MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Wi-Fi based Indoor Monitoring Enhanced by Multimodal Fusion

Hori, Chiori; Wang, Pu; Rahman, Mahbub; Vaca-Rubio, Cristian; Khurana, Sameer; Cherian, Anoop; Le Roux, Jonathan

TR2024-012 March 07, 2024

Abstract

Indoor monitoring systems are in high demand to protect vulnerable people, especially when they are alone at home, in nursing homes, hospitals, etc. Although surveillance systems in public spaces use cameras and microphones to find incidents, indoor monitoring in personal spaces needs to protect privacy. Such systems thus need to understand scenes without relying on direct sensing information, e.g., from audio-visual sensors, instead using indirect sensing information that is difficult to interpret by humans and may be insufficient to understand ongoing events precisely. To mitigate this draw- back, this paper proposes a new indoor monitoring approach that at- tempts to realize scene understanding using only indirect sensors by transferring the learned inductive bias of a multimodal fusion model trained using direct and indirect sensing information to a model that uses only indirect information during inference. We collected direct (audio-visual) and indirect (infrared and Wi-Fi) sensing information of indoor human actions in daily life and manually annotated event captions. We build models that can generate event captions from various combinations of indirect and direct sensor data, and show that our transfer learning approach leads to significant improvements in caption quality when only indirect information is used at inference time.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

WI-FI BASED INDOOR MONITORING ENHANCED BY MULTIMODAL FUSION

Chiori Hori, *Pu Wang*, *Mahbub Rahman*[*], *Cristian Vaca-Rubio*[*],
Sameer Khurana, *Anoop Cherian*, *Jonathan Le Roux*

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

ABSTRACT

Indoor monitoring systems are in high demand to protect vulnerable people, especially when they are alone at home, in nursing homes, hospitals, etc. Although surveillance systems in public spaces use cameras and microphones to find incidents, indoor monitoring in personal spaces needs to protect privacy. Such systems thus need to understand scenes without relying on direct sensing information, e.g., from audio-visual sensors, instead using indirect sensing information that is difficult to interpret by humans and may be insufficient to understand ongoing events precisely. To mitigate this drawback, this paper proposes a new indoor monitoring approach that attempts to realize scene understanding using only indirect sensors by transferring the learned inductive bias of a multimodal fusion model trained using direct and indirect sensing information to a model that uses only indirect information during inference. We collected direct (audio-visual) and indirect (infrared and Wi-Fi) sensing information of indoor human actions in daily life and manually annotated event captions. We build models that can generate event captions from various combinations of indirect and direct sensor data, and show that our transfer learning approach leads to significant improvements in caption quality when only indirect information is used at inference time.

***Index Terms*—** indoor monitoring, multimodal scene understanding, audio-visual, Wi-Fi, infrared, student-teacher learning

1. INTRODUCTION

Surveillance systems using cameras and microphones are very common in public spaces for security purposes, and event recognition is intensively investigated to find incidents promptly. To advance scene understanding through object and event recognition, in [1] we introduced multimodal fusion approaches to event captioning and scene-aware interaction using natural language by combining various kinds of sensing information. Such a capability rapidly raises demands for indoor monitoring at home, in hospitals, and in elderly-care centers to monitor vulnerable people and protect them from incidents. To protect privacy, indoor monitoring systems need to understand events without relying on direct sensing information, such as that from audio-visual sensors, using instead indirect sensing information.

The first attempt at event captioning using indirect sensing information was made using mmWave signals to caption in-home daily life [2]. The approach is based on a supervised captioning model trained from paired mmWave signals and text captions to generate event captions for human actions. Since the mmWave signal by itself is not sufficient to describe events in detail, the authors combined it with the floor map information by embedding a human body skeleton and an unpaired data alignment loss.

[*] Work performed as MERL interns

The ambient Wi-Fi signals can also be used for lower-level localization and device-free human sensing [3,4]. Earlier attempts use coarse-grained receiver signal strength indicator (RSSI) for device localization, while state-of-the-art pipelines leverage fine-grained channel state information (CSI). CSI measurements were used for human gait identification [5], person identification [6, 7], gesture recognition [8,9], activity recognition [10], human behavior prediction [11], emotion sensing [12, 13], face expression detection [14], and breathing rate monitoring during sleep [13]. More recently, Person-in-WiFi [15] used annotations from camera images to train fine-grained CSI measurements at 3×3 antennas at 100 fps for downstream tasks such as segmentation mask and skeleton estimation. [16] further extended the skeleton tracking from 2D to 3D. Extracted environment-independent Doppler profile [17] from the CSI phase may enhance the robustness. However, Wi-Fi-based performance is still limited by the temporal and spatial resolution and its generalization capability to new environments.

Indirect information may however be difficult to interpret by humans and insufficient to understand ongoing events precisely. To mitigate such a drawback, this paper proposes a new indoor monitoring approach that attempts to realize scene understanding using only indirect sensors by transferring the power of a multimodal fusion model trained from direct and indirect sensing information to a model uses only direct sensing information during inference. We previously introduced the contribution of audio features for video captioning and proposed a multimodal attention approach to fuse audio and visual features [18]. The multimodal event captioning framework deploying Audio Visual Transformer [19] can enhance the performance of audio-visual scene-aware dialog (AVSD) [20–23] and robot action sequence generation from instruction videos by combining instruction speech [24]. This paper extends the multimodal scene understanding based on event captioning to indoor monitoring tasks.

In this work, we describe our collection of direct (audio-visual) and indirect (Wi-Fi and Infrared) sensing information about indoor human actions in daily life and manually annotated event captions. Figures 2, 3, and 4 show RGB images and a corresponding Depth image, a spatio-temporal image captured by the Wi-Fi sensors, and a heat map image. We build multimodal fusion models that can generate event captions from various combinations of direct and indirect sensing information to show the power of multimodal fusion. Furthermore, we attempt to train the multimodal fusion model and a model trained from only the direct sensing information using Joint Student-Teacher Learning [25]. The student model can generate event captions at inference time using only the indirect sensing information leveraged by the teacher model trained from the multimodal fusion model. The experimental results show that our transfer learning approach leads to significant improvements in caption quality when only indirect information is used at inference time.

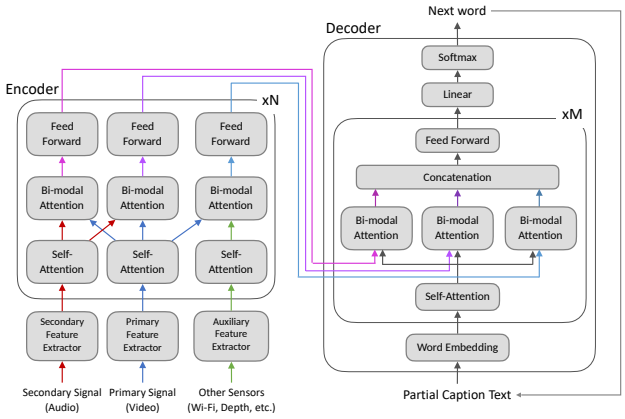


Fig. 1: Multimodal Transformer for indoor monitoring.

2. MULTIMODAL SCENE UNDERSTANDING

In this section, we introduce the task of multimodal scene understanding, in which a system takes a multimodal input consisting of RGB with depth, audio, infrared, and Wi-Fi signals, and needs to output event captions. The evaluation of event captioning follows standard segment-matching natural language evaluation metrics. In particular, we compare the prediction and ground truth using metrics such as BLEU, METEOR, and ROUGE. The model we use for this approach is based on the audio-visual Transformer model [19], which contains an audio-visual encoder, a caption decoder, and an event proposal generator. The audio-visual encoder has self-attention layers for each modality and cross-attention layers across modalities to better encode audio-visual features. The caption decoder is an auto-regressive Transformer decoder, which generates words by attending both to audio and visual encodings. In this work, we extend the audio-visual Transformer to a multimodal Transformer that accepts multiple inputs from more than two modalities. However, we skip the training of the proposal generator, i.e., we use ground-truth video segments in the experiments to focus on multimodal encoding and caption generation. The evaluation including the event proposal will be addressed in future work.

Figure 1 shows the multimodal Transformer for caption generation, where we assume three signals, primary, secondary, and auxiliary signals. The primary signal can be a video segment, the secondary signal can be an audio segment, and the auxiliary signal may be some additional information such as Wi-Fi signal. If the model accepts only the primary and secondary inputs, the architecture is the same as the audio-visual Transformer. To encode more sensing information, we can add multiple auxiliary encoders for the additional sensors. As shown in the figure, the primary and secondary encoders interact with each other through cross-attention layers, while the auxiliary encoders interact with only the primary encoder. This restriction has the advantage of keeping the cross-attention complexity linear with the number of additional modalities.

Let H_m^0 be the feature vector sequence extracted from the m-th input signal. The n-th encoder block computes hidden vector sequences as

$$\bar{H}_m^n = H_m^{n-1} + \mathrm{MHA}(H_m^{n-1}, H_m^{n-1}, H_m^{n-1}), \quad (1)$$

$$\tilde{H}_m^n = \bar{H}_m^n + \mathrm{MHA}(\bar{H}_m^n, \bar{H}_k^n, \bar{H}_k^n), \quad (2)$$

$$H_m^n = \tilde{H}_m^n + \mathrm{FFN}(\tilde{H}_m^n), \quad (3)$$

where MHA and FFN denote multi-head attention and feed-forward network, respectively. MHA takes three arguments, query, key, and value vector sequences [26]. Equation (2) represents cross-attention, with $k = 2$ if $m = 1$ (primary modality cross-attends to secondary modality), otherwise $k = 1$ (all other modalities cross-attend to the primary one). Layer normalization [27] is applied before each MHA and FFN layers, but omitted from the equations for simplicity.

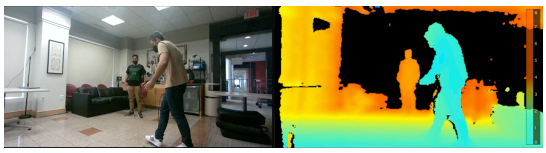


Fig. 2: RGB and corresponding depth captured by the Stereo Camera: Intel RealSense Depth Camera D455.

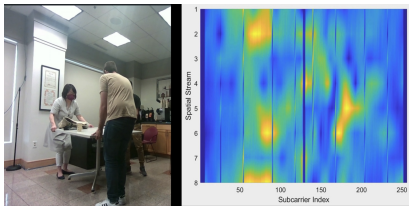


Fig. 3: RGB and corresponding spatio-temporal features captured by the Wi-Fi sensors: ASUS AC2900 WiFi Router (RT-AC86U).

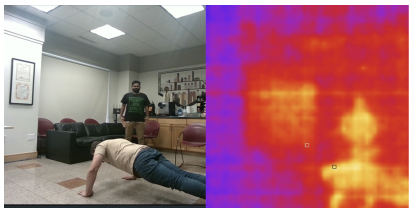


Fig. 4: RGB and corresponding heat map image captured by Infrared sensor: MedlDir MIR8060.

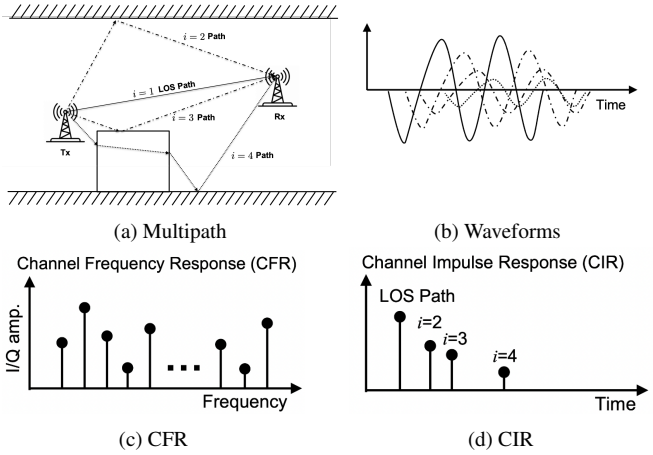
3. INTER-MODAL TRANSFER LEARNING

The goal of this step is to obtain a student model that does not make use of direct sensors, and is trained to mimic a teacher model that has already been trained using all available sensors. Accordingly, the student model can be used to generate captions without relying on direct sensors, while hopefully achieving similar performance to the teacher model. We refer to this technique as inter-modal transfer learning (IMTL).

The student network is trained to minimize the KL-divergence loss, which corresponds to the cross-entropy loss when using the output of the teacher network as a *soft* target. Reducing the KL divergence makes the output distribution of the student model closer to that of the teacher model. The KL-divergence loss is computed as

$$\mathcal{L}_{\mathrm{ST}}(X, Y) = -\sum_{i=1}^{|Y|} \sum_{y \in \mathcal{V}} P_T(y|\hat{y}_{1:i-1}, \hat{X}_{1:\hat{M}}) \log P_S(y|\hat{y}_{1:i-1}, X_{1:M}), \quad (4)$$

where $P_T(y|\hat{y}_{1:i-1}, \hat{X}_{1:\hat{M}})$ denotes the probability distribution for the ith word obtained by the teacher network given the preceding ground-truth word sequence $\hat{y}_{1:i-1}$ and the feature vector sequences $\hat{X}_{1:\hat{M}}$ from $\hat{M}$ sensors, and $P_S(y|\hat{y}_{1:i-1}, X_{1:M})$ is the posterior distribution obtained from the student network currently being trained, given $\hat{y}_{1:i-1}$ and the feature vector sequence $X_{1:M}$ corresponding to M indirect sensors (which typically will be a subset of the $\hat{M}$



(a) Multipath (b) Waveforms

(c) CFR (d) CIR

Fig. 5: Wi-Fi CSI channel measurements at 5-GHz frequency bands in 802.11ac.

sensors in $\hat{X}_{1:\hat{M}}$, although not necessarily).

4. WI-FI FEATURE EXTRACTION

4.1. Channel State Information (CSI)

In a typical indoor setting in Fig. 5 (a), wireless channel can be modeled as a temporal linear filter with contributions from line-of-sight (LOS) path, reflecting, and even penetrating paths. Mathematically, it can be described in terms of a channel impulse response (CIR)

$$h(\tau) = \sum_{i=1}^{N} a_i e^{-j\eta_i} \delta(\tau - \tau_i), \tag{5}$$

where τ_i is the delay of the i-th path, and $\delta(\cdot)$ is the Dirac delta function. The corresponding CIR for the 4 illustrating paths of Fig. 5 (a) is shown in Fig. 5 (d), where the LOS path is the strongest with the smallest delay profile. Then the received signal $r(t)$ is the temporal convolution of the preamble $s(t)$ and $h(t)$: $r(t) = s(t) \circledast h(t)$.

Rather than directly obtaining the CIR, each Wi-Fi receiving RF chain estimates its equivalent channel frequency response (CFR) as $H(f) = S^{-1}(f)R(f)$ where $R(f)$ is the Fourier transform of $r(t)$ and similarly for $S(f)$. In commercial Wi-Fi devices, a group of sampled CFRs at a list of subcarriers are measured as

$$H(f_k) = |H(f_k)|e^{j\angle H(f_k)}, \; k = 1, 2, \ldots, K, \tag{6}$$

where f_k is the k-th frequency subcarrier; see Fig. 5 (c) for an illustration. For a typical 802.11ac Wi-Fi device, the bandwidth is about $B = 80$ MHz, resulting in a delay-domain resolution of 12.5 ns and a distance resolution of 3.75 meters. With L receiving antennas, the phase difference between two consecutive antennas $\Delta\phi_k = \angle\{H_l(f_k)H^*_{l+1}(f_k)\}$ is linearly proportional to the angle-of-arrival (AoD) profiles θ_i.

In summary, the CSI measurements may contain features related to path profiles fully described by the delay τ_i, the angle θ_i, and its propagation amplitude a_i.

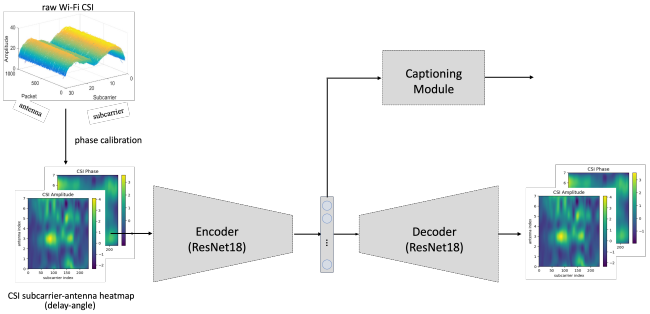


Fig. 6: Spatial-temporal feature extraction of calibrated CSI heatmaps over the antenna-subcarrier domain via an autoencoder.

4.2. Wi-Fi Spatial-Temporal Features Extraction

Our CSI feature extraction network is pretrained using a large amount of unlabelled CSI data (around 100K frames) that are totally separated from the Wi-Fi data recorded for captioning.

4.2.1. CSI Calibration

Since the CSI data are extracted from commercial routers (ASUS RT-AC86U in our case), it is known that the raw CSI data contain artifacts and noise due to power control uncertainty, sampling time offset (STO) between the transmitter and receiver, carrier frequency offset (CFO), and phase offsets between RF chains.

We leverage state-of-art amplitude and phase calibration procedure to mitigate the above hardware-induced distortion. Particularly, we follow a standard averaging operation for the amplitude calibration, and a phase fitting approach as used in the SpotFi approach [28, Algorithm 1] for the phase calibration. The amplitude- and phase-compensated CSI, referred to as the 2-channel CSI angle-delay heatmap $\mathbf{H} \in \mathbb{R}^{L\times K\times 2}$, can serve as an indirect spatial-temporal representation of the multi-path propagation.

4.2.2. CSI Feature Pretraining

To match with input dimension to the ResNet18 network, we further upsample the CSI angle-delay heatmap from $\mathbf{H} \in \mathbb{R}^{L\times K\times 2}$ to $\tilde{\mathbf{H}} \in \mathbb{R}^{224\times 224\times 2}$. As shown in Fig. 6, we then feed the augmented CSI heatmap (cyclic-shifted along the subcarrier or antenna axis) to an encoder to generate bottleneck feature maps $\mathbf{z} = \mathcal{E}(\tilde{\mathbf{H}})$, where the encoder is a 5-layer ResNet18 network (from conv1 to conv5) with a modified input layer accounting for the modified input channel dimension, and $\mathbf{z} \in \mathbb{R}^{7\times 7\times 512}$. The bottleneck feature $\mathbf{z}$ is fed to the decoder to reconstruct the input CSI heatmap: $\hat{\mathbf{H}} = \mathcal{D}(\mathbf{z})$, where $\mathcal{D}$ is a reversed ResNet18 network to gradually roll the bottleneck feature map back to the CSI angle-delay heatmap $\hat{\mathbf{H}} \in \mathbb{R}^{224\times 224\times 2}$.

To pretrain the autoencoder with unlabeled CSI data, we adopt a regularized mean squared error (MSE) as the loss function:

$$\mathcal{L}_{\text{AE}} = \sum_{i=1,2} \left\| \tilde{\mathbf{H}}_i - \hat{\mathbf{H}}_i \right\|_2^2 + \lambda \left\| \text{PD}(\tilde{\mathbf{H}}_2) - \text{PD}(\hat{\mathbf{H}}_2 \right\|^2, \tag{7}$$

where the subindex denotes the amplitude or phase channel of the CSI heatmap, and λ is the regularization weight. The regularization term uses a phase difference (PD) operator that averages out sequential phase differences between two consecutive antennas, resulting in a phase difference vector over subcarriers, and computes their distance between the input and reconstructed phase channels.

Table 1: Caption quality generated by multi-modal sensor combination.

	Direct sensors		Indirect sensors			Metrics		
	Video	Audio	Depth	Thermal	Wi-Fi	BLEU-4	METEOR	ROUGE_L
Baseline (Visual only)	✓					0.122	0.266	0.530
Audio-Visual	✓	✓				0.139	0.287	0.523
+Depth & Thermal	✓	✓	✓	✓		0.142	0.292	0.559
+Wi-Fi	✓	✓			✓	0.139	0.289	0.561
All sensors	✓	✓	✓	✓	✓	**0.147**	**0.293**	**0.567**
Depth & Thermal only			✓	✓		0.068	0.226	0.498
Wi-Fi only					✓	0.088	0.257	0.539
Indirect only			✓	✓	✓	0.097	0.267	0.531

Table 2: Caption quality with/without Inter-Modal Transfer Learning (IMTL) for indirect sensors.

	IMTL	BLEU-4	METEOR	ROUGE_L
Depth+Thermal		0.068	0.226	0.498
Depth+Thermal	✓	0.071	0.229	0.499
Wi-Fi only		0.088	0.257	0.539
Wi-Fi only	✓	0.101	0.261	0.543
All indirect features		0.097	0.267	0.531
All indirect features	✓	**0.113**	**0.277**	**0.535**

5. EXPERIMENTS

5.1. Indoor Monitoring Testbed

We collected the multimodal sensing information for nine sessions, each for 10 minutes, at one of three different stations in the same room. Humans captioned the events in the videos, and we used them for this study. The vocabulary size used for event captions was 214 after cutting-off words that occurred once. The training and test sets have 1330 and 457 event captions. We split the test data into 230 and 227 event captions for cross-validation. The data collection system consists of multiple commercial 802.11ac-compliant routers and devices in a configuration as described below for the downstream captioning task. The data collection system is deployed in standard indoor room settings. RGB with depth heat map images were captured at 30 fps using a stereo camera (Intel RealSense Depth Camera D455), audio was recorded at 16 kHz using an 8-microphone array (miniDSP UMA-8/USB Microphone Array), and thermal heat map images with 80x60 pixels were obtained using an infrared sensor (Mitsubishi Electric's MelDIR MiR8060).

CSI from 802.11ac devices: We use ASUS RT-AC86U AC2900 WiFi routers with 3 external and 1 internal antennas to extract the CSI measurements at 5 GHz and modified its firmware using the Nexmon CSI Extractor Tool of [29]. It allows per-frame CSI extraction for up to 4 spatial streams using all four receive chains on Broadcom and Cypress Wi-Fi chips with up to 80 MHz bandwidth in both 2.4- and 5-GHz bands. In addition, it supports MIMO antenna configurations, up to 4×4 spatial streams. In our in-house testbed, we use 2 pairs of Wi-Fi TX-RX settings (4 routers in total) in a diagonal configuration to record $M = 8$ spatial streams over $K = 234$ subcarriers. The two pairs share the same time clock with a workstation via the NTP server. The 5-GHz CSI are recorded in the routers and sent to the workstation via Ethernet cables.

5.2. Conditions

We evaluate our proposed approach with a newly collected in-house dataset for indoor human daily action. We extract video features with Omnivore [30], and audio features with the audio spectrogram Transformer (AST) [31]. We extract audio, RGB, Depth, and Infrared-thermal features with ImageBind [32]. The CSI features are extracted as described in Section 4.2.2. The video and image features are concatenated and projected to a single video feature sequence which is fed to the encoder. The decoder uses Glove word embedding [33] for initialization. The number of dimensions of the audio, visual, depth, thermal, and Wi-Fi features are 768, 1024, 1024, 1024, and 512, respectively, where the Wi-Fi features $\mathbf{z} \in \mathbb{R}^{7\times7\times512}$ are projected to 512-dimensional vectors before feeding them to the encoder.

The multimodal Transformer contains multiple encoders with two-layer blocks, where the multi-head attention dimension d^m_{model} for each modality m is the same as the corresponding feature dimension. The dimensions of the feed-forward layers are set as $d^m_{ff} = 4 \times d^m_{model}$. The caption decoder consists of two-layer blocks, where $d^D_{model} = 300$. The number of attention heads is 4 for all the Transformer layer blocks.

5.3. Results

Table 1 shows the quality of the event captions generated from the different models using BLEU-4, METEOR, and ROUGE_L scores. The performance using the indirect information is worse than that using the direct information. The teacher model using all sensors achieved the best performance for all metrics. Table 2 shows a comparison of results obtained with and without inter-modal transfer learning (IMTL), where we removed the direct features at testing. The relative improvement of the student model with only Wi-Fi leveraged by the teacher model using IMTL is 15% of BLEU-4. Combining depth and thermal information with Wi-Fi gains a further 12% BLEU-4.

6. CONCLUSIONS

This paper proposed a method for Wi-Fi-based indoor monitoring enhanced by multimodal fusion. We used a multimodal Transformer that converts multimodal features to event captions. Additionally, we utilize inter-modal transfer learning (IMTL) to transfer multimodal fusion power combining direct and indirect features to indirect feature sensor use only. Experiments with a newly collected in-house dataset for multimodal indoor monitoring demonstrated that our proposed method (All indirect features w/ IMTL) improves the quality of event captions by 28% in BLEU-4 score compared to those with Wi-Fi feature only. The evaluation results show the potential of indoor monitoring using only indirect sensors such as Wi-Fi. Future work includes further data collection to mitigate data sparseness and enhance the quality of pre-trained models for Wi-Fi.

References

[1] C. Hori and A. Vetro, "At last, a self-driving car that can explain itself," *IEEE Spectrum*, Feb. 2022.

[2] L. Fan, T. Li, Y. Yuan, and D. Katabi, "In-home daily-life captioning using radio signals," in *Proc. ECCV*, 2020.

[3] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with Wi-Fi!" in *Proc. MobiCom*, 2014, p. 593–604.

[4] D. Wu, D. Zhang, C. Xu, H. Wang, and X. Li, "Device-free WiFi human sensing: From pattern-based to model-based approaches," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 91–97, Oct 2017.

[5] W. Wang, A. X. Liu, and M. Shahzad, "Gait recognition using Wifi signals," in *Proc. ACM UbiComp*, 2016, p. 363–373.

[6] Y. Zeng, P. H. Pathak, and P. Mohapatra, "WiWho: WiFi-based person identification in smart spaces," in *Proc. IPSN*, Apr. 2016, pp. 1–12.

[7] H. Zou *et al.*, "WiFi-based human identification via convex tensor shapelet learning," in *Proc. AAAI*, 2018, pp. 1711–1718.

[8] Y. Zhang *et al.*, "Widar3.0: Zero-effort cross-domain gesture recognition with Wi-Fi," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[9] C. Li, M. Liu, and Z. Cao, "WiHF: Gesture and user recognition with WiFi," *IEEE Transactions on Mobile Computing*, vol. 21, no. 2, pp. 757–768, 2022.

[10] H. Zou *et al.*, "Deepsense: Device-free human activity recognition via autoencoder long-term recurrent convolutional network," in *Proc. ICC*, 2018, pp. 1–6.

[11] Z. Wang, K. Jiang, Y. Hou, W. Dou, C. Zhang, Z. Huang, and Y. Guo, "A survey on human behavior recognition using channel state information," *IEEE Access*, vol. 7, pp. 155 986–156 024, 2019.

[12] Y. Gu *et al.*, "Emosense: Data-driven emotion sensing via off-the-shelf WiFi devices," in *Proc. ICC*, May 2018, pp. 1–6.

[13] ——, "Besense: Leveraging WiFi channel data and computational intelligence for behavior analysis," *IEEE Computational Intelligence Magazine*, vol. 14, no. 4, pp. 31–41, Nov. 2019.

[14] Y. Chen, R. Ou, Z. Li, and K. Wu, "WiFace: Facial expression recognition using Wi-Fi signals," *IEEE Transactions on Mobile Computing*, vol. 21, no. 1, pp. 378–391, 2022.

[15] F. Wang, S. Zhou, S. Panev, J. Han, and D. Huang, "Person-in-wifi: Fine-grained person perception using WiFi," in *Proc. ICCV*, Jul. 2019.

[16] W. Jiang *et al.*, "Towards 3d human pose construction using WiFi," in *Proc. MobiCom*, 2020.

[17] F. Meneghello *et al.*, "Environment and person independent activity recognition with a commodity ieee 802.11 ac access point," *arXiv preprint arXiv:2103.09924*, 2021.

[18] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proc. ICCV*, Oct. 2017.

[19] V. Iashin and E. Rahtu, "A better use of audio-visual cues: Dense video captioning with bi-modal transformer," in *Proc. BMVC*, 2020.

[20] H. Alamri, V. Cartillier, A. Das, J. Wang, A. Cherian, I. Essa, D. Batra, T. K. Marks, C. Hori, P. Anderson *et al.*, "Audio visual scene-aware dialog," in *Proc. CVPR*, Jun. 2019.

[21] C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. G. Lopes, A. Das *et al.*, "End-to-end audio visual scene-aware dialog using multimodal attention-based video features," in *Proc. ICASSP*, May 2019, pp. 2352–2356.

[22] C. Hori, T. Hori, and J. Le Roux, "Low-latency online streaming VideoQA using audio-visual transformers," in *Proc. Interspeech*, 2022.

[23] A. Shah, S. Geng, P. Gao, A. Cherian, T. Hori, T. K. Marks, J. Le Roux, and C. Hori, "Audio-visual scene-aware dialog and reasoning using audio-visual transformers with joint student-teacher learning," in *Proc. ICASSP*, 2022, pp. 7732–7736.

[24] C. Hori, P. Peng, D. Harwath, X. Liu, K. Ota, S. Jain, R. Corcodel, D. Jha, D. Romeres, and J. Le Roux, "Style-transfer based Speech and Audio-visual Scene understanding for Robot Action Sequence Acquisition from Videos," in *Proc. Interspeech*, 2023, pp. 4663–4667.

[25] C. Hori, A. Cherian, T. K. Marks, and T. Hori, "Joint student-teacher learning for audio-visual scene-aware dialog," in *Proc. Interspeech*, 2019.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, Dec. 2017, pp. 5998–6008.

[27] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," in *Proc. NeurIPS Deep Learning Symposium*, 2016.

[28] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "Spotfi: Decimeter level localization using WiFi," in *Proc. SIGCOMM*, 2015, p. 269–282.

[29] F. Gringoli, M. Schulz, J. Link, and M. Hollick, "Free your CSI: A channel state information extraction platform for modern Wi-Fi chipsets," in *Proc. WiNTECH*, 2019, pp. 21–28.

[30] R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra, "Omnivore: A Single Model for Many Visual Modalities," in *Proc. CVPR*, 2022.

[31] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech*, 2021, pp. 571–575.

[32] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proc. CVPR*, Jun. 2023, pp. 15 180–15 190.

[33] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.