TOWARDS COLLABORATIVE MULTIMODAL FEDERATED LEARNING FOR HUMAN ACTIVITY RECOGNITION IN SMART WORKPLACE ENVIRONMENTS

Seyed Mohammad Sheikholeslami[1], *Pai Chet Ng*[1], *Huan Liu*[2], *Yuanhao Yu*[2], *Konstantinos N. Plataniotis*[1]

[1] Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto
[2] Huawei Noah's Ark Lab, Canada

ABSTRACT

This paper aims to improve human activity recognition (HAR) with multimodal data across multiple consumer devices in a smart workplace environment. By leveraging the sensor-rich capabilities of smartphones, smartwatches, and smart speakers, we propose Collaborative Multimodal Federated Learning (CoMFL) algorithm to facilitate efficient feature encoding on lightweight local models implemented on consumer devices while fusing these encoded features for training a super model on a personalized local server, all within the private zone of a user. Federated learning aggregates model updates across users without compromising privacy, resulting in a generalized global super model. Additionally, we address the challenge of missing modality by incorporating a feature reconstruction network. This network attempts to reconstruct missing modalities prior to feature fusion, improving performance when dealing with missing features. Our proposed CoMFL achieves significant performance gains with multimodal HAR systems.

Index Terms— Human activity recognition, multimodal machine learning, federated learning,

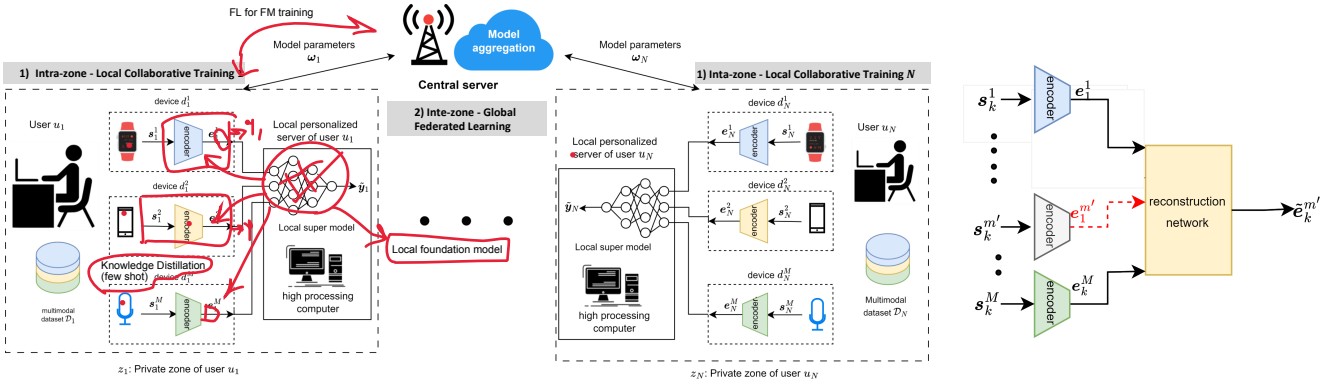
1. INTRODUCTION

Human activity recognition (HAR) has witnessed rapid advancement with the proliferation of consumer devices embedded with heterogeneous sensors that capture multimodal data providing richer contextual information for HAR in real-world scenarios. While deep learning methods have been widely used in HAR [1, 2], most of the recent works rely on gathering data in one silo for training which violates the privacy of users. With the growing concern for data privacy and the limitations of centralized training, federated learning (FL) [3] emerges as an attractive solution widely applied to HAR.

Previous research has mostly focused on exploiting either unimodal data (e.g., inertial sensor [4, 5]) or multimodal data on a single device (e.g., inertial + physiological signals [6, 7], inertial + wifi [8, 9]) for HAR. While these research works have demonstrated promising results, they fail to capture the full complexity of human activities, as different devices offer unique perspectives on user behavior. Intuitively, leveraging multimodal data fusion has the potential to enhance the accuracy and robustness of HAR models, especially in the context of complex human activities in a smart environment. However, implementing a multimodal HAR system across heterogeneous devices within smart workplace environments presents several challenges: 1) effective fusion of multimodal data to build a unified model for all modalities is challenging particularly due to cross-modal data heterogeneity and the asynchronous nature of data sampling across devices, and 2) inaccessibility to diverse data from different users degrades the generalization performance of the HAR model.

Motivated by the above challenges, this paper proposes **Collaborative Multimodal Federated Learning (CoMFL)** algorithm to address the challenges in implementing a multimodal HAR system across multiple users each owning multiple consumer devices in smart workplace environments. As illustrated in Fig. 1(a), CoMFL operates in two zones: 1) Intra-zone training consists of multiple devices within the same user's private zone working collaboratively to train a "super-local model" (in short super model) on a local personalized server, facilitating efficient processing of multiple modalities. 2) Inter-zone training orchestrating through the central server to learn a global model by aggregating the super models, leveraging collective knowledge from participated users. While federated learning holds promise for preserving data privacy, it can encounter issues such as missing modalities resulting from communication failures. To address these missing modalities, we introduce a feature reconstruction network that works on the local personalized server. This network adapts to reconstruct missing features, ensuring a seamless reconstruction of the missing modalities before feature fusion for accurate activity recognition.

Contributions of this paper are: **[C1]** the proposed CoMFL framework enables resource-efficient collaboration of multiple devices with different sensing inputs belonging to the same user to train a multimodal HAR for smart environments. Additionally, the cross-user cooperation of different users enriches the global model without sharing their private multimodal data, **[C2]** we design a robust feature reconstruction network to adaptively reconstruct the missing data. The capability of reconstructing the missing features prior to feature fusion and HAR significantly improves model performance in the event of missing modality, and **[C3]**



(a) (b)

Fig. 1: (a) Our proposed CoMFL consists of Intra-zone for **local collaborative training** between local devices enabling efficient processing of multiple modalities from multiple devices, and Inter-zone for facilitating **global federated learning**, (b) modality reconstruction network to reconstruct the feature (indicated by the red dashed line) lost due to the communication failure.

we evaluate CoMFL by conducting a comparative analysis between unimodal and multimodal approaches within both centralized and distributed settings. This evaluation takes into account potential missing features resulting from communication failures. Our experimental results indicate significant performance gain of CoMFL in addressing the HAR problem in smart workplace environments.

2. RELATED WORK

We categorize the previous works into unimodal FL for HAR [10, 11, 12], and multimodal FL for HAR [13, 14].

Unimodal FL for HAR: The earlier works on HAR using FL, such as [15, 10], showed that FedAvg [3] outperforms alternatives such as [16] and [17]. [11] uses meta-learning to reduce the heterogeneity of clients by training an embedding network. Authors in [18] use knowledge distillation to compress a model in the server for HAR in asynchronous FL. [12] proposes a similarity-aware FL framework for increasing the accuracy and reducing the communication overhead.

Multimodal FL for HAR: The proposed algorithm in [13] uses FL for privacy support in multimodal HAR based on wireless signals. Authors in [14] consider multimodal data for HAR using FL, and fuse all data modalities in a single vector before fitting them to their proposed deep learning model. However, these works consider multimodal data on a single device. In addition, as the number of modalities increases, early fusion can lead to a substantial escalation in both the dimensionality and complexity of the models.

3. SYSTEM MODEL AND PROPOSED COMFL

This section presents the adapted system model introducing private zones with multiple devices surrounding a user and then our proposed CoMFL is described.

3.1. Private Zone's System Model

Considering N users, indexed by $\mathbb{U} = \{u_1, ..., u_N\}$, we define a *private* zone z_k for each user u_k, composed of a set of M smart devices denoted by $\mathbb{D}_{u_k} = \{d_k^1, ..., d_k^M\}$ with different set of sensing inputs $\boldsymbol{s}_k^1, ..., \boldsymbol{s}_k^M$. Each z_k is assumed to have a local personalized server with high communication and computation capabilities, such as a personal computer. We posit that the exchange of data among distinct devices within the individual private zone of u_k remains viable without compromising privacy. Specifically, each user u_k has access to a multimodal local dataset denoted by $\mathcal{D}_{u_k} = (\boldsymbol{x}_k, \boldsymbol{y}_k)$, where $\boldsymbol{x}_k = (\boldsymbol{s}_k^1, ..., \boldsymbol{s}_k^M)$ denotes the multimodal input captured from different devices and $\boldsymbol{y}_k$ is the corresponding activity label.

Due to privacy limitations, the data transfer across separate private zones is deemed impermissible. Accordingly, we leverage FL to enable global model learning across different private zones, leading to our proposed CoMFL algorithm.

3.2. Collaborative Multimodal Federated Learning

Our proposed CoMFL consists of two training zones: 1) intra-zone training, and 2) inter-zone training. Additionally, we introduce a feature reconstruction network to address missing features resulting from communication failures.

3.2.1. Intra-zone training

Leveraging multimodal data distributed among M devices in a z_k, we devise a collaborative training strategy, employing modality-specific encoders and a common classifier as depicted in Fig.1(a). Encoders are strategically positioned on local devices to extract features from raw modalities, while the super model is hosted on the locally personalized device, endowed with substantial computational capacity.

Feature extraction: At user u_k, a distinct encoders is assigned for each modality m to extracting salient features E_k^m that maps the vectorized sensory inputs $\boldsymbol{s}_k^m \in \mathbb{R}^{|S_k^m|}$ to a feature representation $\boldsymbol{e}_k^m \in \mathbb{R}^{e^m}$.

Feature fusion: We employ an intermediate fusion to amalgamate the encoded features across all modalities, resulting

in a concatenated feature vector $\boldsymbol{e}_k = [\boldsymbol{e}_k^1, ..., \boldsymbol{e}_k^M]$. Subsequently, this feature vector $\boldsymbol{e}_k$ serves as the input for the activity classifier c_k to obtain the estimated activity class $\tilde{\boldsymbol{y}}_k$

Proposed feature reconstruction: As discussed, smart devices may fail to transmit encoded features to the personalized local server due to communication failure. To address this challenge, we aim to reconstruct the missing feature given the available modalities. Accordingly, each user applies unsupervised learning to train a feature reconstruction network, as shown in Fig. 1(b). Let m' denote the missing modality in the inference phase. We use the pre-trained encoders for the available modalities $m = 1, ..., M, m \neq m'$ from Fig. 1(a) as the backbone of a reconstruction network. In this regard, the local loss function for feature reconstruction at each user u_k is given as

$$\mathcal{L}_k^{MR} = \sum_{i=1}^{|\mathcal{D}_{u_k}|} L(\tilde{\boldsymbol{e}}_k^{m'}[i], \boldsymbol{e}_k^{m'}[i]) \tag{1}$$

where $\tilde{\boldsymbol{e}}_k^{m'}[i]$ denotes the i^{th} reconstructed features, and $L(\cdot)$ is the distance between the estimated and actual modality, e.g., mean squared error.

Local update: Representing the local model of user u_k parameterized by $\boldsymbol{\omega}_k$ as $f_k(., \boldsymbol{\omega}_k)$, goal of each personalized local server is to determine the weight parameters $\boldsymbol{\omega}_k$ that minimize the loss function over the local data $\mathcal{D}u_k$. Accordingly, the local personalized server at user u_k tackles the following optimization problem:

$$\min_{\boldsymbol{\omega}_k} \quad \mathcal{L}_k = \sum_{i=1}^{|\mathcal{D}_{u_k}|} \ell(\tilde{\boldsymbol{y}}_k[i], \boldsymbol{y}_k[i])), \tag{2}$$

where $\tilde{\boldsymbol{y}}_k[i] = f_k(\boldsymbol{x}_k[i], \boldsymbol{\omega}_k)$ represents the predicted activity for data sample i of user k, and $\ell(.)$ denotes a metric to quantify the dissimilarity between the predicted and actual activity.

3.2.2. Inter-zone training

After updating the local super models, the personalized servers cooperate with a central server to construct a global model. This will enrich the HAR model leading to more data diversity and generalization. In this regard, the objective of the global server is to update the global weight parameters $\boldsymbol{\omega}$ by minimizing the weighted sum of local loss functions defined as $\mathcal{L}^G(\boldsymbol{\omega}) = \sum_{k=1}^{N} p_k \mathcal{L}_k$, where p_k is the importance weight of user k chosen based on the number of data samples at each user. Due to inaccessibility to private local datasets, we use vanilla Federated Averaging (FedAvg) [3] algorithm to update the model parameters as

$$\boldsymbol{\omega}^{t+1} = \sum_{k=1}^{N} p_k \boldsymbol{\omega}_k^t \tag{3}$$

where superscript t designates the index of the communication round. The updated model parameters are sent back to users for further training. CoMFL iterates multiple intra-zone and inter-zone updates until the FL model achieves convergence.

Table 1: List of activities in the multimodal HAR dataset.

Simple activities	Complex activities
(a1) walking around	**(a8)** walking and talking on the phone
(a2) talking on the phone	
(a3) reading book	
(a4) stretching	
(a5) writing	**(a9)** writing and talking on the phone
(a7) browsing the monitor	
(a6) typing on the keyboard	

4. EXPERIMENTAL RESULTS

In this section, our objective is to evaluate the performance of the proposed CoMFL framework. We first introduce the simulation setup, encompassing the multimodal HAR dataset and the employed network architecture. Next, we compare uni-modal and multimodal scenarios in both centralized and distributed learning settings. Then, we evaluate the performance when one of the input modalities is missing.

4.1. Simulation setup

Dataset: We examine a scenario involving $N = 28$ individuals engaged in 9 distinct activities. This data was collected through our data collection campaign, with the informed consent of all participants. Table 1 shows the 7 simple activities and 2 complex activities performed by the participants. Notably, the complex activities involve a combination of any two simple activities (e.g., **(a8)** is a combination of **(a1)** and **(a2)**), which adds complexity to the HAR dataset, making it challenging to classify. Each user possesses $M = 3$ separate devices: a smartphone, a smartwatch, and a smart speaker. Notably, among these users, a subset of 20 individuals referred to as "active users", possess annotated data and actively participate in the training phase. We randomly select 20% of the local data set of active users as local test sets and keep the remaining 80% as training data. The remaining 8 users (referred to as "passive users") serve the purpose of evaluating the cross-person generalization of trained models.

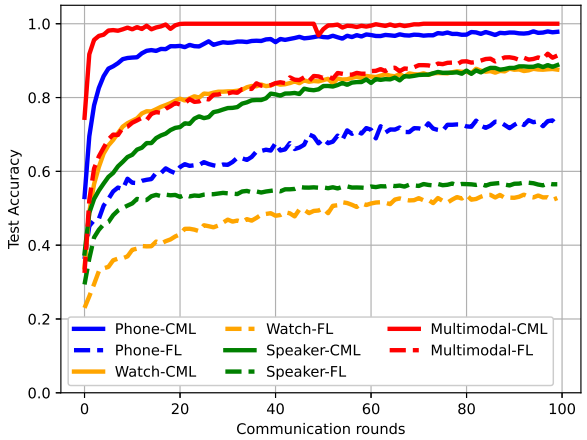
Network: For each device, we use 2 CNN layers with 64 and 32 and filters of size 3×3 and max-pooling 2×2. The super model is composed of two fully-connected layers with 64 and 32 neurons, respectively.

4.2. Performance evaluation

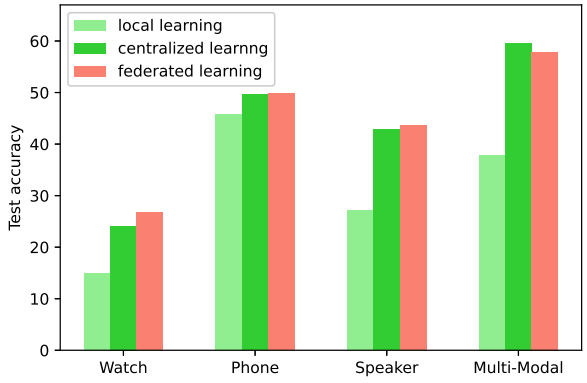
We aim to compare the unimodal and multimodal HAR training for both centralized and distributed scenarios. In the centralized setup, all individual datasets are aggregated on a central server for model training. Conversely, within the proposed CoMFL framework, outlined in Section 4, model

Table 2: Performance of the multimodal HAR models with missing modality.

training strategy \ modality	missing modality						full modality
	watch		phone		speaker		
	baseline	proposed	baseline	proposed	baseline	proposed	
MM centralized learning model	83.21%	90.83%	38.11%	46.17%	94.54%	98.14%	99.29%
MM federated learning model	74.03%	78.44%	28.33%	54.26%	79.21%	85.79%	91.37%



(a)



(b)

Fig. 2: (a) Test accuracy versus number of communication rounds/epochs in distributed and centralized settings, and (b) cross-person test accuracy of trained models.

parameters are exchanged with the server, ensuring data privacy without direct data sharing. Taking the aforementioned consideration into account, 2(a) illustrates the average accuracy over the local test sets versus the number of communication rounds in centralized and FL settings for unimodal and multimodal.

Unimodal HAR: As it can be seen from Fig. 2(a), among the unimodal models, phone modality reaches the highest test accuracy in both centralized and distributed settings with 97.83% and 72.46% test accuracy, respectively. Additionally, the smartwatch model performs better than the model trained with the smart speaker in the centralized setting with 83.29% and 59.59% test accuracy. Conversely, in the FL setting, the smart speaker model performs slightly better than the watch model with 56.50% and 52.78%, respectively. This observation highlights that the data heterogeneity in smart speakers is relatively lower compared to that of smartwatches.

Multimodal HAR: Fig. 2(a) shows that the multimodal HAR algorithm achieves higher performance compared to all unimodal settings. The test accuracy for the centralized setting is 99.29% compared to 91.37% of the FL setting.

Cross-person generalization: FL facilitates enhancing data diversity without the need to share private data. In this regard, we aim to compare the cross-person generalization of the proposed CoMFL with centralized and local training by evaluating the trained models over the test set from passive users. In this regard, Fig. 2(b) demonstrates the cross-person test accuracy with unimodal and multimodal data in (i) local training, (ii) centralized training, and (iii) distributed training. As can be seen from the figure, the proposed CoMFL achieves comparable cross-person generalization with the centralized setup whereas local training has low generalization. The main explanation for this is that CoMFL, similar to the centralized setup, makes use of the various data types available on all user devices while maintaining data privacy. Furthermore, compared to unimodal scenarios, multimodal data exhibits stronger generalization, emphasizing the fact that multimodal data contains more information for HAR.

Missing modalities: We consider a scenario where one of the modalities is missing due to hardware or communication failure. In this case, the network structure in Fig. 1(a) fails to deliver an accurate decision, and hence we use Fig. 1(b) to reconstruct the missing modality. Table 2 reports the performance of the multimodal model when one of the modalities is missing for two cases: (i) baseline where we replace the missing modality with zeros, and (ii) the missing data is reconstructed using the proposed method. As can be seen from the table, phone modality has the highest effect on the model performance. However, in practice, we observed that the smartwatch usually fails to provide its captured data.

5. CONCLUSION

By efficiently leveraging the sensor capabilities of smartphones, smartwatches, and smart speakers, CoMFL enables privacy-preserving feature encoding and fusion. The incorporation of federated learning and feature reconstruction further enhances performance, even in the presence of missing data. CoMFL demonstrates significant improvements in multimodal HAR systems, showcasing its potential for future applications in privacy-preserving, smart workplace environments.

6. REFERENCES

[1] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos, "Multimodal human action recognition in assistive human-robot interaction," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2702–2706.

[2] Yong Li and Luping Wang, "Human activity recognition based on residual network and bilstm," *Sensors*, vol. 22, no. 2, 2022.

[3] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Aarti Singh and Jerry Zhu, Eds. 20–22 Apr 2017, vol. 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, PMLR.

[4] Hamza Ali Imran, "Ultanet: An antithesis neural network for recognizing human activity using inertial sensors signals," *IEEE Sensors Letters*, vol. 6, no. 1, pp. 1–4, 2022.

[5] Ali Nouriani, Alec Jonason, James Jean, Robert McGovern, and Rajesh Rajamani, "System-identification-based activity recognition algorithms with inertial sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 7, pp. 3119–3128, 2023.

[6] Dae Yon Hwang, Pai Chet Ng, Yuanhao Yu, Yang Wang, Petros Spachos, Dimitrios Hatzinakos, and Konstantinos N. Plataniotis, "Hierarchical deep learning model with inertial and physiological sensors fusion for wearable-based human activity recognition," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 21–25.

[7] Narit Hnoohom, Sakorn Mekruksavanich, and Anuchit Jitpattanakul, "Physical activity recognition based on deep learning using photoplethysmography and wearable inertial sensors," *Electronics*, vol. 12, no. 3, 2023.

[8] Muhammad Muaaz, Ali Chelli, Ahmed Abdelmonem Abdelgawwad, Andreu Català Mallofré, and Matthias Pätzold, "Wiwehar: Multimodal human activity recognition using wi-fi and wearable sensing modalities," *IEEE Access*, vol. 8, pp. 164453–164470, 2020.

[9] Valentina Bianchi, Marco Bassoli, Gianfranco Lombardo, Paolo Fornacciari, Monica Mordonini, and Ilaria De Munari, "Iot wearable sensor and deep learning: An integrated approach for personalized human activity recognition in a smart home environment," *IEEE Internet of Things Journal*, vol. 6, no. 5, 2019.

[10] Sannara Ek, François Portet, Philippe Lalanda, and German Vega, "Evaluation of federated learning aggregation algorithms: Application to human activity recognition," New York, NY, USA, 2020, UbiComp-ISWC '20, p. 638–643, Association for Computing Machinery.

[11] Chenglin Li, Di Niu, Bei Jiang, Xiao Zuo, and Jianming Yang, "Meta-har: Federated representation learning for human activity recognition," in *Proceedings of the Web Conference 2021*, New York, NY, USA, 2021, WWW '21, p. 912–922, Association for Computing Machinery.

[12] Xiaomin Ouyang, Zhiyuan Xie, Jiayu Zhou, Jianwei Huang, and Guoliang Xing, "Clusterfl: A similarity-aware federated learning system for human activity recognition," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, New York, NY, USA, 2021, MobiSys '21, p. 54–66, Association for Computing Machinery.

[13] Alex Iacob, Pedro P. B. Gusmão, Nicholas D. Lane, Armand K. Koupai, Mohammud J. Bocus, Raúl Santos-Rodríguez, Robert J. Piechocki, and Ryan McConville, "Privacy in multimodal federated human activity recognition," 2023.

[14] Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu, "Cross-modal federated human activity recognition via modality-agnostic and modality-specific representation learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 3063–3071, Jun. 2022.

[15] Konstantin Sozinov, Vladimir Vlassov, and Sarunas Girdzijauskas, "Human activity recognition using federated learning," in *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications*, 2018.

[16] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni, "Federated learning with matched averaging," in *International Conference on Learning Representations*, 2020.

[17] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary, "Federated learning with personalization layers," *CoRR*, vol. abs/1912.00818, 2019.

[18] Pranjali Jain, Shreyas Goenka, Saurabh Bagchi, Biplab Banerjee, and Somali Chaterji, "Federated action recognition on heterogeneous embedded devices," *ArXiv*, vol. abs/2107.12147, 2021.