WiFi and Vision Multimodal Learning for Accurate and Robust Device-Free Human Activity Recognition

Han Zou[†], Jianfei Yang[‡], Hari Prasanna Das[†], Huihan Liu[†], Yuxun Zhou[†], Costas J. Spanos[†]
[†]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA
[‡] School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore
{hanzou, hpdas, liuhh, yxzhou, spanos}@berkeley.edu, yang0478@ntu.edu.sg

Abstract

Human activity recognition plays an indispensable role in a myriad of emerging applications in context-aware services. Accurate activity recognition systems usually require the user to carry mobile or wearable devices, which is inconvenient for long term usage. In this paper, we design WiVi, a novel human activity recognition scheme that is able to identify common human activities in an accurate and device-free manner via multimodal machine learning using only commercial WiFi-enabled IoT devices and camera. For sensing using WiFi, a new platform is developed to extract fine-grained WiFi channel information and transform them into WiFi frames. A tailored convolutional neural network model is designed to extract high-level representative features among the WiFi frames in order to provide human activity estimation. We utilized a variant of C3D model for activity sensing using vision. Following this, WiVi performs multimodal fusion at the decision level to combine the strength of WiFi and vision by constructing an ensembled DNN model. Extensive experiments are conducted in an indoor environment, demonstrating that WiVi achieves 97.5% activity recognition accuracy and is robust under unfavorable situations, as each modality provides the complementary sensing when the other faces its limiting conditions.

1. Introduction

Nowadays, various emerging applications in the field of human-computer interaction, home automation, gaming, and smart healthcare usually require the knowledge of human activity for human-centric design [2]. For instance, adaptive air conditioning based on human activity could not only improve the user's thermal comfort but also reduce the building energy consumption. Continuous human activity sensing could also enable personalized context-aware services and improve individual well-being in terms of comfort and health via activity analysis.

The sensing modalities for human activity recognition can be classified into two categories: device-based and device-free, from the user perspective. Device-based systems need cooperation from the user to perform inference. In particular, the embedded sensors on mobile and wearable devices (e.g. accelerometers, gyroscopes, and proximity sensors) are leveraged to identify activities [5]. Requiring user to carry or wear a particular set of devices is inconvenient for long-term usage even though they usually achieve high recognition accuracy. Device-free systems infer the activities in a non-intrusive manner. Vision constitutes the most popular sensing modality in the device-free category. Advanced convolutional neural networks (CNNs) performing image feature extraction have been proposed in recent years and have enhanced the accuracy of human activity identification significantly. But it also faces certain limitations. Cameras may not capture valid appearance information under poor illumination conditions, as well as with occlusion [1]. Moreover, the RGB camera may also expose users privacy. To address the privacy concern, researchers proposed to utilize depth [6] or thermal infrared camera [8]. Radio frequency (RF) wireless signals from specialized hardware platforms have been exploited for device-free activity recognition since human body movements alter the propagation of the signals [17]. However, the high cost of the dedicated infrastructure involved with these modalities and the intensive labor needed for their installations hinder them from large-scale application.

In this paper, we propose **WiVi**, a novel device-free human activity recognition scheme that is able to identify common human activities via multimodal machine learning with commercial off-the-shelf (COTS) **WiFi**-enabled Internet of Things (IoT) devices and COTS RGB **Vision** camera. The reason for selection of WiFi as a sensing modality for our scheme is that WiFi infrastructure is already widely available in indoor environments, and can be opportunistically utilized. We exploit a fine-grained channel measurement from WiFi physical layer, namely Channel State Information (CSI), which describes the detailed propaga-

978-1-7281-2506-0/19/$31.00 ©2019 IEEE
DOI 10.1109/CVPRW.2019.00056

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on September 05,2021 at 07:41:01 UTC from IEEE Xplore. Restrictions apply.

tion of WiFi signals from the transmitter (TX) to receiver (RX) through multiple paths at the granularity of Orthogonal Frequency Division Multiplexing (OFDM) subcarriers [20]. CSI is able to reveal human activity in a non-intrusive manner because the body movements during different activities would interfere with the signal propagation paths and give rise to distinct variations of CSI. Thus, it's the ideal modality to provide complementary information to vision sensing especially in poor lighting conditions. We first develop a WiFi CSI sensing platform to obtain the CSI measurements from COTS IoT devices directly, and then transform them into WiFi CSI frames for multimodal fusion.

WiVi consists of a WiFi sensing module and a vision sensing module that process WiFi frames and visual frames for unimodal inference, followed by a multimodal fusion module. A dedicated CNN architecture is designed as the feature extractor and classifier of the WiFi sensing module. We leverage a pre-trained C3D model on the Sports1M Dataset [11] as the architecture for the vision sensing module and fine-tune it with our vision dataset. WiVi conducts multimodal fusion at the decision phase (after both WiFi and vision has made a classification) because the mechanism is more flexible and robust to unimodal failure compared to feature level fusion. We concatenate the outputs of the SoftMax layer from the two modules and feed them into a four-layer deep neural networks (DNN) model to exploit the correlations and interactions between the modality estimations. The output of the DNN model is considered as the final activity estimation of WiVi. Real-world experiments are conducted in an indoor environment, demonstrating that WiVi recognizes common human activities with an accuracy of 97.5% by leveraging only two commercial WiFi routers and one camera. It also provides consistent activity inference under unfavorable lighting conditions. In general, WiVi makes substantial steps towards device-free human activity recognition using existing and pervasive vision and WiFi infrastructure for context-aware service applications.

2. Related Work

2.1. Sensing Modalities for Activity Recognition

The device-based human activity recognition systems require active cooperation from the user. For instance, the built-in inertial sensors on mobile devices, including accelerometers, gyroscopes and proximity sensors, are leveraged to identify various activities [5]. Smart watches and wristbands have also been exploited for activity recognition [7]. Although fine-grained activity recognition can be obtained from these sensors, the users have to carry these devices all the time, which is inconvenient and invasive for long-term usage.

The most popular sensing modality for device-free systems is visual frames captured by cameras. With the recent development of deep CNN and the availability of large labeled visual dataset, the accuracy of vision-based activity recognition has improved tremendously [2]. The fundamental challenges of vision-based approaches are poor illumination conditions and occlusion, since the camera requires line-of-sight for sensing, as well as appropriate brightness level. More importantly, using cameras for large-scale activity recognition raises severe privacy concerns. Although researchers also proposed to use depth camera [6] and thermal infrared camera [8] to overcome the privacy issue, the high cost of these cameras hinder them for ubiquitous implementation. In addition to visual frames, wireless signals have been exploited for device-free human activity recognition. RF signals from specialized hardware platform (e.g. USRP and FMCV) are utilized to identify human activities and poses [17]. Its limitations include requirement of dedicated wireless transmitters and receivers.

Meanwhile, WiFi has been acknowledged as the most pervasive wireless signal in indoors. With the booming development of IoT, billions of WiFi enabled IoT devices, e.g. thermostats, smart switch, sound bar, and smart TV are en route to being ubiquitous in buildings. Due to the low cost and privacy preserving properties, WiFi has been recognized as the primary sensing modality for occupancy sensing in indoors [15]. Numbers of occupancy sensing applications, e.g. occupancy detection [21], crowd counting [20], location estimation [18, 9], human identification [22], and activity recognition [14], have been realized. The basic rationale behind this is, when we perform different activities, the movement of our human body alters WiFi signal propagation paths between the TX and RX. Thus, human activities can be inferred by analyzing these changes and variations at the RX without user instrumentation or extra infrastructure. Furthermore, CSI, a fine-grained reading in WiFi PHY layer became accessible recently, which can capture the subtle variations of WiFi signals caused by human activities and can be analyzed at the RX. In this work, we aim to explore the potentials of using WiFi CSI readings from COTS WiFi-enabled IoT devices for device-free human activity recognition.

2.2. Multimodal Machine Learning

Multimodal machine learning is a modelling approach which aims to extract the novelty of multiple sensing modalities by processing and relating information from them via fusion, co-learning and other methodologies. Multimodal machine learning derives its motivation from the fact that unimodal models have their own shortcomings making them perform sub-optimally, e.g. vision-based models do not work well when the images have issues with respect to illumination, camera angle or background clutter. In such cases, an ensemble of data from different modalities prove beneficial. Additionally, human experience of

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on September 05,2021 at 07:41:01 UTC from IEEE Xplore. Restrictions apply.

the world is multimodal, with vision, audio, language and olfactory receptors among others, which encourages introduction of multimodal learning.

Multimodal machine learning involves two major techniques, *Fusion* and *Co-learning*, which essentially contribute towards the novelty that it offers. *Fusion* involves combining information from multiple sensing modalities, either at the feature or decision level, to perform a prediction. Feature based fusion integrates the features from different modalities immediately after they are extracted, whereas decision-based fusion performs integration after each of the modalities has made a decision (e.g., classification or regression). The feature-based fusion learns to exploit the correlations between low-level features of different modalities and is simple to train as it involves training a single model. On the other hand, decision-based fusion fuses the unimodal decisions using some mechanism such as averaging, voting schemes or a learned model. In retrospect, decision-based models have several important advantages. First, they allow more flexibility by allowing different models for different modalities. Second, they are more robust to loss of data from single or multiple modalities. We adopt the decision-based fusion for the current research. *Co-learning* enables modelling of a resource poor (lack of annotated data, noisy input) modality by exploiting knowledge from other resource rich modalities. It achieves the capability by utilizing state of the art transfer learning and domain adaptation methods [12, 23]. Current day machine learning algorithms rely heavily on annotated data for their training, whose availability needs significant human effort. So, there has been a lot of discussion on improvising unsupervised learning methods. In such scenario, co-learning proves worthy of its existence. We introduce newer avenues where co-learning finds important application combined with our proposed sensor fusion method.

3. Method

In this section, we first elaborate the sensing methodology and designed models for WiFi and vision sensing modules. Then, we introduce the proposed ensemble learning method for WiFi and vision multimodal fusion.

3.1. WiFi Sensing Module

3.1.1 WiFi CSI Sensing Platform

Due to the complexity of the indoor environment, WiFi signals usually travel through more than one path between TX and RX. The signals get scattered and reflected by furniture and human movements [16]. Since multiple antennas are commonly equipped with commercial WiFi IoT devices, signals obtained from them provide significant information for data analytics. CSI provides sophisticated information (e.g. amplitude attenuation, phase shift and time delay) about how signals propagate between TX and RX at each OFDM subcarrier through multiple paths [19]. We model the WiFi signal as a channel impulse response $h(\tau)$. In the frequency domain, since a sampled version of the signal spectrum on each subcarrier can be obtained from the RX, the CSI measurements can be summarized as complex numbers: $H_i = \|H_i\|e^{j\angle H_i}$, where $\|H_i\|$ denotes the amplitude attenuation and $\angle H_i$ represents the phase shift at the i^{th} subcarrier.

Most of the conventional CSI-based sensing systems leverage the Intel 5300 NIC tool [4] to collect CSI measurements. A laptop and an external dedicated WiFi adapter are required to construct the receiver for CSI data acquisition [13], which is impractical for pervasive implementation. To relieve the requirement of introducing extra hardware, we develop an OpenWrt firmware that can run on commercial WiFi routers for CSI data acquisition. With our firmware, the routers can analyze the data packets transmitted in the WiFi traffic and extract the CSI measurements from those frames. By upgrading the Atheros CSI Tool [13], our platform can also provide CSI readings from all 114 subcarriers for 40MHz channel when the routers are operating at 5 GHz. Thus, we obtain $N_{TX} \times N_{RX} \times 114$ CSI streams at each time for analyzing how human activities interfere with WiFi signals, where N_{TX} and N_{RX} represent the number of antennas on TX and RX, respectively.

3.1.2 WiFi Frames of Human Activities

To validate whether human activity information can be explored from WiFi CSI data, an experiment was conducted in a $(6.1m \times 4.4m)$ conference room. We upgraded the firmware of 2 TPLINK N750 WiFi routers to our WiFi CSI sensing platform (served as TX and RX, put 3 meters apart). A volunteer performed a series of common human activities, including sitting, standing, and walking near the TX-RX pair. The CSI phase difference readings across 2 RX antennas during the experiment are depicted in the second row of Fig. 1. It can be seen from Fig. 1 that distinct perturbations on CSI measurements are generated by different human activities. The readings were relatively smooth under sitting and standing than walking. Based on these observations, we can conclude that unique information associated with each activity can be revealed from CSI measurements. Since the platform does not require user to carry or wear any device, it is another ideal information source for device-free human activity recognition. Furthermore, as depicted in Fig. 1, the 2D WiFi CSI time-series data encapsulate the same human activities as video, but from a different perspective.

Thus, we divide the CSI streams into small chunks with a sliding window Δt and transform them into WiFi frames. Each frame includes $n \times m$ WiFi CSI pixels, which n is

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on September 05,2021 at 07:41:01 UTC from IEEE Xplore. Restrictions apply.

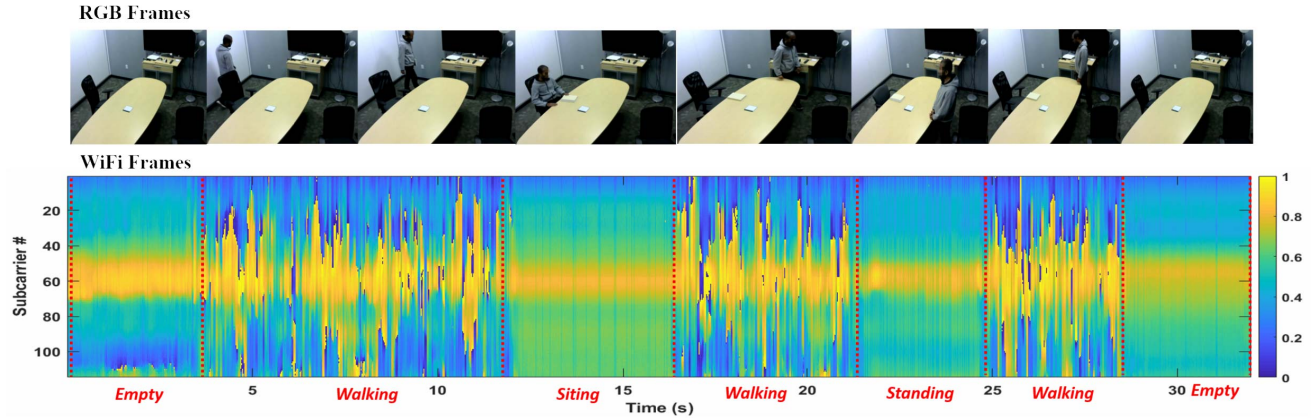

Figure 1. The synchronized RGB frames and WiFi CSI phase difference frames of a series of human activities.

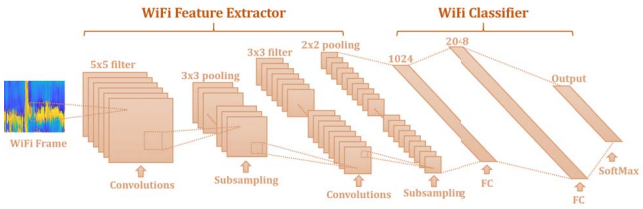

Figure 2. Designed CNN architecture of WiFi sensing module.

the number of consecutive samples and m represents the number of subcarriers. These WiFi frames are leveraged as the input dataset for constructing the WiFi sensing module via deep learning.

3.1.3 Deep Learning Model for WiFi Sensing

The WiFi sensing module consists of the WiFi feature extractor and the WiFi classifier as demonstrated in Fig. 2. The sampling rate at RX is 100 packets/s. By measuring the CSI phase difference data from 114 subcarriers, we constructed a WiFi frame with a size (100×114) every time window Δt. In order to extract the most discriminative local features from the raw WiFi frames. WiFi feature extractor constitutes two pairs of convolutional layer and subsampling layer in a cascaded manner. The convolutional layer extracts local features by using multiple filters slided over the input followed by nonlinear activation functions. The ith feature map $\theta_i^{l_c}$ in layer l_c is given by $\theta_i^{l_c} = \sigma\left(\sum_{m \in S_{l-1}} w_{i,m}^{l_c} * \theta_m^{l_c-1} + b_i^{l_c}\right)$, where σ denotes the activation function. We use rectified linear unit (ReLU) [10] $f(z) = max(0, z)$ because it has better gradient propagation than sigmoid activation function and at the same time is also scale-invariant. S_{l-1} represents a set of the feature maps in the previous layer connected to the current feature map, $w_{i,m}^{l_c}$ is the convolutional kernel for feature map generation and $b_i^{l_c}$ is the bias of the ith feature map in the current layer l_c. In order to reduce the number of features and the computational complexity of the network, we connect a subsampling layer after the convolutional layer to progressively reduce the spatial size of the feature maps by downsampling. The features are split into several partitions. Then, in each partition, we use max operator to generate the output. The activation function of above max-pooling operation is: $\theta_i^{l_c} = max_{k=1}^{r}(\theta_k^{l_c-1})$. It preserves the scale-invariant features from the previous convolutional layer. It makes the detection of features invariant to changes in scale [3].

Fig. 2 demonstrates the proposed CNN architecture for the WiFi sensing module. In the first convolutional layer, we leverage 32 filters with kernel size 5×5 to generate 32 feature maps with size 96×110. Following this, we reduce the dimensionality of the data while guaranteeing the invariance of feature maps by max pooling with size 3×3 in the subsampling layer. We further process the data with one more pair of convolutional and subsampling layers, and obtain 64 feature maps with size 15×16 as shown in Fig. 2. The WiFi feature extractor is concatenated with a WiFi classifier, which consists of two fully connected (FC) layers and a SoftMax transformation layer. The output of the SoftMax layer (C_w) are used (WiFi sensing estimation) for the multimodal learning module of WiVi.

As an end-to-end network architecture, WiFi feature extractor and the WiFi classifier are trained jointly via backpropagation [3] to minimize the cross-entropy loss via Adam optimizer.

3.2. Vision Sensing Module

Vision based human activity recognition has been extensively studied in recent years and has achieved encouraging accuracy due to the blooming development of CNNs. Since human activity usually consists of a sequence of body movements, a video that includes a sequence of RGB frames is more informative than individual frame to capture the temporal dependencies. Therefore, we utilize the Convolutional 3D (C3D) model [11] as the architecture for the

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on September 05,2021 at 07:41:01 UTC from IEEE Xplore. Restrictions apply.

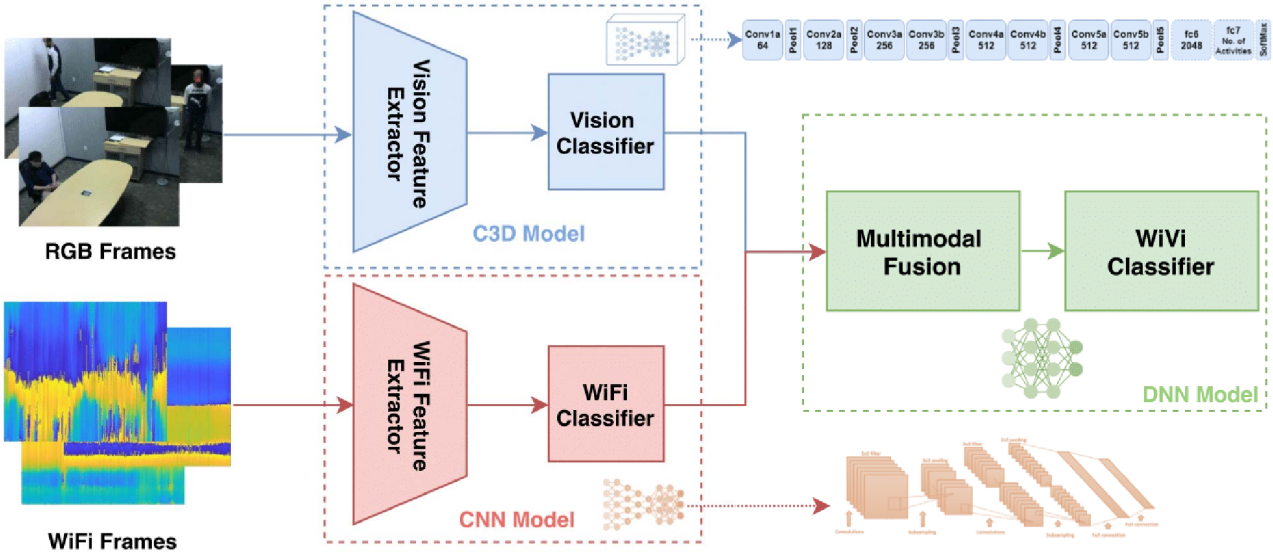

Figure 3. Architecture of WiVi.

Figure 4. Modified C3D model for vision sensing module. The network parameters in solid line boxes are fixed and those in dashed line boxes are trained.

vision sensing module (vision feature extractor and vision classifier). 3D CNNs are more suitable for spatiotemporal feature learning in videos compared to conventional 2D CNNs. C3D is a 15-layer 3D CNN architecture dedicated for action recognition.

In order to train our vision model in a more efficient manner, the parameters are initialized with a pre-trained C3D model on the Sports1M Dataset. We resize our RGB frames to 128×171 and cluster them into sequences of 16 frames, the default input size for the pre-trained model. All the parameters in the C3D model remain fixed except the last two FC layers. As presented in Fig. 4, these layers are replaced with a FC layer (No. of neurons: 2048) and a FC layer (No. of neurons: No. of activity categories). The parameters in these layers are tuned via backpropagation with SGD as the optimizer to minimize the loss function. The output of the SoftMax layer (C_v) is used as the input (vision sensing estimation) for the multimodal learning module of WiVi.

3.3. Multimodal Learning Framework

With the activity estimation from the WiFi sensing module (C_w), as well as the inference from the vision sensing module (C_v), WiVi conducts multimodal fusion at the decision phase via ensemble learning. Suppose there are k synchronized WiFi and vision data samples with c activity categories. Then, the output of the SoftMax layer of each sensing module is a $k \times c$ matrix, which represents the categorical distribution over class labels. As depicted in

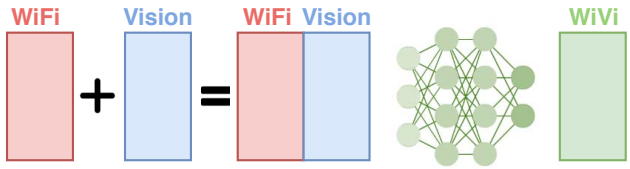

Figure 5. Multimodal fusion with ensemble learning at decision phase.

Fig. 5, we concatenate the outputs of the two modules as a new input for the ensemble classifier with a size of $k \times 2c$. After comprehensive evaluations of popular ensemble classifiers, including majority voting, weighted average, logistic regression and DNN, we found that the performance of DNN achieves the highest classification accuracy. Thus, a four-layer DNN model (128-256-128-No. of activity categories with ReLu as activation function) is leveraged as the ensemble classifier for WiVi. During the training phase, the DNN model is trained by leveraging the concatenated estimations from WiFi and vision ($k \times 2c$) with ground truth being the categorical labels of the training dataset ($k \times c$). During the testing phase, the synchronized RGB frames and WiFi frames are fed into the proposed C3D model and CNN model to obtain unimodal estimations. Then, their estimations are fed into the ensemble DNN model, and the output of the DNN model is the final activity inference provided by WiVi, C_{WiVi}.

4. Experiments

4.1. Experimental Setup

To evaluate the human activity recognition performance of WiVi, we deployed two TPLINK N750 WiFi routers for WiFi data acquisition and one Logitech C270 webcam to

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on September 05,2021 at 07:41:01 UTC from IEEE Xplore. Restrictions apply.

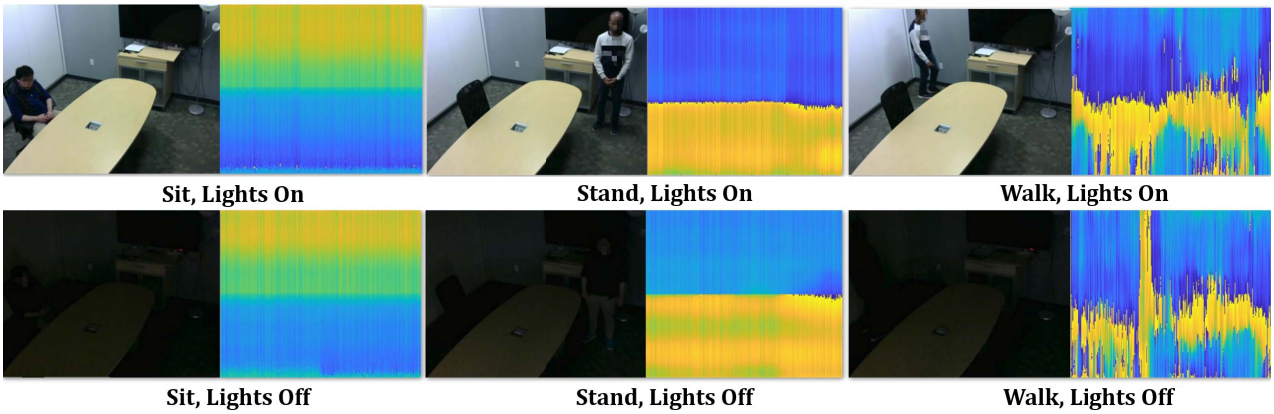

Figure 6. RGB frames (*left*) and WiFi frames (*right*) of various human activities under different illumination conditions.

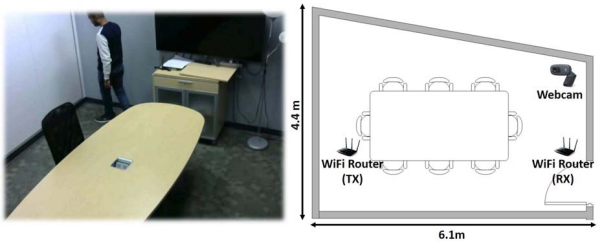

Figure 7. A sample RGB frame and the floorplan of the testbed, the locations of the camera and 2 WiFi router (one serves as transmitter (RX) and one servers as receiver (RX).

obtain the vision data in a $(6.1m \times 4.4m)$ conference room. The floorplan of the conference room, as well as the locations of the routers and cameras, are depicted in Fig. 7. The firmware of the routers was upgraded to our CSI IoT platform and they formed a WiFi TX-RX pair (3 meters apart) to sense the nearby human activities. Existing WiFi networks were operated as usual during the entire experiments. The webcam was installed below the ceiling of the room. The visual frame size is (640×480). As shown in Fig. 7, it captured the majority of human activities but with partial occlusions, where parts of the human body, such as legs and arms, were hidden by furniture during some phase of the experiment.

Two volunteers participated in the experiments and performed three common activities, sitting, standing and walking, at arbitrary locations in the conference room. one volunteer conducted one of the activities for 5 - 6 seconds each trace and we collected 50 traces per activity per volunteer. During the experiments, both WiFi and video data were collected simultaneously with only 4 milliseconds average synchronization error. In addition to the common light-on scenario, we also conducted all the experiments under the light-off mode (to mimic poor lighting conditions) as demonstrated in Fig. 6. Thus, there were 600 pairs of synchronized WiFi and vision samples in total to fully evaluate the performance of WiVi. The WiFi sensing module and vision sensing module were trained with the proposed CNN model and C3D model respectively as introduced in Section 3.1.3 and Section 3.2. The multimodal fusion DNN model is tuned afterwards. 80% of the dataset are used for training and remaining 20% are leveraged for evaluation.

4.2. Accuracy Analysis

We first validate WiVi's activity recognition performance in terms of True Positive Rate (TPR), which is the ratio of the number of times for identifying an activity correctly to the total number of activities performed, when both WiFi and vision modules are operating in normal condition (light-on). According to our experimental results, WiVi achieves 97.5% cross-validation accuracy on average for 3 different activities with 2 volunteers. By inheriting the strengths of both WiFi and vision sensing modalities, it outperforms each individual modality (WiFi: 95.83%; vision: 95%). Fig. 8 demonstrates the confusion matrices for human activity recognition using (a) WiFi, (b) Vision and (c) WiVi, respectively. It can be observed from Fig. 8 (a) that the WiFi sensing module can distinguish walking from other activities easily since the variance of WiFi CSI introduced by walking is much higher than sitting and standing (as depicted in Fig. 6). On the other hand, since both sitting and standing are relative stationary activities, there are some misclassifications in each category. With the appropriate lighting condition, vision sensing module achieves high recognition accuracy, especially for sitting and standing as depicted in Fig. 8 (b). But it misclassified walking to standing by 15% due to the frame similarities between standing and certain moments during walking. As shown in Fig. 8 (c), WiVi integrates the strength of WiFi module (strong identification capability for walking), and the advantage of vision module (strong classification capability among sitting and standing) and achieve outstanding human activity recognition performance. Moreover, the results presented in the bar chart in Fig. 9 also validates that its performance is consistent across different users.

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on September 05,2021 at 07:41:01 UTC from IEEE Xplore. Restrictions apply.

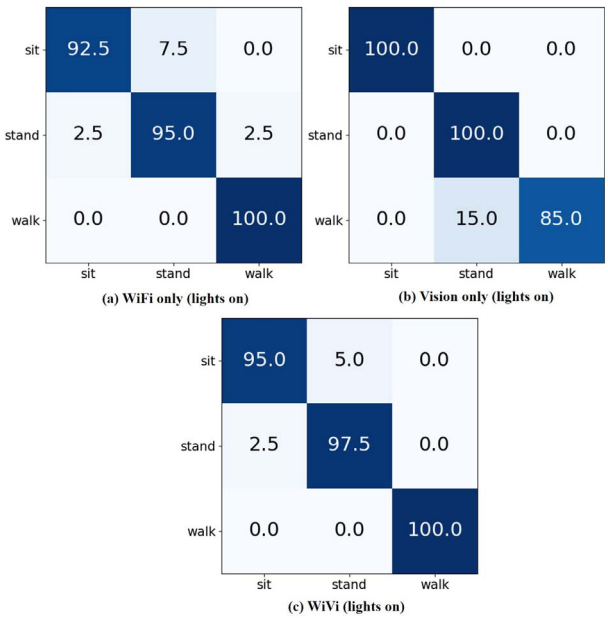

Figure 8. The confusion matrices for human activity recognition using (a) WiFi, (b) Vision and (c) WiVi under *lights on* scenario.

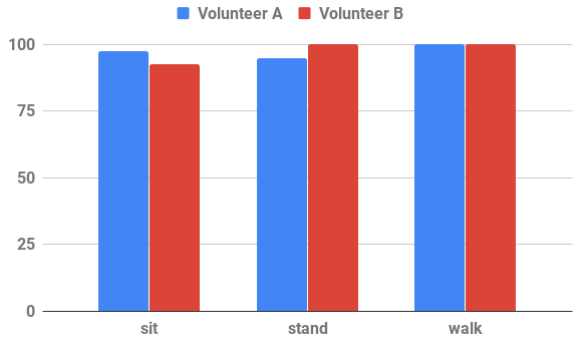

Figure 9. WiVi's human activity recognition accuracy (%) of two volunteers under *lights on* scenario.

4.3. Robustness Analysis

To validate the robustness of WiVi under unfavorable sensing circumstance, we turned off all the lamps in the conference room to create a poor lighting condition. The second row of Fig. 6 demonstrates the sample RGB frames and WiFi frames captured for different activities under the light off scenario. As it can be seen from these figures, although the table is still partially looming due to the reflection of the light from outside the conference room, other places are comparably dark. This directly led to performance degradation of vision sensing module as presented in Fig. 10 (a). Its recognition accuracy is only 46.67%, which cannot meet the basic requirement of any context-aware services.

On the other hand, by comparing the WiFi images under lights on and lights off conditions in Fig. 6, it can be observed that the degree of perturbations on WiFi readings

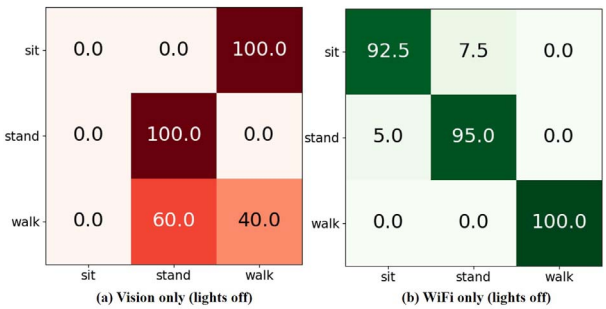

Figure 10. The confusion matrices for human activity recognition using (a) Vision and (b) WiFi under *lights off* scenario.

of same human activities are similar with each other and distinct from other activities, regardless the variations of brightness level. By leveraging our designed inference network, the WiFi sensing module provides 95.83% activity recognition accuracy as shown in Fig. 10 (b). Thus, WiVi contains a brightness detection module, that ceases the estimation process of the vision sensing module when poor illumination is detected and makes WiVi rely on the estimates from the WiFi sensing module to ensure the high recognition accuracy.

Similarly, the performance of WiFi sensing module may also degrade under certain unfavorable conditions, e.g. when a huge metal object blocks the antennas of TX or RX, or the RX IoT device experiences severe wireless interferences or network congestions that cannot parse the data normally. So, we use the averaged signal-to-noise ratio (SNR) at RX as an indicator to determine whether the WiFi sensing module should be operated or not. If the SNR reading is less than a predefined empirical threshold, the WiFi module is postponed and the estimation of the vision module is the output of WiVi. In this manner, WiVi guarantees the activity recognition performance when one of the modalities is malfunctioned under adversarial circumstances.

5. Conclusion

In this paper, we proposed WiVi, an innovative human activity recognition scheme that is able to identify common human activities in an accurate and robust manner by conducting multimodal fusion of the pervasive WiFi signals from COTS IoT devices and RGB frames from cameras via deep learning and ensemble learning. A WiFi CSI sensing platform is developed that enables CSI data acquisition from COTS IoT devices and also transforms them into WiFi frames. A tailored CNN model is designed to extract representative features among the WiFi frames and provide estimations from the WiFi sensing module. A modified C3D model is fine-tuned as the vision sensing module to fit our activity dataset. Following that, we designed a DNN model as the ensemble classifier to reveal the correlations and interactions between the inferences provided by the two

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on September 05,2021 at 07:41:01 UTC from IEEE Xplore. Restrictions apply.

modalities. The output of DNN is the final activity estimation of WiVi. We implemented WiVi using only two commercial routers and one camera and connected experiments in real-world indoor space. According to the experimental results, WiVi achieves 97.5% human activity recognition accuracy which inherits the strengths of both WiFi and vision, and its performance is consistent even under harsh or unfavorable conditions. Future works include multimodal fusion with other modalities (e.g. acoustic signals), cross-modal supervision and transfer learning across modalities.

Acknowledgments

This work is supported by a 2018 Seed Fund Award from CITRIS and the Banatao Institute at the University of California.

References

[1] N. Carissimi, P. Rota, C. Beyan, and V. Murino. Filling the gaps: Predicting missing joints of human poses using denoising autoencoders. In *European Conference on Computer Vision*, pages 364–379. Springer, 2018. 1

[2] N. C. Garcia, P. Morerio, and V. Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. 1, 2

[3] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. 4

[4] D. Halperin, W. Hu, A. Sheth, and D. Wetherall. Tool release: gathering 802.11 n traces with channel state information. *ACM SIGCOMM Computer Communication Review*, 41(1):53–53, 2011. 3

[5] B. Huang, G. Qi, X. Yang, L. Zhao, and H. Zou. Exploiting cyclic features of walking for pedestrian dead reckoning with unconstrained smartphones. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 374–385. ACM, 2016. 1, 2

[6] E. P. Ijjina and K. M. Chalavadi. Human action recognition in rgb-d videos using motion sequence information and deep learning. *Pattern Recognition*, 72:504–516, 2017. 1, 2

[7] M. Jin, H. Zou, K. Weekly, R. Jia, A. M. Bayen, and C. J. Spanos. Environmental sensing by wearable device for indoor activity and location estimation. In *Industrial Electronics Society, IECON 2014-40th Annual Conference of the IEEE*, pages 5369–5375. IEEE, 2014. 2

[8] V. V. Kniaz, V. A. Knyaz, W. G. Kropatsch, and V. Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *European Conference on Computer Vision*, pages 606–624. Springer, 2018. 1, 2

[9] X. Lu, H. Wen, H. Zou, H. Jiang, L. Xie, and N. Trigoni. Robust occupancy inference with commodity wifi. In *2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 1–8. IEEE, 2016. 2

[10] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 4

[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2, 4

[12] R. Volpi, P. Morerio, S. Savarese, and V. Murino. Adversarial feature augmentation for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5495–5504, 2018. 3

[13] Y. Xie, Z. Li, and M. Li. Precise power delay profiling with commodity wifi. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 53–64. ACM, 2015. 3

[14] J. Yang, H. Zou, H. Jiang, and L. Xie. Carefi: Sedentary behavior monitoring system via commodity wifi infrastructures. *IEEE Transactions on Vehicular Technology*, 2018. 2

[15] J. Yang, H. Zou, H. Jiang, and L. Xie. Device-free occupant activity sensing using wifi-enabled iot devices for smart homes. *IEEE Internet of Things Journal*, 5(5):3991–4002, 2018. 2

[16] J. Yang, H. Zou, H. Jiang, and L. Xie. Fine-grained adaptive location-independent activity recognition using commodity wifi. In *Wireless Communications and Networking Conference (WCNC), 2018 IEEE*, pages 1–6. IEEE, 2018. 3

[17] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, 2018. 1, 2

[18] H. Zou, M. Jin, H. Jiang, L. Xie, and C. Spanos. Winips: Wifi-based non-intrusive ips for online radio map construction. In *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1081–1082. IEEE, 2016. 2

[19] H. Zou, Y. Zhou, R. Arghandeh, and C. J. Spanos. Multiple kernel semi-representation learning with its application to device-free human activity recognition. *IEEE Internet of Things Journal*, 2019. 3

[20] H. Zou, Y. Zhou, J. Yang, W. Gu, L. Xie, and C. Spanos. Freecount: Device-free crowd counting with commodity wifi. In *Global Communications Conference (GLOBECOM), 2017 IEEE*. IEEE, 2017. 2

[21] H. Zou, Y. Zhou, J. Yang, W. Gu, L. Xie, and C. Spanos. Freedetector: Device-free occupancy detection with commodity wifi. In *Sensing, Communication and Networking (SECON Workshops), 2017 IEEE International Conference on*, pages 1–5. IEEE, 2017. 2

[22] H. Zou, Y. Zhou, J. Yang, W. Gu, L. Xie, and C. Spanos. Wifi-based human identification via convex tensor shapelet learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1711–1718, 2018. 2

[23] H. Zou, Y. Zhou, J. Yang, H. Liu, H. Das, and C. Spanos. Consensus adversarial domain adaptation. In *AAAI Conference on Artificial Intelligence*, 2019. 3

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on September 05,2021 at 07:41:01 UTC from IEEE Xplore. Restrictions apply.