XModal-ID: Using WiFi for Through-Wall Person Identification from Candidate Video Footage

Belal Korany*
UC Santa Barbara
belalkorany@ece.ucsb.edu

Chitra R. Karanam*
UC Santa Barbara
ckaranam@ece.ucsb.edu

Hong Cai*
UC Santa Barbara
hcai@ece.ucsb.edu

Yasamin Mostofi
UC Santa Barbara
ymostofi@ece.ucsb.edu

ABSTRACT

In this paper, we propose XModal-ID, a novel WiFi-video cross-modal gait-based person identification system. Given the WiFi signal measured when an unknown person walks in an unknown area and a video footage of a walking person in another area, XModal-ID can determine whether it is the same person in both cases or not. XModal-ID only uses the Channel State Information (CSI) magnitude measurements of a pair of off-the-shelf WiFi transceivers. It does not need any prior wireless or video measurement of the person to be identified. Similarly, it does not need any knowledge of the operation area or person's track. Finally, it can identify people through walls. XModal-ID utilizes the video footage to simulate the WiFi signal that would be generated if the person in the video walked near a pair of WiFi transceivers. It then uses a new processing approach to robustly extract key gait features from both the real WiFi signal and the video-based simulated one, and compares them to determine if the person in the WiFi area is the same person in the video. We extensively evaluate XModal-ID by building a large test set with 8 subjects, 2 video areas, and 5 WiFi areas, including 3 through-wall areas as well as complex walking paths, all of which are not seen during the training phase. Overall, we have a total of 2,256 WiFi-video test pairs. XModal-ID then achieves an 85% accuracy in predicting whether a pair of WiFi and video samples belong to the same person or not. Furthermore, in a ranking scenario where XModal-ID compares a WiFi sample to 8 candidate video samples, it obtains top-1, top-2, and top-3 accuracies of 75%, 90%, and 97%. These results show that XModal-ID can robustly identify new people walking in new environments, in various practical scenarios.

*Co-primary student authors.

MobiCom '19, October 21–25, 2019, Los Cabos, Mexico

ACM ISBN 978-1-4503-6169-9/19/10…$15.00
https://doi.org/10.1145/3300061.3345437

CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile computing*; • **Computing methodologies** → *Modeling and simulation*; • **Hardware** → Wireless devices.

KEYWORDS

Through-wall, person identification, gait analysis, WiFi

ACM Reference Format:
Belal Korany, Chitra R. Karanam, Hong Cai, and Yasamin Mostofi. 2019. XModal-ID: Using WiFi for Through-Wall Person Identification from Candidate Video Footage. In *The 25th Annual International Conference on Mobile Computing and Networking (MobiCom '19), October 21–25, 2019, Los Cabos, Mexico.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3300061.3345437

1 INTRODUCTION

Person identification is an important problem that has been widely studied and implemented in various modalities, e.g., fingerprints, iris, and voice. Recently, there has been extensive work establishing that a person's gait can serve as a unique signature for identification [5]. Gait-based identification is attractive as it does not require a person to perform any specific active task (e.g., fingerprint scanning) and can automatically recognize a person based on his/her way of walking. This is very useful for many applications: smart buildings, personalized services, and security/surveillance.

Given the importance of gait-based person identification, there has been considerable research in using either videos or Radio Frequency (RF) signals to extract a person's gait for identification purposes. Vision-based approaches extract the walking person's silhouette and calculate various gait features to learn people's identities [5]. However, they require an unobstructed view of the person in good lighting and camera coverage everywhere, which are not always feasible. On the other hand, RF-based approaches are more versatile as RF signals can pass through walls/obstacles, and are not affected by lighting conditions. Additionally, RF signals are more ubiquitous due to the increasing presence of wireless devices. However, all existing RF-based gait identification approaches rely on extensive training with prior instances of the same people walking in the same area [30, 32, 33]. This significantly limits the practical use of this technology on

data of new people and in new locations. In addition to these technical limitations, RF-only approaches are not applicable to an important class of identification applications in the security domain, where, for instance, only a crime-scene video footage of a suspect that is being looked for is available.

In this paper, we propose a novel WiFi-video cross-modal person identification system, which we call **XModal-ID** (pronounced: *Cross-Modal-ID*). More specifically, given WiFi measurements of an unknown person walking in an unknown area, and the video footage of a walking person in another area, XModal-ID is able to determine whether it is the same person in both the WiFi area and the video footage. **One key characteristic of XModal-ID is that it does not require any prior wireless or video data of either the person to be identified or the area where the identification is to be conducted.** In other words, it does not need to be trained on prior WiFi or video data of the person being identified, or the identification area. It also does not need any knowledge of the test area or the person's track. Moreover, **it only uses CSI magnitude measurements of a pair of off-the-shelf WiFi transceivers**. Finally, **it can identify people through walls.** To the best of our knowledge, such a cross-modal gait-based identification system has not been studied before. This new technology can enable a wide range of new real-world applications that would not be possible with existing technologies. We next briefly describe two broad sets of applications that this system can be used for.

- **Security and Surveillance:** Consider the scenario where the footage of a crime is available and the police is searching for the suspect. A pair of WiFi transceivers outside a suspected hide-out building can use XModal-ID to detect if this person is hiding inside. Moreover, the existing WiFi infrastructure of public places can further be used to report the presence of the suspect. **To the best of our knowledge, there is currently no existing technology that can enable such applications.**
- **Personalized Services:** Consider a smart home, where each resident has personal preferences (e.g., lighting, music, and temperature). The home WiFi network can use XModal-ID and one-time video samples of the residents to recognize the person walking in any area of the house and activate his/her preferences, **without the need to collect wireless/video data of each resident for training purposes.** New residents can also be easily identified without a need for retraining. This is in contrast to the existing technologies that would require training with the wireless data of every resident collected in all areas of the house.

In order to achieve such cross-modal identification capabilities, XModal-ID compares the gait characteristics of a given WiFi measurement to that of a given video footage, and deduces their similarity. More specifically, given the video footage of a walking person, XModal-ID constructs a 3D mesh of the person from the video and then calculates the corresponding WiFi signal that would have been generated by this person walking in the area where a pair of WiFi transceivers are present (it does so without any knowledge of the person's track or the area). It then compares this simulated WiFi signal to the real WiFi signal measured in the area where the person-of-interest walks. Based on the similarity between the simulated WiFi signal and the real WiFi one, the system determines whether the person walking in the WiFi area is the same person in the video. Once XModal-ID is trained on a pool of data, it can be deployed in any new, unseen area and can perform cross-modal identification of new people, of whom it has no prior knowledge during training. We next explicitly discuss the contributions of this paper.

Statement of Contributions:

1. We propose a new approach to simulate the WiFi signal that would have been measured by a pair of transceivers, based on the video footage of a person walking. More specifically, we extract a 3D mesh model of the person in the video and apply Born approximation to simulate the corresponding WiFi CSI magnitude measurements if the person in the video was walking in a WiFi area.
2. We propose a new framework and set of features that capture the gait characteristics of a person based on WiFi CSI magnitude signals. More specifically, we utilize a combination of Short-Time Fourier Transform and Hermite functions to generate a spectrogram, and extract a key set of features that are subsequently used for identification. We further propose a way to extract key parts of the spectrogram as well as the direction of motion, which allows us to do identification, without the need to know the track of the person.
3. We extensively evaluate our proposed framework using a large test set, where all the test subjects and test areas are completely unknown in the training phase, thus allowing us to demonstrate the generalizability of XModal-ID to new, unknown people and environments. In the test set, there are 8 subjects, 2 video areas, and 5 WiFi areas, including 3 areas where the transceivers are placed behind a wall and scenarios where the walking paths are complex. The walking paths are further assumed unknown in all the experiments. Overall, the test set contains a total of 2,256 pairs of WiFi and video samples to be identified. Given a pair of video and WiFi samples, XModal-ID achieves a binary classification accuracy of 85%, in judging whether the two samples belong to the same person. Furthermore, given a queried WiFi data sample and 8 candidate video samples, XModal-ID achieves top-1, top-2, and top-3 accuracies of 75%, 90%, and 97%, respectively, in ranking the video samples.

We discuss the current limitations and future extensions of our system in Sec. 8.

2 RELATED WORK

Existing gait-based identification work can be broadly classified into two categories: RF-based and video-based.

2.1 RF-Based Person Identification

RF-based approaches utilize RF signals to deduce information about the gait of a person. RF signals from the transmitter reflect off of different parts of the body of a walking person and reach the receiver, thereby implicitly carrying information about the movement of various body parts.

Radar-based: Various radar-based approaches have utilized dedicated hardware and/or wideband signals for gait analysis. For instance, in [19, 26], the authors utilize radar signatures to extract stride rate and velocities of different body parts. Orovic et al. [20] classify various body part motions using the received radar signals and Hornsteiner et al. [12] characterize the gait features in a time-frequency analysis using a 24 GHz radar. In [27], the authors use a 77 GHz radar to extract micro-Doppler signatures from a walking person for identification.

WiFi-based: Recently, there has been considerable interest in using off-the-shelf WiFi devices for gait-based person identification. WiFiU [30] uses WiFi CSI to generate spectrogram-based gait features, which are then used to classify the identities of a pre-defined set of people. WiWho [32] uses the time-domain signals measured during people's motion to identify people. Similarly, a few other papers [18, 31, 33, 35] identify a person from a priorly-known set of people. In addition to walking, Wang et al. [29] show that respiration patterns can also be used for identification. WiID [34] uses the CSI measured while a person performs several actions for identification, using two links in the area. Shi et al. [23] identify a person based on his/her daily habits. All these existing approaches require the transceivers to be in the same area as the person, with a line-of-sight view at all times, with the only exception of Hoble [17], which uses a Software Defined Radio to identify people in both line-of-sight and through-wall settings in a known area.

All these existing RF-based papers identify people from a pre-defined group and require prior wireless measurements of these people for training. In other words, they cannot handle new people without retraining. They also require the training and test walking paths/actions and locations to be the same. Thus, a model that is trained in one location and on one type of path cannot be used in other scenarios. The radar-based approaches further require extensive hardware setup. Moreover, aside from [17], none of the existing methods have through-wall identification capabilities. In this paper, on the other hand, we propose a novel person identification system that does not require training with prior measurements of the subjects/areas, does not require the test areas/tracks to be known, and can identify people through walls. **Finally, our proposed system enables a new set of applications not possible before, i.e., given a video footage of a person, it can detect if this person is present in a WiFi area.**

2.2 Video-Based Person Identification

Video-based person identification using gait is a well-studied problem in the computer vision literature. There are broadly two types of approaches: model-based, where gait features are extracted by fitting a walking human model to the video frames, and model-free, where features are extracted directly from the video. Model-based methods include fitting a walking human with ellipses [16], estimating the lengths of body parts and joint angles [28], and estimating the joint trajectories [25]. Model-free approaches rely on the person's silhouette in the video. Commonly-used features include the silhouette key frames [4] and gait energy image [9]. These features are then fed into machine learning pipelines for training. We refer the readers to [5] for a detailed survey. Overall, video-based methods require installing cameras everywhere and lack through-wall identification capabilities.

3 PROBLEM FORMULATION AND SYSTEM OVERVIEW

In this paper, we propose a WiFi-video cross-modal person identification system. More specifically, given the WiFi CSI magnitude measurements of a pair of WiFi transceivers, obtained in an area where a person is walking, and the video footage of a person in another area, we propose a system that determines whether this given pair of video and WiFi measurements correspond to the same person or not. Unlike existing RF-based person identification systems, our system does not need prior wireless or video measurements of the person-of-interest for training purposes. It further does not need prior measurements in the operation environments.

The overall architecture of our proposed XModal-ID system and the various steps involved in the pipeline are shown in Fig. 1, and briefly described below:

- Given the video footage of a person, we construct a 3D mesh model of the person. We then propagate this mesh model over time and use Born approximation to simulate the corresponding received WiFi signal if the person was walking near a pair of WiFi transceivers. We then use the signal magnitude to generate the spectrogram of the signal, using Short-Time Fourier Transform (STFT). It is noteworthy that we do not need to know the track of the person or details of the operation area.
- In the operation area where a person is walking, a WiFi receiver (Rx) measures the CSI magnitude of the received signal in the transmission from a WiFi transmitter (Tx). We then generate the corresponding spectrogram from this CSI magnitude measurement (using a combination

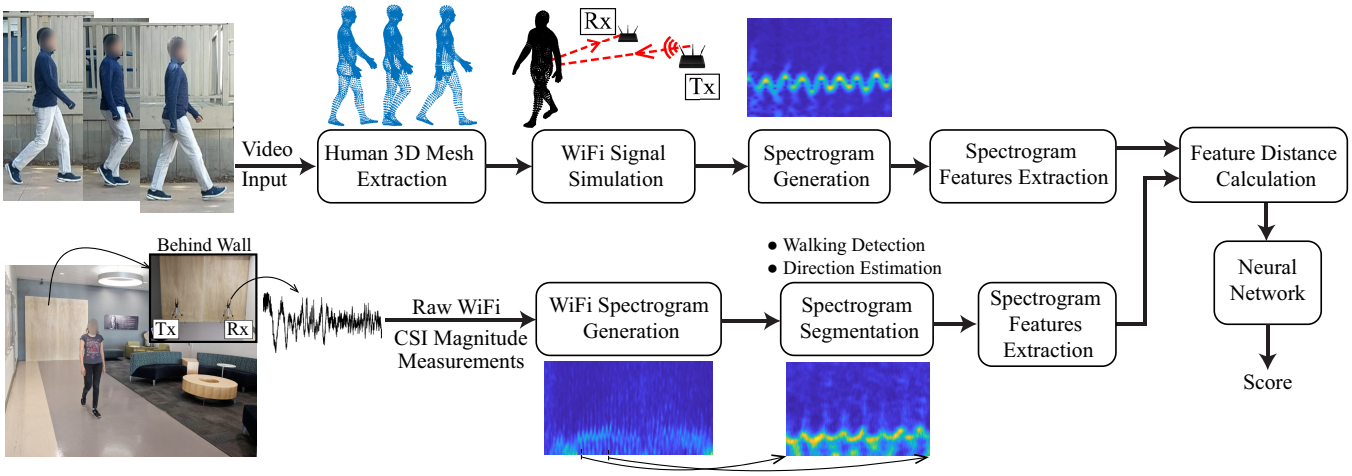


Figure 1: System architecture showing the various steps involved in the video and WiFi pipelines of XModal-ID. We refer readers to the color pdf for optimal viewing of the sample spectrograms.

Figure 2: (Right) Three sample HMR algorithm output meshes for (left) different snapshots of a walking person.

of STFT and Hermite functions) and segment it to obtain the parts most informative for identification, and further estimate the direction of motion.

- We then show how to extract key features from the spectrograms generated from both the WiFi and video data, and calculate the distance between them. The feature distances of a training set are fed into a small 1-layer neural network, which, after training, outputs a score indicating the similarity between any pair of real and simulated spectrograms, thus indicating if the person in a video is the same person in a WiFi area.

4 PROPOSED XMODAL-ID SYSTEM

In this section, we lay out the details of our proposed system, which is shown in Fig. 1. We first show how we can use a video footage of a walking person to generate a simulated wireless signal, which would have been measured if that person walked in a WiFi-covered area. Then, we show how to process the raw WiFi magnitude data measured in a real WiFi-covered area in which a person is walking. We mathematically model the wireless signals reflected from the person's body and apply time-frequency analysis techniques to generate a *spectrogram*, which captures the gait attributes of the person. We further focus on extracting the informative parts of the spectrogram as well as the direction of motion. We finally show how we can utilize the simulated wireless signal from the video footage to generate a corresponding spectrogram of the person based on the video. In Sec. 5, we then introduce a set of key features and show how we can use them to quantify the similarity between the two spectrograms to determine if they belong to the same person or two different people.

4.1 Video-to-WiFi Gait Modeling

In this section, we show how we can use a video footage of a walking person to generate a simulated WiFi signal, which would have been measured by a pair of WiFi transceivers if this person walked in their vicinity. Note that we do not assume that the real WiFi transceivers are in the same area where the video footage was taken.

Given one video frame (snapshot) of a person, we first utilize the Human Mesh Recovery (HMR) algorithm of [13] to produce a dense 3D mesh, which contains a large number of 3D points describing the outer surface of the human body. Given a video clip of a person, we then construct a set of 3D points for each frame. The sequence of such sets then captures the gait of the person. Fig. 2 shows a few sample video snapshots with their corresponding 3D mesh models.

Denote by $\mathcal{M}(t) = \{\mathbf{x}_m(t) \in \mathcal{R}^3, m = 1, \ldots, M\}$ the set of generated 3D mesh points of the human body at time t.[1] In the real WiFi environment, a WiFi Tx is located at $\mathbf{x}_T \in \mathcal{R}^3$, and a WiFi Rx is located at $\mathbf{x}_R \in \mathcal{R}^3$, as shown in the bottom row of Fig. 1. In order to simulate the WiFi signal that would have been received if the person in the video was walking in the WiFi area, we utilize the Born approximation [3] to

[1]Note that the HMR method outputs 3D points in the pixel space. Transforming these points to real-world 3D coordinates only requires a one-time calibration of the camera upon fixation, using the coordinates of a few known points in the real world that are identified within the camera frame. See [10] for more details.

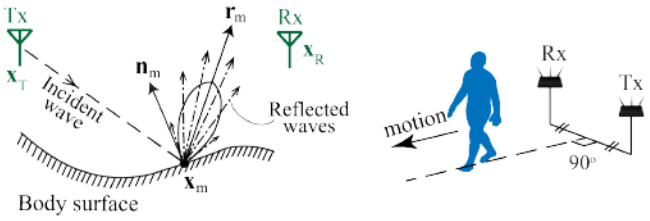


Figure 3: (a) Quasi-specular reflection model of the human body. An incident wave on a point $\mathbf{x}_m$ on the body is reflected to different directions with different amplitudes, with the strongest reflection being in the direction $\mathbf{r}_m$ determined by the normal to the body surface at $\mathbf{x}_m$. (b) Walking path of the generated human mesh to simulate the WiFi signal.

model the WiFi reflections off of the generated human mesh surface. More specifically, the simulated received WiFi signal at time t can be written as,

$$s_v(t) = \underbrace{g(\mathbf{x}_T, \mathbf{x}_R)}_{\substack{\text{direct signal} \\ \text{from Tx to Rx}}} + \sum_{m \in \mathcal{M}'(t)} \underbrace{A_m G_m g(\mathbf{x}_T, \mathbf{x}_m) g(\mathbf{x}_m, \mathbf{x}_R)}_{\text{reflected signal from point } \mathbf{x}_m}, \quad (1)$$

where $g(\mathbf{x}, \mathbf{y})$ is the Green's function from point $\mathbf{x}$ to point $\mathbf{y}$ in $\mathcal{R}^3$, and is given by $g(\mathbf{x}, \mathbf{y}) = \frac{\exp(j\frac{2\pi}{\lambda}\|\mathbf{x}-\mathbf{y}\|)}{4\pi\|\mathbf{x}-\mathbf{y}\|}$, where $\|.\|$ is the Euclidean norm of the argument, and λ is the wavelength of the wireless signal. $\mathcal{M}'(t) \subset \mathcal{M}(t)$ is then the subset of all points in the human mesh that are visible to both the Tx and Rx, since only these points will reflect the signal to the Rx. We determine $\mathcal{M}'(t)$ by applying the Hidden Point Removal (HPR) algorithm [15] to $\mathcal{M}(t)$.

The strength of the signal reflected from point $\mathbf{x}_m$ is determined by two factors: the surface area and the orientation of the body part to which $\mathbf{x}_m$ belongs. For instance, the human torso has a higher reflectivity than the other body parts since it has a larger surface area. This factor is captured by the scale A_m. The orientation of the body part then determines the direction in which an incident signal would be reflected. A perfect reflector would reflect the incident wave at $\mathbf{x}_m$, only in the direction $\mathbf{r}_m = \frac{\mathbf{x}_m - \mathbf{x}_T}{\|\mathbf{x}_m - \mathbf{x}_T\|} - 2\frac{(\mathbf{x}_m - \mathbf{x}_T)^T \mathbf{n}_m}{\|\mathbf{x}_m - \mathbf{x}_T\|}\mathbf{n}_m$, where $\mathbf{n}_m$ is the normal vector to the body at point $\mathbf{x}_m$ (see Fig. 3 (a)). However, the human body is best modeled as a *quasi-specular reflector* [1], which reflects the signal into many directions with different amplitudes, with the strongest in the $\mathbf{r}_m$ direction (as shown in Fig. 3 (a)). The amplitude of the reflection from $\mathbf{x}_m$ towards the Rx will then be inversely related to the angle between the vectors $\mathbf{x}_R - \mathbf{x}_m$ and $\mathbf{r}_m$. Based on our empirical studies, we capture this relation using a Gaussian mask $G_m = \exp\left(-\left(\cos^{-1}\frac{(\mathbf{x}_R - \mathbf{x}_m)^T \mathbf{r}_m}{\|\mathbf{x}_R - \mathbf{x}_m\|}\right)^2 / 2\sigma_a^2\right)$.

We simulate the received wireless signal for the case where the person in the video is walking away from the link, on the line that is the perpendicular bisector of the Tx-Rx link, as shown in Fig. 3 (b). We shall see in Sec. 4.3 why we do not need to know the real track of the person in the WiFi area and that simulating the receptions on only the aforementioned path will be sufficient for our XModal-ID system.

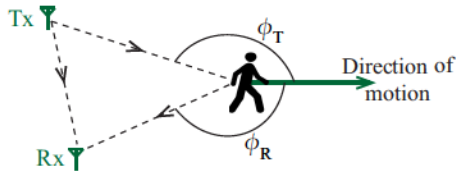


Figure 4: A pair of WiFi transceivers are used to identify the person.

4.2 WiFi-Based Gait Modeling

In this section, consider the WiFi-covered area where a person is walking, as shown in Fig. 4. A WiFi Tx emits a wireless signal that reflects off of different parts of the human body and is received by a WiFi Rx. The complex baseband received signal $s_b(t)$ can be written as follows [14],

$$s_b(t) = \alpha_s e^{j\theta_s} + \sum_m \alpha_m e^{j\left(\frac{2\pi}{\lambda}\psi v_m(t)t + \frac{2\pi}{\lambda}d_m\right)}, \quad (2)$$

where $\alpha_s e^{j\theta_s}$ is the complex received signal including the impact of both the direct path and the static paths, α_m is the amplitude of the signal path reflected off of the m^{th} part of the body, d_m is the length of that path at time $t = 0$, $v_m(t)$ is the speed of the m^{th} body part at time t, and $\psi = \cos\phi_R + \cos\phi_T$ where ϕ_R and ϕ_T are as illustrated in Fig. 4.

Denoting by $s(t)$ the magnitude square of the baseband signal $s_b(t)$ and assuming that $|\alpha_s| \gg |\alpha_m|$, $s(t)$ can be written as follows,

$$s(t) = P + \sum_m \beta_m \cos\left(\frac{2\pi}{\lambda}(\psi v_m(t)t + d_m) - \theta_s\right), \quad (3)$$

where $P = |\alpha_s|^2 + \sum_m |\alpha_m|^2$ is the DC component of $s(t)$ and $\beta_m = 2|\alpha_s\alpha_m|$. Note that ψ can be time-varying.

4.3 Spectrogram Generation Based on Measured Wireless Signals

It can be seen from Eq. 3 that the signal $s(t)$ is the sum of multiple sinusoids whose frequencies are linearly related to the respective speeds of different body parts of the moving person. Hence, estimating the instantaneous frequency components of the signal $s(t)$ provides information about how the person walks. To this end, we utilize the Short-Time Fourier Transform (STFT), which is a commonly-used time-frequency analysis technique in the RF-based gait analysis literature. In STFT, a short moving window of length T_{win} is applied to $s(t)$ and the Fourier Transform is applied to each instance of the moving window to estimate the frequency components, resulting in a signal *spectrogram*. More

specifically, we have,

$$STFT(t,f) = \left| \int_{t}^{t+T_{\text{win}}} s(\mu) e^{-j2\pi f \mu} d\mu \right|. \quad (4)$$

Fig. 5 (a) shows a sample STFT spectrogram of a walking person, which is generated from the received WiFi signal when a person walks away from a WiFi link, on a path perpendicular to it. A strong reflection (indicated by brighter colors) can be seen in the spectrogram at ~25 Hz, which corresponds to a speed of 0.72 m/s. This is caused by the motion of the torso, the reflection of which is stronger due to its larger surface area. Weaker reflections (indicated by darker colors) of the faster body parts (e.g., legs) appear periodically at higher frequencies in the spectrogram.

While the STFT provides valuable information about the instantaneous speeds of different body parts, it has been shown in the literature that the corresponding time-frequency resolution trade-off can affect the quality of this information [20]. Multi-window Hermite Spectrograms (HS) were then proposed, in the Radar literature [20], to improve the concentration of STFT spectrograms. In a Hermite spectrogram, multiple Hermite functions are used as windows for the time-frequency analysis. More specifically,

$$HS(t,f) = \frac{1}{2\pi} \sum_{k=0}^{K-1} b_k(t) \left| \int s(\mu) \chi_k(\mu - t) e^{-j2\pi f \mu} d\mu \right|^2, \quad (5)$$

where $\chi_k(t)$ is the k^{th} Hermite function, and $b_k(t)$ are weighting coefficients obtained by solving the system

$$\sum_{k=0}^{K-1} b_k(t) \frac{\int |s(t+\mu)|^2 \chi_k^2(\mu) \mu^{n-1} d\mu}{\int |s(t+\mu)|^2 \chi_k^2(\mu) d\mu} = \begin{cases} 1, & \text{if } n = 1 \\ 0, & \text{if } n \in \{2, \ldots, K\} \end{cases} \quad (6)$$

Fig. 5 (b) shows a Hermite spectrogram (with $K = 3$ Hermite functions) generated from the same data as the STFT spectrogram in Fig. 5 (a). These two transformations, however, are utilized independently in different literature. In order to combine the desirable concentration properties of the HS and the ability of STFT to detect minute reflections from different body parts, we propose to generate the final WiFi spectrogram $S(t,f)$ by combining the two spectrograms as follows,

$$S(t,f) = STFT(t,f) + HS(t,f). \quad (7)$$

Essentially, $S(t,f)$ is a multi-window spectrogram that utilizes the rectangular window as well as the hermite function windows. We have observed that this combination considerably improves the visibility of the gait information in the Fourier domain. We then normalize the resulting spectrogram at each time instant, with respect to the sum of the values over all the frequencies at that time instant.

To visualize the impact of combining the spectrograms, consider the spectrograms shown in Fig. 5 (a) for STFT and Fig. 5 (b) for HS. As can be seen, the reflections of the person's limbs are clearly visible in the STFT, while the concentration of the torso reflection is clearer in the Hermite spectrogram. Consequently, the combined spectrogram in Fig. 5 (c) captures both these aspects of the gait. Additionally, Fig. 5 (d) shows the combined spectrogram of another subject walking on the same path, showing differentiable gait attributes.

Remark 1. *From Eq. 3, a detected reflection at a frequency f in the spectrogram is caused by a moving object with the speed $v = f\lambda/\psi$. Hence, static multipath due to the background environment appears at $f = 0$ and thus does not affect the gait motion information, which appears at non-zero frequencies.*

Remark 2. *We extract the gait information from the frequency of the reflected signal, and not from its power. Hence, as long as the power of the reflected signal is above the noise floor, the gait information can be extracted from the spectrogram. This is particularly attractive for through-wall settings, where the wall attenuates the signal power, but does not affect the gait motion information.*

4.3.1 **Spectrogram Segmentation:** As described in Eq. 3, two parameters determine the instantaneous frequencies of the different sinusoidal components of $s(t)$: the direction of motion (represented by ψ) and the instantaneous speeds of different body parts (v_m). In this section, we describe how we segment the spectrogram $S(t,f)$ and extract the part in which ψ can be considered constant. When the WiFi Tx and Rx are close to each other, as compared to the distance of the person to the link, a segment with an approximately constant ψ is obtained whenever the person walks on a straight line towards or away from the midpoint of the Tx-Rx line. In such a segmented spectrogram, the frequency information mainly contains the gait attributes of the person, since it depends only on the speeds of the different body parts (v_m). **As such, it can be very informative for person identification, without requiring the knowledge of the track of the person.** We henceforth refer to such a segment as a *constant-ψ segment*, and utilize it for our XModal-ID system. Note that we do not need the whole track to be on a straight line towards/away from the midpoint of the link. The person can take any track, and as long as there is even a small part of the track (e.g., 3 sec or longer) that satisfies this condition, then the proposed approach can be utilized.

In order to extract a constant-ψ segment from the spectrogram, we search for a segment (with a minimum width of T_{min}) that satisfies two conditions. First, the spread of the energy distribution across frequency inside the segment, $V(t) = \int f^2 S(t,f) df - \left(\int f S(t,f) df \right)^2$, should remain below a certain threshold V_{th}, since a higher value of $V(t)$ indicates that the spectrogram is close to being flat at time t, which implies that there is no walking detected within this

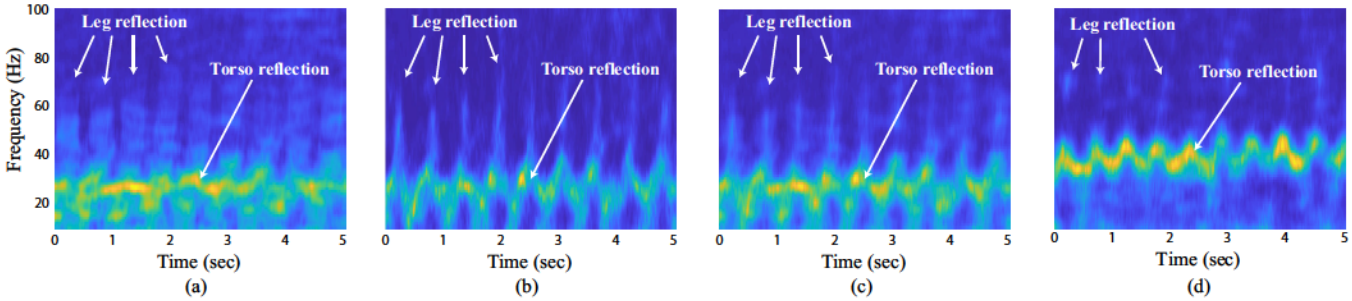


Figure 5: (a) Spectrogram based on Short-Time Fourier Transform (STFT) for a person walking away from the link, on a straight line perpendicular to the Tx-Rx line. (b) Spectrogram of the same data based on the Hermite method. (c) Combined Spectrogram $S(t, f)$ of STFT and the Hermite method. (d) Combined spectrogram $S(t, f)$ of another person walking on the same path. It can be seen from (c) and (d) that the combined spectrograms of different people are well differentiable, e.g., the torso speed, leg speed, and gait cycles are different. See the color pdf for optimal viewing of the spectrograms.

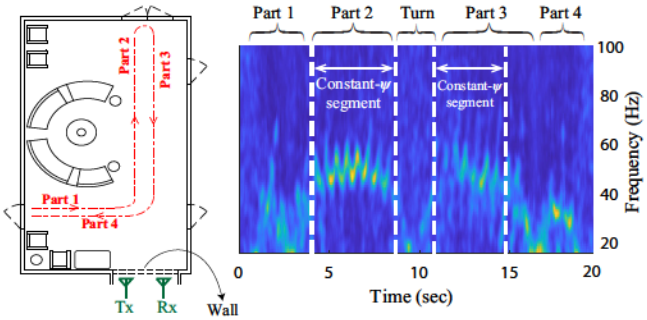


Figure 6: Working example of the spectrogram segmentation algorithm. (Left) A floor plan with a 4-part path where a person walks (experiment area of Fig. 9 (g)), with WiFi Tx-Rx placed behind a wall. (Right) The spectrogram of the measured WiFi data, showing different parts of the walk. The dashed lines show two extracted constant-ψ segments.

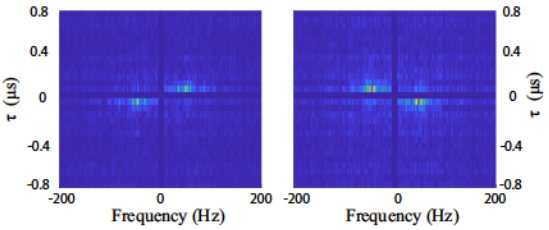


Figure 7: Plots of $Z(f, \tau)$ when a person is walking (left) towards the link, and (right) away from the link. Energy distribution of Z over the four quadrants indicates the motion direction. See the color pdf for better viewing.

segment. Secondly, the variations of the average torso speed within this segment should remain below a certain threshold v_{th}. Since the average torso speed of a walking person is constant in a small time window, a varying average torso speed in the spectrogram is due to a varying ψ. The average torso speed can be calculated from the spectrogram, as we shall see in Sec. 5. When a segment satisfies the aforementioned conditions, it is declared as a constant-ψ segment.[2]

Next, we consider what would be a good value for T_{min} (the minimum acceptable width of the segment). A small T_{min} would result in many false positives, in which ψ could be falsely considered constant simply because the segment was too short. On the other hand, a large T_{min} would require the person to walk for a long time in order to be identified. We observe that using $T_{\text{min}} = 3$ sec is a good trade-off, which provides a sufficient number of gait cycles (for a casual walk) for extracting meaningful gait features.

Fig. 6 shows an example of the spectrogram segmentation algorithm for the walking experiment depicted on the left, where the constant-ψ portion corresponds to parts 2 and 3 of the track. The figure on the right shows the un-segmented spectrogram $S(t, f)$ of the entire walking experiment, as well as the constant-ψ segments detected by our algorithm.

4.3.2 **Walking Direction Estimation:** We have observed that the segments of the spectrogram corresponding to a person walking away from the link have clearer gait patterns than those corresponding to walking towards the link (for instance, compare parts 2 and 3 in Fig. 6). Similar observations have been made in [22]. Therefore, we propose to utilize only the spectrogram segments corresponding to when the person is walking away from the link in our subsequent processing pipeline. Let $S_w(t, f)$ denote a constant-ψ spectrogram segment detected by the spectrogram segmentation algorithm. The information about the direction of motion, i.e., whether the person is walking towards or away from the link, is theoretically contained in the sign of ψ. However, this information cannot be extracted from $S_w(t, f)$ since both a positive and a negative ψ would result in the same spectrogram, given that we only use signal magnitude measurements. In this section, we then propose a new method that can estimate the walking direction.

Despite the absence of the information about the sign of ψ in $S_w(t, f)$, we can still determine the walking direction by exploiting the fact that a WiFi signal spans frequencies in the

[2]Note that there can be multiple constant-ψ segments in one spectrogram, depending on the track of the person, which we assume unknown.

band $[f_c - B/2, f_c + B/2]$, where f_c is the carrier frequency and B is the WiFi bandwidth, as we shall see next. Based on Eq. 3, the magnitude squared WiFi signal, $s(t;\rho)$, measured in a very short time on a frequency range $f_c + \rho$, for $\rho \in [-B/2, B/2]$, can be written as,

$$\begin{aligned} z(t;\rho) &= s(t;\rho) - P \\ &= \sum_m \beta_m \cos\left(\frac{2\pi}{c}(f_c+\rho)(v_m\psi t + d_m) - \theta_s\right) \\ &\approx \sum_m \beta_m \cos\left(\frac{2\pi}{c}(f_c v_m \psi t + f_c d_m + \rho d_m) - \theta_s\right), \end{aligned} \quad (8)$$

where c is the speed of light. The approximation in the last line of Eq. 8 relies on the fact that, in a very short time window, the product ρt is negligible as compared to the other terms in the cosine argument. By taking the Fourier Transform of $z(t;\rho)$ along the t dimension and the inverse Fourier Transform along the ρ dimension, we get,

$$\begin{aligned} Z(f;\tau) &= \left| \iint z(t;\rho) e^{-j2\pi f t} e^{j2\pi\tau\rho} dt d\rho \right| \\ &= \sum_m \frac{\beta_m}{2}\left(\delta\left(f - \frac{v_m\psi}{\lambda_c}, \tau - \frac{d_m}{c}\right) + \delta\left(f + \frac{v_m\psi}{\lambda_c}, \tau + \frac{d_m}{c}\right)\right), \end{aligned} \quad (9)$$

where $\delta(.,.)$ is the 2D Dirac Delta function, and $\lambda_c = c/f_c$. By examining Eq. 9, we can see that for a positive ψ (a person moving towards the link), the components of Z lie in the first and third quadrants of the (f,τ) space, while, for a negative ψ (a person moving away from the link), the components of Z lie in the second and fourth quadrants of the (f,τ) space. Therefore, we can determine the walking direction of the person according to the following decision rule,

$$\frac{\int_0^\infty \int_0^\infty |Z(f;\tau)|^2 df d\tau}{\int_0^\infty \int_{-\infty}^0 |Z(f;\tau)|^2 df d\tau} \underset{\text{away}}{\overset{\text{towards}}{\gtrless}} 1 . \quad (10)$$

Fig. 7 shows an example of the direction estimation output $Z(f;\tau)$ when a person is walking towards the link (left), and away from the link (right). It can be clearly seen that the energy of $Z(f;\tau)$ is concentrated in the first and third quadrants for the former case, and in the second and fourth quadrants for the latter case.

4.4 Video-Based Spectrogram Generation

In Sec. 4.1, we proposed a way of simulating the wireless signal based on the video footage of the person, using the HMR algorithm, HPR algorithm, and Born approximation. Since we are only interested in the constant-ψ parts of the spectrogram of the real WiFi measurement, as discussed in Sec. 4.3, we only need to simulate the corresponding wireless signal of Eq. 1 when a person walks away from the link, on the line that is the perpendicular bisector of the link (see Fig. 3 (b)), while being far enough from the link, as compared

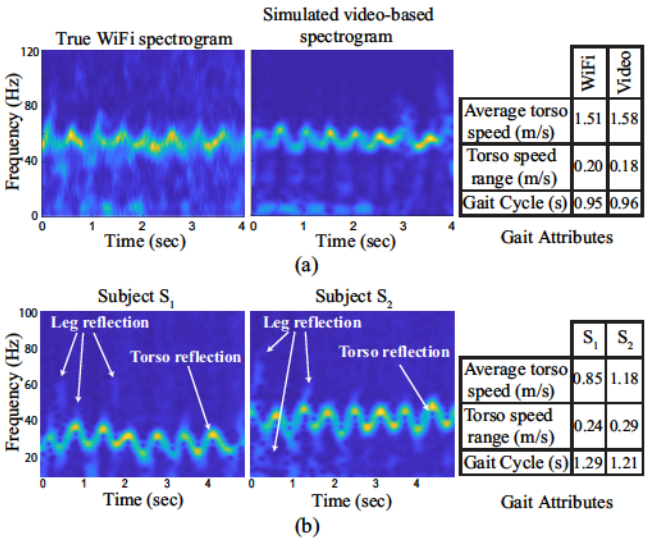


Figure 8: (a) Spectrograms of real WiFi data and of the video-based simulated one for the same person, showing similar gait attributes. (b) Video-based simulated spectrograms of two different people, showing their distinct gait attributes.

to the distance between the Tx and Rx (e.g., at least 2 m when the Tx and Rx are 1.5 m apart). As such, no knowledge of the track or operation area is needed. After the calculation of $s_v(t)$ using Eq. 1 on the aforementioned path, a spectrogram of the simulated WiFi signal $S_v(t,f)$ is generated from $|s_v(t)|^2$ via STFT, using the procedure described by Eq. 4.

To validate our framework of video-based spectrogram generation, we conduct a preliminary experiment where a subject walks away on a path that perpendicularly bisects the WiFi link, while simultaneously being videotaped. Fig. 8 (a) shows the spectrogram of the real WiFi data as well as the video-based simulated spectrogram. The figure further shows some gait attributes extracted from both spectrograms. It can be seen that our video-based spectrogram closely resembles the real WiFi data spectrogram, demonstrating the accuracy of our proposed framework. Additionally, Fig. 8 (b) shows two sample video-based spectrograms of two different subjects. It can be seen that the gait attributes of the two subjects are well differentiable in the two spectrograms.

Next, we show how to measure the similarity between the video-based simulated spectrogram and the spectrogram obtained from the real measured WiFi data, in order to identify whether the video and the WiFi data correspond to the same person or two different people.

5 FEATURE EXTRACTION AND SIMILARITY PREDICTION

So far, we have described our approach to extract spectrograms from both WiFi data and video data. In order to determine whether a WiFi sample and a video sample belong to

the same person, we extract several features from the corresponding spectrograms. We then compute a set of distances between the WiFi-based and video-based features. Given these feature distances, we train a simple 1-layer neural network to properly combine the distances and determine if a pair of WiFi and video samples belong to the same person. After training, the network not only provides this binary prediction, but also provides a score indicating the similarity between the WiFi and video samples. Once the network is trained, we use it on unseen WiFi and video data. In other words, none of the test data and locations is used for training.

5.1 Spectrogram Features

We have identified 12 features that are key for capturing the main characteristics of a person's gait. We compute each feature on both the WiFi and video spectrograms, and use a distance metric to measure the difference between the two spectrograms with respect to each feature. More specifically, we look at the frequency and time dimensions of the spectrogram, which carry different types of gait signatures that can be used for identification, as we describe next.

The frequency dimension carries information about the speeds of different body parts. We use the following frequency-related features:

• **Frequency distribution (FD):** This feature is obtained by averaging the spectrogram over time. This feature captures the distribution of frequency components, or equivalently, the speeds of different body parts, during the person's walk.

• **Frequency distribution in 4 gait phases (FD4):** Similar to the previous feature, we calculate the time-average of the spectrogram for each of the 4 phases of the gait cycle, resulting in 4 corresponding feature vectors [30].

• **Average torso speed:** We calculate the average of the torso speed curve, which can be extracted from the spectrogram using the method in [30].

• **Average of the range of torso speed:** After extracting the torso speed curve, we calculate the range of the torso speed variation in one gait cycle. This range is then averaged over all the gait cycles in the spectrogram.

The time dimension carries temporal information (e.g., periodicity patterns) about a person's gait. We capture the temporal signatures using the following features:

• **Autocorrelation (AC):** Given a spectrogram, we compute the autocorrelation across time (with a maximum lag of 2 sec) for each frequency bin, resulting in an autocorrelation matrix. We then compute a weighted sum of this matrix over the frequency dimension based on the energy distribution over the frequencies. This feature carries information on the gait cycle and the periodicity of the walk.

• **FFT of spectrogram over time:** Similar to the method in [22], we calculate the Fast Fourier Transform (FFT) of the spectrogram over time for each frequency bin. We then compute a weighted sum over the frequency dimension based on the energy distribution over the frequencies.

• **Histogram of autocorrelation gradient:** This is the histogram of the gradient of the AC feature vector.

• **Histogram of torso speed gradient:** We calculate the histogram of the torso speed gradient, which carries information on how the torso speed changes with time.

• **Stride length:** This is obtained by multiplying the average torso speed by the gait cycle length, which can be extracted from the torso speed curve.

Given the 12 features of a WiFi spectrogram and the 12 features of a video-based simulated spectrogram, we compute the distance between each corresponding feature in WiFi and video. This results in a vector of 12 feature distances. More specifically, for the frequency distribution (FD), we use the Kullback-Leibler Divergence (KLD) as the distance metric. For the frequency distributions over 4 gait phases (FD4), for each gait phase, we first align the WiFi-based and video-based features by offsetting their respective average torso speeds. We then use KLD as the distance metric between the two aligned features. The alignment removes the effect of area-dependent average speeds (see Sec. 8) and places more focus on the relative speeds of body parts. For autocorrelation (AC), we use the cosine similarity. For all the other features, we use the Euclidean distance.

5.2 Similarity Prediction

Given a pair of WiFi and video data samples, we compute a set of 12 distances as described previously. We then utilize a simple neural network to combine these distances into a final decision on whether this WiFi-video pair belongs to the same person. We train the network on WiFi/video data and locations disjoint from the test subjects and areas (more details in Sec. 6). During training, these 12 distances are fed into the neural network, which has 1 hidden layer with 30 units, along with a binary label indicating whether these two samples belong to the same person. After training, the network can provide a binary decision on a given pair and a confidence score indicating the similarity between the pair.

6 EXPERIMENTAL SETUP AND DATA COLLECTION

In this section, we describe the experimental setup for collecting data (both wireless and video) and validating our proposed methodology. We then show how we construct the training set for training a small neural network described in Sec. 5.2, and the test set for evaluating our proposed system.

6.1 Experiment Subjects

In order to collect WiFi and video data, we have recruited a total of 18 subjects. We divide them into two disjoint sets

of 10 and 8 subjects, for training and test, respectively. As a result, the test set consists of the walking data of people that have never been seen during training, which allows us to evaluate the proposed system's ability to generalize to new people. In the training set, the 10 subjects (referred to as the training subjects) consist of 9 males and 1 female, with heights ranging from 163 cm to 186 cm. In the test set, the 8 subjects (referred to as the test subjects) consist of 6 males and 2 females, with heights ranging from 160 cm to 186 cm. The speeds of the test subjects have a mean of 1.43 m/s and a standard deviation of 0.26 m/s, while their gait cycles have a mean of 1.06 sec and a standard deviation of 0.17 sec.

6.2 WiFi Data Collection

In this part, we describe the experiments where we use a pair of WiFi transceivers to collect the WiFi data of the subjects.

6.2.1 **Experiment Platform and Data Processing:** For the WiFi data collection process, we use two laptops equipped with Intel 5300 WLAN Network Interface Cards (NICs). We mount $N_{\text{Tx}} = 2$ omni-directional antennas to a tripod of height 85 cm, and connect them to two antenna ports on the Tx laptop, which transmits WiFi packets on WiFi channel 36 with a carrier frequency of 5.18 GHz. Similarly, we mount $N_{\text{Rx}} = 2$ receiving antennas to a tripod of the same height, located 1.5 m away from the Tx antennas, and connect them to two antenna ports on the Rx laptop, which logs the CSI information on 30 subcarriers with a rate of 2,000 packets/sec. The data is then processed offline to extract the CSI information using Csitool [8]. The setup results in a total of $N_{\text{Tx}} \times N_{\text{Rx}} \times 30 = 120$ streams of data which we process using the method in [30]. More specifically, we denoise the data using Principal Component Analysis (PCA). We first generate spectrograms of the first 15 PCA components of the measured signal, using time windows of $T_{\text{win}} = 0.4$ sec, with a shift of 4 ms. We then average these 15 spectrograms to obtain the final spectrogram. The frequency axis ranges from 15 Hz to 125 Hz (which translates to speeds of 0.4 m/s to 3.6 m/s). For the spectrogram segmentation algorithm, we set $V_{\text{th}} = 0.8$ for indoor areas and 0.88 for outdoor areas. These values were determined by using the experimental data of 3 training subjects. We also set the allowable change in average speed to $v_{\text{th}} = 0.3$ m/s.

6.2.2 **Experiment Scenarios:** In the WiFi experiments, we use three different settings for collecting the WiFi CSI data, as described below and shown in Fig. 9:

• **Line-of-Sight Straight-Path (LOS-SP):** In this setting, a WiFi link is deployed in the environment where the person is walking, with a direct view of the person. In each experiment, the subject walks from a starting point that is at least 8 m from the link and towards the link. The subject turns around when he/she is ∼1 m away from the link and then walks back to the starting point. This setting captures how people typically walk in a hallway or a pathway environment. The corresponding areas are shown in Fig. 9 (a) - (d). Areas of Fig. 9 (a) and (b) are only used for training while areas of Fig. 9 (c) and (d) are only used for testing.

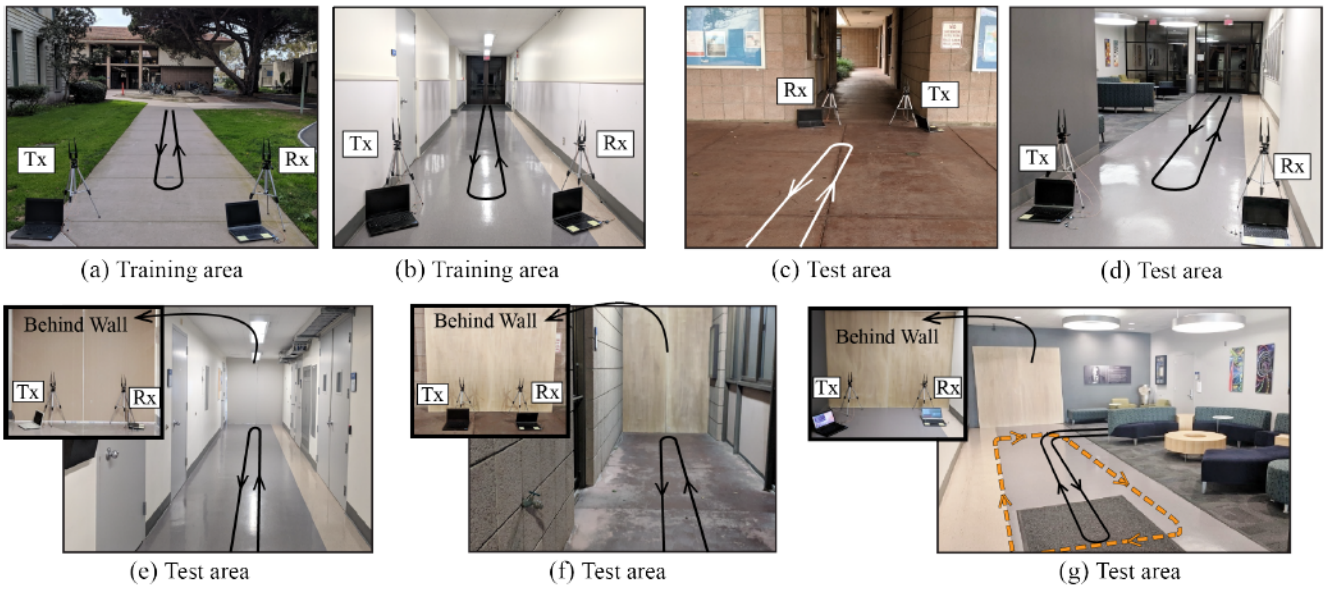
• **Through-Wall Straight-Path (TW-SP):** In this setting, the subjects walk on a path similar to the LOS-SP setting. However, in this case, the WiFi Tx and Rx are placed behind a wall, without any view of the walking subject. We use plywood and drywall for the through-wall experiments, which are used for the walls of ∼90% of residential and small commercial buildings in the U.S. [24], hence, showing the applicability of our proposed system to typical through-wall environments. Our two TW-SP areas are shown in Fig. 9 (e) and Fig. 9 (f). TW-SP areas are only used for testing.

• **Through-Wall Complex-Path (TW-CP):** In this setting, the WiFi Tx and Rx are placed behind a wall. Unlike the previous straight-path settings, the subjects walk on more general and complex paths. As shown in Fig. 9 (g), in the TW-CP experiments, the subjects are asked to walk on two different complex paths that are representative of how people would typically walk in a lounge environment. TW-CP area and complex paths of Fig. 9 (g) are only used for testing.

6.2.3 **Experiment Areas (see Fig. 9):** We use the walking data of the 10 training subjects in 2 LOS-SP areas (Fig. 9 (a) and (b)) for training the neural network. The walking data of the remaining 8 test subjects in the remaining 5 areas, 3 through-wall (2 TW-SP and 1 TW-CP) and 2 line-of-sight, is then used for testing. The training and test areas all vary in size and geometry. In order to create more statistics and avoid biasing the results to a particularly favorable or unfavorable data point, each test subject walks back and forth in each area twice. Each such data instance (one back-and-forth) is then treated independently in the data pool.

6.3 Video Data Collection

In order to train and test XModal-ID, we collect the video data of the 18 subjects walking in an area. For training, we have collected videos of the 10 training subjects walking in one area, shown in Fig. 10 (a). For testing, we have collected videos of the 8 test subjects walking in two different areas shown in Fig. 10 (b). The video data collection areas are completely disjoint from the WiFi experiment locations. In each video area, a subject walks back and forth on a 7-m straight path and a side-view video (with a frame rate of 60 fps) is recorded. The videos are then manually clipped such that each resulting video clip contains a subject walking on a straight path in one direction. Overall, each video clip has an average duration of 4.7 sec. Each video clip is then treated independently in the data pool. We collected a total of 100 such video clips of the training subjects and 96 clips of



(a) Training area (b) Training area (c) Test area (d) Test area

(e) Test area (f) Test area (g) Test area

Figure 9: (a) – (b) WiFi training areas, Line-of-Sight Straight-Path (LOS-SP) setting: we collect WiFi CSI data of the training subjects in these two areas. (c) – (d) WiFi test areas, Line-of-Sight Straight-Path (LOS-SP) setting. (e) – (f) WiFi test areas, Through-Wall Straight-Path (TW-SP) setting. (g) WiFi test area, Through-Wall Complex-Path (TW-CP) setting, with two complex routes indicative of how people generally walk in this lounge area.

the test subjects (by having them repeat the back and forth path a number of times).

We process the frames of each video clip using the algorithm described in Sec. 4.1. The HMR algorithm outputs a total of 2,300 mesh points on the human body for each frame. The number of mesh point sets (frames) is then upsampled to have a frame rate of 250 fps. Based on the surface area values mentioned in [7], we approximate the reflectivity of the torso points to be 3 times the reflectivity of other body parts (which are all taken to have the same reflectivity). For the quasi-specular reflection beam, we set $\sigma_a^2 = 40$ based on the data of 3 training subjects.

To obtain the final video-based features of a walking person, we average the 12 features (described in Sec. 5) over 4 randomly-selected video-based spectrograms of that person (i.e., over 4 video clips of that person). Such averaging is feasible in practice as these 4 spectrograms can be generated from chunks of a longer video or from a few short video clips of the same person. In this paper, the 4 spectrograms amount to a total video duration of 18.8 sec on average.

6.4 Training and Test Sets

Given the collected WiFi and video data, we construct a training set and a test set. For both sets, we first generate the spectrograms for the WiFi data samples and the video clips as described in Sec. 4. After the spectrogram generation, each training or test instance consists of a WiFi data sample and a video data sample (drawn from the corresponding training or test pools), a distance vector between their corresponding features, and a label indicating whether they belong to the same subject. A positive label indicates that the pair belongs to the same person and a negative label denotes otherwise.

The training set is based on the 10 training subjects walking in the 2 WiFi training areas in the LOS-SP setting (Fig. 9 (a) and (b)) and in the 1 video training area (Fig. 10 (a)). The training set consists of a total of 7,280 pairs of WiFi-video instances. As we have a different number of pairs with positive and negative labels, we utilize oversampling [2] to obtain a balanced training set. The neural network discussed in Sec. 5.2 is then implemented in PyTorch [21].

The test set is based on the 8 test subjects' data in the 5 WiFi test areas (Fig. 9 (c) - (g)) and the 2 video test areas (Fig. 10 (b)). The test set includes all the 3 scenarios: LOS-SP (Fig. 9 (c) and (d)), TW-SP (Fig. 9 (e) and (f)), and TW-CP (Fig. 9 (g)) in the WiFi experiments. In the test set, each WiFi sample is paired with several randomly-selected video samples. Overall, we have a total of 2,256 instances (i.e., pairs of WiFi and video data samples) in the test set, with 768 pairs in the LOS-SP setting, 744 pairs in the TW-SP setting, and 744 pairs in the TW-CP setting. In addition to binary classification (i.e., does a WiFi-video pair belong to the same person or not?), we also test the ranking accuracy of our proposed system (see Sec. 7.1). In each ranking test, a WiFi sample serves as a query and 8 video samples serve as the candidates, with one of them containing the same subject as in the WiFi sample. We have a total of 282 such ranking tests in the test set, with 96 in the LOS-SP setting, 93 in the TW-SP setting, and 93 in the TW-CP setting.

7 SYSTEM EVALUATION

In this section, we present extensive experimental evaluations of our proposed system in various practical settings using a large test set. Unlike existing studies on WiFi-based person identification, our test set only contains subjects and areas that have never been seen during the training process.

7.1 Evaluation Criteria

We use the following two evaluation criteria, which are both relevant in different applications:

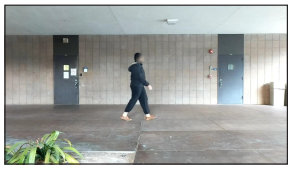
1. Binary classification accuracy: In this setting, we evaluate our proposed system by using pairs of WiFi and video samples. Given a pair of WiFi and video data samples, the system predicts whether they belong to the same person or not. The resulting binary classification accuracy is used as the evaluation metric. As we have different numbers of test instances with positive (same-person) labels and negative (different-people) labels, we report the balanced classification accuracy, i.e., the average of the respective accuracies over the same-person and different-people pairs.

2. Ranking Accuracy: In each ranking test, the system is given a WiFi sample of a test subject and the video samples of several subjects from the test set. Among these candidate video samples, only one of them belongs to the person corresponding to the queried WiFi sample, to which we refer as the correct video sample. The system then ranks the video samples based on their similarity to the WiFi sample. We report the top-1, top-2, and top-3 ranking accuracies in this setting, where the top-k accuracy is defined as the percentage of cases where the correct video sample is ranked among the top k positions of all the video samples in a test.

Remark 3. *Note that if the number of subjects in the ranking test is 2, the system determines which one of the two video samples belongs to the person in the queried WiFi sample. This is different from the binary classification task, which determines whether a video sample and a WiFi sample belong to the same person or not.*

7.2 Performance Evaluation

In this section, we evaluate our proposed system on our extensive test set, which only has experimental areas and subjects that are not seen during the training phase. We further extensively test our system in through-wall scenarios and with complex paths. See Sec. 6.4 for the details of the test set. It is noteworthy to re-emphasize that our system does not need to know the track of the subject, or the details of the test area, as we discussed in Sec. 4.3. Furthermore, all the test videos are from areas of Fig. 10 (b) (disjoint from the WiFi areas), as discussed earlier. Table. 1 summarizes all the results that we shall discuss in this section.



(a) Video training area

(b) Video test areas

Figure 10: Sample snapshots for videos in (a) the training video location, and (b) the two test video locations.

Area	Binary class. accuracy	Ranking accuracy		
		Top-1	Top-2	Top-3
Line-of-Sight Straight-Path setting				
Area of Fig. 9 (c)	90%	87%	96%	98%
Area of Fig. 9 (d)	86%	70%	83%	95%
Average	88%	78%	90%	96%
Through-Wall Straight-Path setting				
Area of Fig. 9 (e)	83%	74%	90%	97%
Area of Fig. 9 (f)	89%	82%	96%	100%
Average	86%	78%	93%	98%
Through-Wall Complex-Path setting				
Area of Fig. 9 (g)	82%	69%	86%	96%
Overall average	85%	75%	90%	97%

Table 1: The binary classification accuracy and top-1 to top-3 ranking accuracies of XModal-ID on the test set, in three different settings. The last row shows the average performance over all the areas/settings.

7.2.1 **Evaluation of Line-of-Sight Scenarios:** We first evaluate XModal-ID in the Line-of-Sight Straight-Path (LOS-SP) setting, consisting of 2 WiFi areas (Fig. 9 (c) and (d)). In this case, XModal-ID achieves a binary classification accuracy of 90% in the area of Fig. 9 (c) and 86% in the area of Fig. 9 (d), resulting in an overall average binary classification accuracy of 88%. In other words, given a pair of WiFi and video samples, both generated from subjects and environments not seen during training, our system has an 88% chance of correctly predicting whether these two samples correspond to the same person or not, in these two areas.

Next, we look at the ranking performance. In the LOS-SP setting, given a queried WiFi sample and 8 candidate video samples of the 8 test subjects, XModal-ID has a success rate of 78% of assigning the highest rank to the correct video sample, in these two areas. Note that a random selection would only result in a success rate of 12.5%. Moreover, in this setting, XModal-ID has top-2 and top-3 accuracies of 90% and 96%, respectively. The ranking accuracy per area is shown in Table 1.

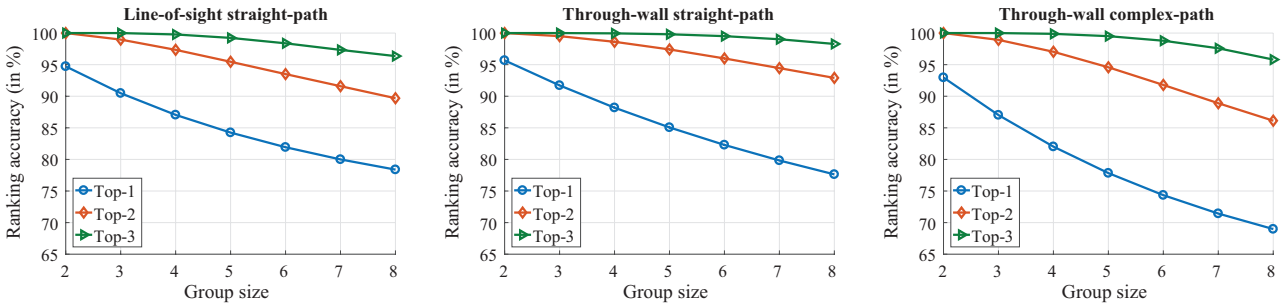


Figure 11: Top-1 to top-3 ranking accuracies when group size varies from 2 to 8, in (a) Line-of-Sight Straight-Path (LOS-SP) areas, (b) Through-Wall Straight-Path (TW-SP) areas, and (c) Through-Wall Complex-Path (TW-CP) area.

7.2.2 **Evaluation of Through-Wall Scenarios:** Next, we consider the Through-Wall Straight-Path (TW-SP) WiFi areas (Fig. 9 (e) and (f)), where the WiFi link is placed behind a wall and does not have any view of the subjects. XModal-ID achieves a binary classification accuracy of 83% in the area of Fig. 9 (e) and 89% in the area of Fig. 9 (f), amounting to an overall average accuracy of 86%. In terms of ranking, XModal-ID achieves top-1, top-2, and top-3 accuracies of 78%, 93%, and 98%, over both areas. In particular, when XModal-ID is deployed in the area of Fig. 9 (f), it includes the correct video sample among the top 3 all the time.

In the Through-Wall Complex-Path (TW-CP) area, shown in Fig. 9 (g), the WiFi link is placed behind a wall and each test subject walks on two sample complex paths (each path treated as a separate experiment). These two paths represent how people would typically walk in this lounge area. In this setting, XModal-ID achieves a binary classification accuracy of 82%. For the case of ranking, our system obtains top-1, top-2, and top-3 accuracies of 69%, 86%, and 96%, respectively, in this area. It is noteworthy that in this TW-CP setting, which showcases challenging real-world application scenarios, the system has a very high probability (0.96) of including the correct video sample within the top 3 ranks.

Overall, XModal-ID achieves a binary classification accuracy of 85%, and top-1, top-2, and top-3 ranking accuracies of 75%, 90%, and 97%, over all 5 areas/scenarios. These results demonstrate that XModal-ID has a robust performance, even when the transceivers are placed behind a wall, without any prior knowledge or view of the person/area, and when the subjects walk on unknown and complex paths.

7.3 Evaluation with Different Group Sizes

In the previous part, we showed the performance of our proposed system on the full test set consisting of 8 subjects. While the binary classification accuracy is independent of the number of subjects, ranking accuracy is a function of the number of subjects. In this section, we then study the performance of XModal-ID by varying the number of subjects in the test set, to which we refer as the group size.

Fig. 11 (a) shows the top-1, top-2, and top-3 ranking accuracies when the group size is varied from 2 to 8, in the LOS-SP setting. For each group size that is smaller than 8, the accuracies are averaged over all the possible subsets of subjects for that group size. As can be seen, as we reduce the group size, the ranking accuracies increase, since, with a smaller group size, it is less likely to have two subjects with similar gaits. When the group size is less than 8, the top-1 ranking accuracy is always greater than 80%.

Fig. 11 (b) and (c) further show the ranking accuracies in the through-wall straight-path and complex-path settings, respectively, as a function of the group size. As can be seen, the accuracies increase as the group size decreases. Notably, when the group size is less than 8, the top-3 accuracy in these two through-wall settings is very close to 100%.

Overall, these evaluation results show that XModal-ID can successfully perform cross-modal person identification even when the test subjects and areas have never been seen before. The test set areas represent a wide variety of real-life scenarios, including through-wall scenarios and cases where the person walks on a complex path (rather than a straight one). Our system does not even need to know the track of the subjects. Overall, our results demonstrate the efficacy of XModal-ID in various real-world scenarios.

8 DISCUSSION

In this section, we discuss a few key aspects of XModal-ID, as well as its limitations and future extensions.

Environment-Dependent Average Speeds: Environmental factors can sometimes affect people's average walking speed [6]. For instance, we noticed that people tend to walk slightly faster in outdoor/open areas, as compared to indoor/closed areas. All existing works on WiFi-based gait identification train and test in the same area, where the subjects mostly maintain the same walking speeds. On the other hand, in XModal-ID, in addition to the overall average speed, we also utilize spectrogram features that are independent of the average speed and only depend on the distribution of

the relative speeds of body parts (see Sec. 5). Hence, XModal-ID can tolerate small changes in the average speeds of the subjects.

Tracks with Varying ψ: XModal-ID does not assume any knowledge of the track of the person. Instead, it uses the spectrogram segmentation algorithm in Sec. 4.3 to extract the part of the person's track where ψ is approximately constant. The constant-ψ parts correspond to parts of track where the subject walks on a straight path towards/away from the midpoint of the Tx-Rx line, for the case where Tx and Rx are close enough to each other (see Sec. 4.3). Since this is a very general condition, most natural tracks will at least have small parts that would satisfy this condition. In fact, XModal-ID only needs a very small part of the track, e.g., 3 sec, to satisfy this condition, as discussed earlier. In the rare case that no part of the track satisfies this condition, the varying ψ can be estimated by existing WiFi-based tracking approaches and XModal-ID can be extended to accommodate the varying ψ.

Applicability to Intruder Detection: XModal-ID can also determine whether a WiFi sample belongs to a new user whose video is not available. It can compare this WiFi sample with each of the available video samples, using the binary classification criterion, and declare an unseen user if the WiFi sample does not match any of the videos. This setting can be relevant in applications such as intruder detection.

Processing Time: A typical duration of a WiFi data sample in our experiments is 25 sec. On a 3.40 GHz Intel Core i7 PC, XModal-ID takes an average of ~19.8 sec to fully process such WiFi data. For videos, XModal-ID takes ~132.5 sec to fully process a video clip of 4.7 sec (average duration) in order to generate a final feature vector. In particular, ~112.8 sec are dedicated to generating the human mesh model, using the publicly available codes of [11, 13] on an NVidia GTX 1070 GPU, while the remaining steps (e.g., WiFi signal simulation) take ~19.7 sec on a 3.40 GHz Intel Core i7 PC.

Limitations and Future Extensions: 1) *Number of People*: Currently, XModal-ID can determine if a pair of WiFi and video samples belong to the same person or not. In addition, it can reliably identify a person from 8 video footage candidates, which enables several real-world applications, such as suspect search and smart-home personalized services, where the number of subjects is typically less than 8. As part of future work, one can scale up XModal-ID to a larger number of people, which can enable other useful applications.

2) *Multi-Person Identification*: XModal-ID assumes that there is only one person walking in the WiFi area. As discussed in Sec. 1, the current system can support several applications (e.g., personalized service provisioning), where there is typically a single user in the WiFi area. When there are multiple users, the spectrogram would contain the impact of all users' motions, thus making it challenging to identify each individual. As part of future work, one can isolate the impact of each person for the purpose of identification.

3) *Stationary People*: XModal-ID identifies people based on their gait. Thus, it requires that the person walks (even briefly) to be identified. If a person remains completely stationary, XModal-ID would not be able to identify him/her. Extensions to include other features more relevant to stationary people/actions is a possible future direction.

4) *Reflection-Based Video-to-Wifi Modeling:* We utilize Born approximation and quasi-specular reflections to model the wireless signals in video-to-WiFi modeling, as discussed in Sec. 4.1. This model is not valid when the person is crossing the link. However, XModal-ID can still robustly work if a person crosses the link occasionally, since the segmentation algorithm will not choose such segments of the spectrogram. However, if a person is mainly blocking the link, or generally, has a motion pattern that does not have any constant-ψ segment, then XModal-ID needs to be extended, as discussed.

9 CONCLUSIONS

In this paper, we proposed XModal-ID, a WiFi-video cross-modal person identification system, which can determine if an unknown person walking in a WiFi-covered area is the same as the person in a video footage. To achieve this, XModal-ID utilizes WiFi CSI magnitude measurements of a pair of WiFi transceivers to identify a person, by matching the gait features captured by the WiFi measurements to those from a video of a walking person. XModal-ID does not need any prior wireless or video data of the person to be identified, or the identification area. It can further identify people through walls and does not need the knowledge of the track of the person. In order to evaluate our proposed system, we constructed a large test set with 8 subjects, 5 WiFi areas, and 2 video areas, all of which were unseen in the training phase. Furthermore, the test set includes 3 areas where the transceivers were placed behind a wall, as well as scenarios with complex paths. XModal-ID achieves an overall binary classification accuracy of 85% in predicting whether a WiFi-video pair belong to the same person or not, and top-1, top-2, and top-3 ranking accuracy of 75%, 90%, and 97%, respectively. This demonstrates that our proposed XModal-ID system can robustly identify unknown people in new environments and through walls.

ACKNOWLEDGMENTS

The authors would like to thank all the participants in our experiments. The authors would also like to thank the shepherd Dr. Swarun Kumar and the anonymous reviewers for their valuable comments and helpful suggestions. This work is funded in part by NSF CCSS award # 1611254 and in part by NSF NeTS award # 1816931.

REFERENCES

[1] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand. 2015. Capturing the human figure through a wall. *ACM Transactions on Graphics* 34, 6 (2015), 219.

[2] N. V. Chawla. 2009. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*. Springer, 875–886.

[3] W. C. Chew. 1995. *Waves and fields in inhomogeneous media*. IEEE press.

[4] R. T. Collins, R. Gross, and J. Shi. 2002. Silhouette-based human identification from body shape and gait. In *Proceedings of IEEE International Conference on Automatic Face Gesture Recognition*.

[5] P. Connor and A. Ross. 2018. Biometric recognition by gait: A survey of modalities and features. *Computer Vision and Image Understanding* 167 (2018), 1–27.

[6] M. Franěk. 2013. Environmental factors influencing pedestrian walking speed. *Perceptual and Motor Skills* 116, 3 (2013), 992–1019.

[7] J. L. Geisheimer, E. F. Greneker, and W. S. Marshall. 2002. High-resolution Doppler model of the human gait. In *Radar Sensor Technology and Data Visualization*.

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

[9] J. Han and B. Bhanu. 2006. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 2 (2006), 316–322.

[10] R. Hartley and A. Zisserman. 2003. *Multiple view geometry in computer vision*. Cambridge university press.

[11] K. He, G. Gkioxari, P. Dollar, and R. Girshick. 2018. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[12] C. Hornsteiner and J. Detlefsen. 2008. Characterisation of human gait using a continuous-wave radar at 24 GHz. *Advances in Radio Science* 6, B.2 (2008), 67–70.

[13] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[14] C. R. Karanam, B. Korany, and Y. Mostofi. 2018. Magnitude-based angle-of-arrival estimation, localization, and target tracking. In *Proceedings of the ACM/IEEE International Conference on Information Processing in Sensor Networks*.

[15] S. Katz, A. Tal, and R. Basri. 2007. Direct visibility of point sets. In *ACM Transactions on Graphics*, Vol. 26. ACM, 24.

[16] L. Lee and W. E. L. Grimson. 2002. Gait analysis for recognition and classification. In *Proceedings of IEEE International Conference on Automatic Face Gesture Recognition*.

[17] Y. Li and T. Zhu. 2016. Using Wi-Fi signals to characterize human gait for identification and activity monitoring. In *Proceedings of the IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*.

[18] J. Lv, W. Yang, D. Man, X. Du, M. Yu, and M. Guizani. 2017. Wii: Device-free passive identity identification via WiFi signals. In *Proceedings of the IEEE Global Communications Conference*.

[19] A. Mehmood, J. M. Sabatier, M. Bradley, and A. Ekimov. 2010. Extraction of the velocity of walking human's body segments using ultrasonic Doppler. *The Journal of the Acoustical Society of America* 128, 5 (2010), EL316–EL322.

[20] I. Orović, S. Stanković, and M. Amin. 2011. A new approach for classification of human gait based on time-frequency feature representations. *Signal Processing* 91, 6 (2011), 1448–1456.

[21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[22] A. Seifert, M. Amin, and A. M. Zoubir. 2019. Toward unobtrusive in-home gait analysis based on radar micro-Doppler signatures. *IEEE Transactions on Biomedical Engineering* (2019).

[23] C. Shi, J. Liu, H. Liu, and Y. Chen. 2017. Smart user authentication through actuation of daily activities leveraging WiFi-enabled IoT. In *Proceedings of the ACM International Symposium on Mobile Ad Hoc Networking and Computing*.

[24] A. Sinha, R. Gupta, and A. Kutnar. 2013. Sustainable Development and Green Buildings. *Drvna Industrija* 64, 1 (2013), 45–53.

[25] A. Świtoński, A. Polański, and K. Wojciechowski. 2011. Human identification based on gait paths. In *Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems*.

[26] D. Tahmoush and J. Silvious. 2009. Stride rate in radar micro-Doppler images. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*.

[27] B. Vandersmissen, N. Knudde, A. Jalalvand, I. Couckuyt, A. Bourdoux, W. De Neve, and T. Dhaene. 2018. Indoor person identification using a low-power FMCW radar. *IEEE Transactions on Geoscience and Remote Sensing* 56, 7 (2018), 3941–3952.

[28] D. K. Wagg and M. S. Nixon. 2004. On automated model-based extraction and analysis of gait. In *Proceedings of the IEEE international conference on Automatic Face and Gesture Recognition*.

[29] J. Wang, Y. Zhao, X. Fan, Q. Gao, X. Ma, and H. Wang. 2018. Device-Free Identification Using Intrinsic CSI Features. *IEEE Transactions on Vehicular Technology* 67, 9 (2018), 8571–8581.

[30] W. Wang, A. X. Liu, and M. Shahzad. 2016. Gait recognition using WiFi signals. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*.

[31] T. Xin, B. Guo, Z. Wang, M. Li, Z. Yu, and X. Zhou. 2016. Freesense: Indoor human identification with Wi-Fi signals. In *Proceedings of the IEEE Global Communications Conference*.

[32] Y. Zeng, P. H. Pathak, and P. Mohapatra. 2016. WiWho: WiFi-based person identification in smart spaces. In *Proceedings of the International Conference on Information Processing in Sensor Networks*.

[33] J. Zhang, B. Wei, W. Hu, and S. S. Kanhere. 2016. WiFi-ID: Human identification using WiFi signal. In *Proceedings of the International Conference on Distributed Computing in Sensor Systems*.

[34] R. Zheng, Y. Zhao, and B. Chen. 2017. Device-Free and Robust User Identification in Smart Environment Using WiFi Signal. In *Proceedings of the IEEE International Symposium on Parallel and Distributed Processing with Applications and the IEEE International Conference on Ubiquitous Computing and Communications*.

[35] H. Zou, Y. Zhou, J. Yang, W. Gu, L. Xie, and C. J. Spanos. 2018. WiFi-based human identification via convex tensor shapelet learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.