 2014, pp. 2688–2696.

[3] K. F. Haque, F. Meneghello, and F. Restuccia, "Wi-bfi: Extracting the ieee 802.11 beamforming feedback information from commercial wi-fi devices," in *Proceedings of the 17th ACM Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization*, Madrid, Spain: Association for Computing Machinery, 2023.

[4] I. Sobron, J. Del Ser, I. Eizmendi, and M. Vélez, "Device-free people counting in iot environments: New insights, results, and open challenges," *IEEE Internet of Things Journal*, 2018.

[5] Y. Yang, J. Cao, X. Liu, and X. Liu, "Wi-count: Passing people counting with cots wifi devices," in *2018 27th International Conference on Computer Communication and Networks*, 2018.

[6] F. Wang, F. Zhang, C. Wu, B. Wang, and K. J. Ray Liu, "Passive people counting using commodity wifi," in *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, 2020.

[7] C. Tang, W. Li, S. Vishwakarma, K. Chetty, S. Julier, and K. Woodbridge, "Occupancy detection and people counting using wifi passive radar," in *2020 IEEE Radar Conference (RadarConf20)*, 2020.

[8] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "Spotfi: Decimeter level localization using wifi," *SIGCOMM Comput. Commun. Rev.*, 2015.

[9] M. Abbas, M. Elhamshary, H. Rizk, M. Torki, and M. Youssef, "Wideep: Wifi-based accurate and robust indoor localization system using deep learning," in *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*, 2019.

[10] F. Meneghello, D. Garlisi, N. D. Fabbro, I. Tinnirello, and M. Rossi, "Sharp: Environment and person independent activity recognition with commodity ieee 802.11 access points," *IEEE Transactions on Mobile Computing*, 2023.

[11] H. Yan, Y. Zhang, Y. Wang, and K. Xu, "Wiact: A passive wifi-based human activity recognition system," *IEEE Sensors Journal*, vol. 20, 2020.

[12] Y. Zhang, Y. Zheng, K. Qian, *et al.*, "Widar3.0: Zero-effort cross-domain gesture recognition with wi-fi," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[13] H. Abdelnasser, M. Youssef, and K. A. Harras, "Wigest: A ubiquitous wifi-based gesture recognition system," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, 2015.

[14] J. Ding, Y. Wang, and X. Fu, "Wihi: Wifi based human identity identification using deep learning," *IEEE Access*, 2020.

[15] C. Li, M. Liu, and Z. Cao, "Wihf: Gesture and user recognition with wifi," *IEEE Transactions on Mobile Computing*, 2022.

[16] R. Zhang, C. Jiang, S. Wu, Q. Zhou, X. Jing, and J. Mu, "Wi-fi sensing for joint gesture recognition and human identification from few samples in human-computer interaction," *IEEE Journal on Selected Areas in Communications*, 2022.

[17] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980.

[19] Z. Yang, Y. Zhang, G. Zhang, Y. Zheng, and G. Chi, *Widar 3.0: Wifi-based activity recognition dataset*, 2020. DOI: 10.21227/7znf-qp86.

[20] Y. Gu, X. Zhang, Y. Wang, *et al.*, "Wigrunt: Wifi-enabled gesture recognition using dual-attention network," *IEEE Transactions on Human-Machine Systems*, 2022.