Vi-Fi: Associating Moving Subjects across Vision and Wireless Sensors

Hansi Liu
Rutgers University
hansiiii@winlab.rutgers.edu

Abrar Alali
Old Dominion University
aalal003@odu.edu
Saudi Electronic University
a.alali@seu.edu.sa

Mohamed Ibrahim
Carnegie Mellon University
miahmed@andrew.cmu.edu

Bryan Bo Cao
Stony Brook University
boccao@cs.stonybrook.edu

Nicholas Meegan
Rutgers University
njm146@scarletmail.rutgers.edu

Hongyu Li
Rutgers University
hongyuli@winlab.rutgers.edu

Marco Gruteser
Rutgers University/Google Research
gruteser@winlab.rutgers.edu

Shubham Jain
Stony Brook University
jain@stonybrook.edu

Kristin Dana
Rutgers University
kristin.dana@rutgers.edu

Ashwin Ashok
Georgia State University
aashok@gsu.edu

Bin Cheng
InfoTech Labs, Toyota Motor North America R&D
bin.cheng@toyota.com

Hongsheng Lu
InfoTech Labs, Toyota Motor North America R&D
hongsheng.lu@toyota.com

ABSTRACT

In this paper, we present Vi-Fi, a multi-modal system that leverages a user's smartphone WiFi Fine Timing Measurements (FTM) and inertial measurement unit (IMU) sensor data to associate the user detected on a camera footage with their corresponding smartphone identifier (e.g. WiFi MAC address). Our approach uses a recurrent multi-modal deep neural network that exploits FTM and IMU measurements along with distance between user and camera (depth information) to learn affinity matrices. As a baseline method for comparison, we also present a traditional non deep learning approach that uses bipartite graph matching. To facilitate evaluation, we collected a multi-modal dataset that comprises camera videos with depth information (RGB-D), WiFi FTM and IMU measurements for multiple participants at diverse real-world settings. Using association accuracy as the key metric for evaluating the fidelity of Vi-Fi in associating human users on camera feed with their phone IDs, we show that Vi-Fi achieves between 81% (real-time) to 91% (offline) association accuracy.

1 INTRODUCTION

Association of cross-domain sensor data is a fundamental need in applications and systems that exploit multi-modal sensor data. With the pervasive use of cameras and wireless devices, one key instance of this problem is the association between objects or persons detected in camera video and wireless data originating from their transmitters, as depicted in Figure 1. Successful association enables use-cases such as localization by fusing depth camera measurements with wireless ranging. It can also improve identification and re-identification when objects reappear in camera view, and tracking since transmitters, with user consent, can send a stable identifier. Moreover, it can provide a means to send messages or notifications to devices observed on camera if part of the shared identifier is a communication address. Such capabilities open new opportunities such as systems that improve the performance of respiratory disease contact tracing or exposure notifications in schools, offices or warehouses. Current Bluetooth-based systems provide coarse distance estimates and cannot reliably determine whether two persons occupied the same room. Stationary cameras that can associate people with their phones could provide more precise distance measurements and the necessary room disambiguation to fill this gap towards more accurate exposure notifications. Other potential use cases include traffic safety applications via vehicle-to-pedestrian (V2P) communication, for example detecting and notifying pedestrians in dangerous situations using road-side or vehicle mounted cameras. Note that we focus here on cases where users provide explicit consent and opt-in, rather than background tracking without the phones cooperation.

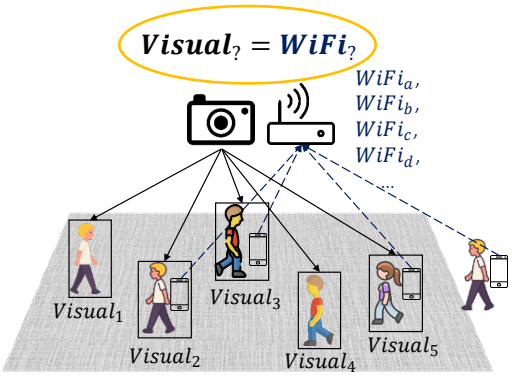


Figure 1: Motivation: Successfully associating vision-wireless information is fundamental to sensor fusion systems. The goal is to associate visually detected participants with corresponding phone identifiers.

Prior Work. Multiple recent projects have addressed data association by extracting features from multi-modal data including image, WiFi signal or smartphone IMU measurements and consider feature similarity as a key of association. Existing approaches to address this association problem tend to rely on color information [27, 56], incur accuracy tradeoffs [16, 37], or cannot provide real-time association due to post processing needs over longer sequences [22, 27, 31, 56, 59]. The reliance on color information creates challenges under varied lighting conditions or in scenarios where workers wear standardized clothing. Moreover, a large group of works [16, 22, 32, 38, 56] assumes users are always visible in the camera view and/or no passerby users exists during the experiment, which significantly reduced the complexity of the association problem.

Approach. We propose *Vi-Fi*, a multi-modal approach that associates visually detected persons, represented through the bounding boxes generated by an object detector, with a smartphone identifier (MAC addresses, for example), by fusing information from vision and wireless domains. A key insight of this approach is that both cameras and wireless receivers are now becoming capable of improved ranging—cameras through RGB-D technology and WiFi through Fine Timing Measurement (FTM) [1, 15]. Since cameras are also increasingly equipped with WiFi transceivers [2–4] this presents the opportunity to generate distance measurements between the camera and the detected persons in both the visual and wireless domain, generating a common reference measurement to facilitate cross-domain fusion. Yet, it remains challenging to match these measurements due to measurement noise and ambiguity. Lighting condition changes and occlusions add noise and uncertainty to camera-based depth measurements. Multi-path fading and shadowing degrades the accuracy of WiFi FTM measurements.

To address the above challenges and achieve the association goal, we explore supervised (data-driven) and model-driven methods that leverage information from WiFi FTM measurements and smartphone inertial measurement unit (IMU) motion sensor data to match each detected participant in the camera view with their smartphone ID. Specifically, we introduce a supervised multi-modal affinity matrix deep learning technique that learns a similarity metric and predicts an affinity matrix for multiple camera-phone pairs. We compare this with a baseline model-driven approach involving a bipartite trajectory matching algorithm that exploits similarities between trajectories recovered from IMU sensor data and trajectories recovered from video. To evaluate *Vi-Fi*'s performance, we collect a large multi-modal dataset of real-world outdoor scenes with multiple participants per scene. The participants have phones that send FTM messages. Numerous passerby pedestrians are in the camera view. The dataset comprises RGB-D video from a mounted camera and smartphone data from participants including WiFi FTM and IMU measurements.

Summary of Contributions. As a summary, *Vi-Fi* makes the following contributions:

- Exploring the design space of associating moving subjects across vision and wireless sensors using multi-modal sensors data including depth, WiFi FTM from a single access point (AP), and IMU measurements from users' smartphones without the need to rely on color and appearance information.
- Designing a new multi-modal deep neural network architecture that learns embedding similarities of the multi-modal sensors data to predict affinity matrices that associate subjects across the camera and wireless domain. The network is capable of handling complex real-world scenarios where multiple passerby pedestrians exist and signal duration is limited.
- Exploring a model-driven bipartite matching algorithm that matches trajectories estimated from camera views with those estimated from wireless phone sensors measurements.
- Presenting a large-scale multi-modal dataset and using this dataset to evaluate and compare the aforementioned approaches. The dataset consists of a total 90 RGB-D video sequences (each of 3 minutes captured at 10 frames/sec, leading to a sample set of about 162,000 images) recorded at 6 different scenarios with different camera perspectives , multiple uncontrolled experimental users and passerby pedestrians. It also contains WiFi FTM measurements, smartphone IMU sensor data and GPS measurements, captured at 3Hz, 50Hz and 1 Hz, respectively.

Artifact Availability: We have made our dataset and methods implementation public [5, 6]. In the dataset we label each experiment subject with a unique identifier to to ensure anonymity. Moreover, we adopt deface [7] to blur the faces of user participants and passerby pedestrians to protect their privacy and biometric information.

2 BACKGROUND AND RELATED WORK

WiFi Fine Timing Measurement (FTM). IEEE 802.11-2016 Standard [15] has included the Fine Timing Measurement (FTM) protocol (802.11REVmc) to perform wireless ranging by measuring the round trip time between an access point and a WiFi station. The protocol subtracts processing times from the round trip time, converts it into a one-way time-of-flight estimate, and uses this to compute an estimated range using typical propagation speed. To achieve higher ranging accuracy, it conducts multiple message exchanges and compute average on the estimated ranges. [29] confirms that the FTM protocol can achieve meter-level accuracy in open space environments but degrades in high multipath environments.

Camera-Wireless Association. Several surveys [51] [23] have summarised the human-sensing methods for localization and identification. Prior works [28, 37, 45, 48, 58, 61, 62] consider the association problem as a sub-problem of person localization with various focuses. XModal-ID [33] identifies a person based on gait features extracted from WiFi CSI and video footage without training, but only in limited WiFi areas where one person is walking. RGB-W [16] adopts a minimum weight bipartite matching algorithm to match visually detected participants with the corresponding WiFi MAC addresses from multiple WiFi APs using spatial information, and achieves 64.0% to 28.6% matching accuracy. Eye-Fi [22] leverages WiFi CSI to estimate Angle of Arrival (AoA) using deep learning to associate vision and AoA based trajectories with 75% accuracy using weighted Euclidean distance. [24, 47] take advantage from both modalities for localization. However, these methods require multiple WiFi APs deployed in the scene, multiple IMU devices attached to a single user, or calibrated environments. Alongside of these works, other sensing modalities have also been explored, including GPS [39], Bluetooth [30], RF [41, 42, 60] and audio signal [36, 40].

Camera-IMU Association. Using inertial sensor readings to estimate motion trajectories is extensively studied. Prior work has used this concept to associate the motion status and direction of motion with a camera image/video feed [52, 54]. Recent methods [19, 26, 59] associate visual tracker trajectories with smartphone IMU measurements. Other works [20, 26, 27, 35, 43, 53, 57] exploit extracted features from motion data to associate with vision data. Multi-modal association can benefit from learning a joint latent space that represents semantic similarity on large datasets [17, 25] such as using camera, smart phone and WiFi data. [37] associates silhouette images and accelerometer data by exploiting deep learning features.

Multi-modal Data Extraction. Existing approaches to address this association problem have a dependency on color information [27, 56], incur accuracy tradeoffs [16, 37], or cannot function in real-time due to dependency on long sequences [59] and thus long post-processing times. The reliance on color information creates challenges under varied lighting conditions or in scenarios where workers wear standardized clothing. One exception is Eye-Fi [22] which performs Angle-of-Arrival based matching using WiFi CSI measurements. Despite many years of research, interfaces for obtaining CSI information are still not widely available, which leads to practical deployment hurdles.

In summary of related works, previous methods either require heavy infrastructure, human activity recognition as an intermediate step [47], or feature representations based on image pixel color that are not robust to lighting or cloth changes in constrained environments. Some works [16, 22, 56] assume the experiment users are always visible in the camera field of view or choose to aggregate individual user trajectories to simulate multi-user scenarios [32, 38]. Another group of works target the association problem as an offline optimization problem (post-processing). The disadvantage of this is that it requires knowledge of significant history (sometimes the entire set) of data measurements.[22, 27, 31, 56, 59] Additionally, most of the existing work assume that all detected pedestrians in the camera view have their corresponding IMU devices, while in reality some passerby users may not be carrying their device or have a device with an unknown ID. In comparison, our dataset and results demonstrate the ability of our proposed deep affinity multi-modal network to overcome these limitations in **diverse scenarios, paving the path to complex, real world deployment**. A summary of the most representative works that focus on vision-wireless association is listed in Table 1.

3 VI-FI OVERVIEW

Figure 2 presents an overview of *Vi-Fi*. Vi-Fi uses a deep learning approach to associate identities across multiple modalities - vision and wireless devices. The system process the vision (camera) feed and detects pedestrians (users). Within wireless devices, we leverage the depth information provided by WiFi FTM and the motion profile from IMU. The users detected on the camera feed are *associated* with their respective smartphone IDs using IMU and WiFi FTM data obtained from their smartphones. A RGB-D camera detects and computes each participant's depth within its field-of-view for each frame. Meanwhile, each phone exchanges WiFi FTM messages with the AP while also gathering IMU data including accelerometer, gyroscope, magnetometer readings.

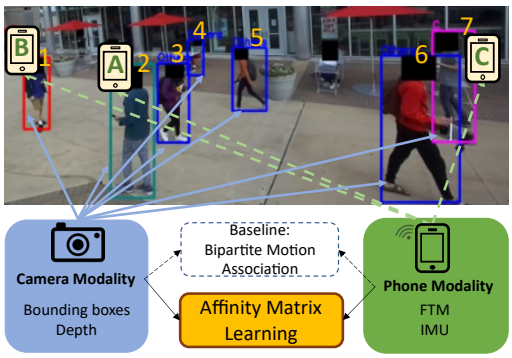


Figure 2: *Vi-Fi* Overview. Our approach collects participants' multi-modal data and associates vision and wireless measurements of the same identity.

For privacy reasons, Vi-Fi is designed to only track and associate consenting users. We consider that phones of consenting users will actively share their measurements with a Vi-Fi server agent that determines the associations using our proposed **affinity matrix deep learning** approach. This approach leverages discriminative embedding features from the multi-modal sensors data to match a visually detected person to the correct smartphone identity. We propose a network architecture that extracts features from camera modality (bounding boxes, depth) and smartphone modality (FTM, IMU). The network further utilizes the learned features to compute affinity matrices where association probabilities are encoded for every camera-phone pair.

Considering that deep learning approaches are highly data driven, we also explore a model-driven approach that does not require any training data. In this approach, referred to as **bipartite motion association**, we compute motion profiles from data gathered from the camera and users' smartphones. Motion profiles comprise trajectory, heading estimates, and distance from a reference point. We compute similarity indices between corresponding pairs of motion features. The similarity metrics are combined in a bipartite graph to determine camera-phone associations based on minimum weight matching. At each time, the algorithm only uses the historical data from that session to compute similarity indices. It leverages users' movements across heterogeneous modalities to create spatio-temporal motion profiles. We note that the bipartite motion approach is based on well-known methods of motion profile and trajectory mapping techniques. Unlike the affinity matrix approach, we do not claim novelty, however, explore this approach merely to study the performance of a standard baseline technique that does not leverage deep learning.

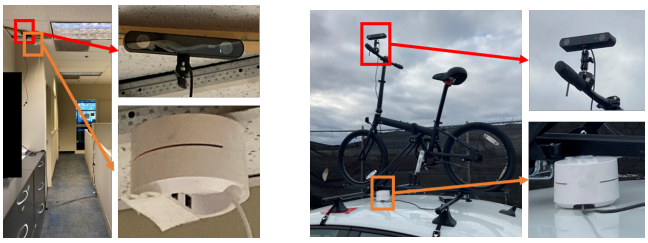
4 MULTI-MODAL DATASET

Data Collection. To investigate the vision-phone association, we collect 90 sequences of multi-modal data through experiments[1]. We divide this dataset into two categories: **Dataset A** constitutes data from experiments conducted in one controlled indoor office space environment involving 5 legitimate users and no passerby; **Dataset**

[1]Experiments with human users were conducted following strict COVID-19 protocols and IRB stipulations. All participating users wear masks and are more than 6 ft apart during data collection.

Work reference	Vision data	Wireless data	Scene (Num. of scenes)	Max. num. of users	Max. num. of passersby	Out-of-view users	User moving pattern	Min. signal duration to make association	Association accuracy	Open source
[16]	Silhouette	RSSI	Indoor outdoor (2)	12	0	No	Unconstrained walking	-	28.6% ∼ 64%	-
[22]	Panoramic RGB	CSI	indoor (2)	10	0	No	Unconstrained walking	≤ 25 s	avg: 75%	-
[56]	RGB	Accelerometer Gyroscope Compass	indoor outdoor	12	0	No	Unconstrained walking	10 s	70% ∼ 90%	-
[38]	Silhouette	Accelerometer	Indoor	10 (aggregated)	0	No	11 activities	3 s	76.30%	data: [8] (modified) code: [9]
[26]	RGB	Accelerometer	Indoor outdoor (2)	8	0	yes	unconstrained walking, running	30 s (avg.)	93.60%	data: [10]
[32]	RGB	Accelerometer	Indoor (1)	12 (aggregated)	0	No	6 activities	60 s	65% ∼ 92%	-
Ours	**Depth**	**Accelerometer Gyroscope Magnetometer FTM**	**Indoor, outdoor (6)**	**5**	**10**	**Yes**	**Unconstrained walking**	**3 s**	**81%**	**data: [5] code: [6]**

Table 1: Summary of the most representative works that focus on associating vision data and wireless data.



(a) Dataset A setup (b) Dataset B setup

Figure 3: Data collection setup. In order to have a proper field of view, the camera is mounted at the height of 2.4 - 2.8 m. For Dataset A, the camera is mounted to the ceiling of the facility that has a height of 2.8 m. A Google Nest WiFi AP is placed next to the camera. For Dataset B, the height of the camera in each scenario is 2.6 m. The AP is placed beneath the camera at the height of 1.4 m. For the purpose of sufficient height and outdoor stability/mobility, we mount the ZED2 camera on the handle of a roof-mounted bike.

B constitutes data from experiments conducted in 5 different real-world outdoor environments with 2-3 user participants and rest of the pedestrians in view are passersby (up to 12 in our dataset). Each video sequence lasts 3 minutes and contains RGB-D frames (captured at 10 frames/sec), FTM, and 9-axis IMU sensor data (accelerometer, gyroscope, and magnetometer) of up a Google Pixel 3a smartphone device. Each of the legitimate users (3 for Dataset A and 5 for Dataset B) carried the Pixel phone. The users were not restricted in how they carried the phones (in hand or in their pocket). Our dataset is representative of a diverse set of scenarios with participants holding smartphone devices exchanging FTM messages with the AP and recording IMU measurements, as well as a varying number of passerby whose phone devices did not opt in to the FTM and IMU recording. The RGB-D camera captures the scene and all the pedestrians under its field of view (FOV). The maximum number of detected pedestrians (phone holders and passerby) at a time is 12. A participant with a phone might exit and re-enter the camera's field of view due to unconstrained walking pattern and limited field of view of the camera. As a result, the number of pedestrians detected in vision modality could be less than, equal to, or greater than the number of participants' phones heard over the wireless channel. This change of cardinality in both modalities poses another challenge to the cross modal association. Figure 4 shows an example of continuous sampled frames from the labeled dataset.

Collection Setup. The setup (Figure 3) consists of a mounted Stereolabs ZED2 [11] (RGB-D) camera set at the height of 2.4 - 2.8 meters with a proper field of view to record video at 10fps, which collects depth information from 0.2m to 20m away from the camera. The smartphones are set to exchange FTM messages at 3 Hz frequency with a Google Nest WiFi Access Point anchored besides the camera. Each smart phone also logs its IMU sensor data at 50 Hz and GPS readings at 1 Hz (in Dataset B only). The smartphones and camera are connected to the Internet to achieve coarse synchronization using network time synchronization.

Ground-Truth Labeling. To mark ground truth for evaluating association accuracy, we manually annotate the participants in the (dataset) video frames with bounding boxes. We use a tracking module from ZED SDK's API to help with annotating the pedestrians in the video scene and perform 2 total passes over the data. During the first pass, the ZED tracker outputs a track ID for each tracked person and a bounding box for that person at each frame, where the ground truth bounding box labels are manually matched with these track IDs every 10 frames. We perform a second pass in which each frame's ground truth label for the pedestrians is manually reviewed and corrected where necessary using the Visual Object Tagging Tool (VoTT) [12]. The two-step method saves time over manually drawing bounding boxes and labeling each pedestrian individually.

Figure 4: Showcasing sampled screenshots from our multi-modal dataset. The dataset contains videos recording one indoor scenario and 5 different outdoor scenarios where multiple participants randomly walk around the venue with a varying number of passersby. Each participant, denoted by a bounding box with an alphabetical character, is holding a smart phone that is exchanging FTM messages with an access point (located close to the RGB-D camera) while logging its IMU sensor data. Other passersby are depicted with a bounding box labeled as "Others". We collect a total of 94 3-minute video sequences across the 6 scenes. (Best viewed zoomed)

5 BASELINE: BIPARTITE MOTION ASSOCIATION

We develop a model-driven approach that computes association based on pairs of similar information from multiple modalities. Our goal is to compute features from each modality to capture the complementary characteristics as well as redundancy between the data streams. Bipartite matching using Linear assignment or Hungarian algorithm [13, 49] is an intuitive yet versatile approach to find associations between two disjoint sets. It has been widely adopted and evaluated in systems that require data association such as [34, 36]. It is also a popular choice among the related works that focus on vision-wireless association [16] To this end, we construct vision-based and wireless-based motion profiles for each entity in the scene and associate them under the paradigm of bipartite matching. Figure 5 shows an overview of bipartite motion association that is explained in the following:

Motion Profile Construction. Motion profiles, computed for the phone's IMU and camera (vision) data, capture the relative movement and positioning of entities in the scene. A motion profile, p, is given by: $p = \{j, \phi, d\}$, where j is the trajectory represented as time-series data, ϕ is the heading measured in degrees, and d is the distance in meters from a shared reference point. We estimate each user's heading from the visual data, ϕ_c by calculating the angle between the user's position in two consecutive frames, with respect

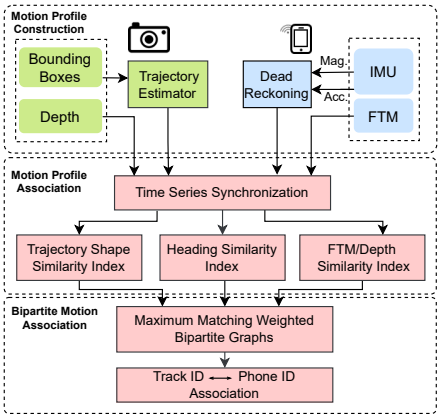


Figure 5: Bipartite motion association overview

to the image x-axis. For each phone, the heading ϕ_p is computed as the rotation around the gravity axis after transformation to global reference frame.

For trajectory estimation, the tracking module from ZED API [11] was found to be most accurate, in comparison to other open-sources methods [18, 55]. Each track is a series of bounding boxes. In order to estimate the trajectory from the phone, we leverage the data from the IMU. We detect steps from the accelerometer signal and employ pedestrian dead reckoning to estimate trajectories. We initialize the starting position as the reference for the dead reckoning process. Then, we use the previously determined heading angle ϕ_p and the average step length for adults (l = 0.8m) to update the phone position over time.

We utilize depth information computed by the RGB-D camera to estimate the distance between the camera and visually tracked persons. On the other hand, WiFi Fine Time Measurements (FTM) [15], currently on Android phones [14], can estimate distance between a phone and a WiFi access point, supporting FTM positioning, with a meter-level [1, 29] accuracy. This provides the timestamped WiFi FTM ranges.

Motion Profile Association. We synchronize the data streams using their timestamps and time offset between their clocks, followed by re-sampling to compute pairwise similarity indices. We compute similarity measures between each pair of motion extracted features. All similarity indices are computed for the entire motion history for the detected people in a video frame, as shown in Figure 6. We use dynamic time warping (DTW) to compute (i) similarity index I_ϕ between heading estimates ϕ_c and ϕ_p; and (ii) similarity index I_j by comparing the shapes of the trajectories j_c and j_p. Even though the measured trajectories are in different coordinate frames, comparing shape allows us to leverage relative changes in users' position and heading to match visual trajectories with IMU trajectories. Lastly, we compute similarity index I_d to correlate FTM and depth measurements. Depth matching addresses the limitations of two-dimensional trajectory matching, which is insufficient when multiple users walk in the same direction. For example, two users walking in a straight line along same direction will exhibit similar trajectory shapes.

Association via Bipartite Graph. To associate phone IDs with visually generated track IDs, we employ a bipartite graph. Bipartite graphs have been used to model 1 : 1 matching problems. In our bipartite graph, $G = \{U, V, E\}$, U represents the set of connected phones, and V is the set of track IDs from the camera at each frame, and E is the set of edges which connect each phone from set U with a person from V. Since track IDs appear and disappear as users move in and out of the camera's field of view and not all users' phone communicate with our system, it often results in V and U having different number of nodes. To ensure 1 : 1 matching, we balance the number of nodes by adding new nodes, labeled "∅". We compute edge scores for G by combining the similarity indices as a weighted sum, $\sum w_i I_i$, over all the motion features. After scanning all possible weight combinations, the edge score is computed as, $E_{uv} = 0.1I_\phi + 0.1I_j + 0.8I_d$. In the case that the depth and FTM similarity I_d is not incorporated in the score, we adjust the weights and the edge score becomes: $E_{uv} = 0.6I_\phi + 0.4I_j$. We do not incorporate I_d to evaluate the performance using IMU data alone as we discuss in the evaluation section. We input the weighted bipartite graphs into the Hungarian algorithm [21] to match the track IDs in the scene to phone IDs. The Hungarian algorithm provides the maximum possible matching between track IDs and phones, preserving the minimum cost, which is the weighted sum in our case. In Figure 6, we demonstrate the bipartite graph construction. In the example shown, bounding boxes 2, 3, and 4 are in the camera FOV at time t_k and there are 3 phones connected to the network A, B and C. As mentioned earlier, the visual tracker generates two separate bounding boxes (1 and 4) for the user carrying phone A. After the bipartite matching, phones A and B are correctly matched to the corresponding bounding boxes, whereas User D who does not carry a phone, is matched with phone node "∅" and Phone C identified as not having a corresponding visual bounding box at time t_k.

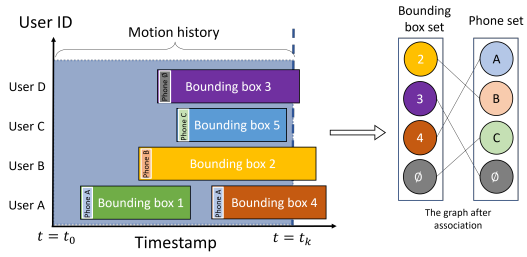


Figure 6: An example of bipartite matching at time t_k. Small vertical boxes overlaid on horizontal boxes denote ground truth associations. The edge weights are computed over the motion history from t_0 to t_k. The algorithm associates bounding Box 4 with Phone A, bounding box 2 with Phone B. Bounding box 3 is associated with "∅" to indicate that the user's phone is not opt-in. Phone C is associated with "∅" because its user is not in the camera's field of view.

6 AFFINITY MATRIX LEARNING

We use the notation of affinity matrix [50] in the context of a deep learning model to facilitate learning advanced feature embedding for the multi-modal data. In our multi-modal association problem, an affinity matrix $\mathcal{M} \in \mathbb{R}^{M \times N}$ quantifies the similarity between instance $i \in [1, M]$ from group A (smartphone information) and instance $j \in [1, N]$ from group B (camera information). Fig. 7(a) and 7(b) show an example of ground truth affinity matrix for 3 smartphones and 7 visually detected bounding boxes.

In our approach, we directly learn an affinity matrix $\mathcal{M}$ to associate all the phone-camera pairs at every timestamp. Each entry of the affinity matrix $\mathcal{M}(i, j)$ represents the association score for the i-th smartphone and the j-th bounding box. Applying the Hungarian algorithm to the affinity matrix or taking row-wise (or column-wise) softmax of the matrix allows us to obtain the association decision for every camera-phone pair, thus successfully associating each bounding box from camera domain to the correct smartphone ID.

6.1 Model Architecture

Fig. 8 depicts the architecture of our multi-modal deep affinity network. The model takes inputs of two modalities as two branches. The upper branch input consists of RGB-D information that comes

(a) An example screenshot from the experiment. 7 pedestrians are detected by human detector. 3 pedestrians' phones are exchanging FTM messages with the WiFi access point while recording IMU measurements.

	1	*2*	*3*	*4*	*5*	*6*	*7*	∅
A	0	**1**	0	0	0	0	0	0
B	**1**	0	0	0	0	0	0	0
C	0	0	0	0	0	0	**1**	0
∅	0	0	**1**	**1**	**1**	**1**	0	0

(b) Vision-wireless ground truth association for the above screenshot represented by an affinity matrix.

Figure 7: An example affinity matrix $\mathcal{M}$ for 3 smartphone devices (A ~ C) and 7 visually detected pedestrians (*1* ~ *7*). The "1" cells suggest that bounding box *1* is associated to phone B, bounding box *2* is associated to phone A and bounding box *7* is associated to phone C. "∅" indicates no association existed for a phone device or bounding box.

from the camera modality. The lower branch of the input consists of FTM and IMU information that come from phone modality. More specifically, for every timestamp of the video sequence, we feed the upper branch with a history of depth measurements d as well as pixel coordinates (x, y) of all the bounding boxes detected by the object tracker. We denote the camera input as $I_c \in \mathbb{R}^{N\times 3\times k}$ where N represents the number of visually detected bounding boxes and k represents the length of time window history. Meanwhile, we feed the lower branch with the corresponding history of FTM measurements (FTM range r, FTM standard deviation std) as well as IMU sensor data (accelerometer $(x_{acc}, y_{acc}, z_{acc})$, gyroscope$(x_{gyr}, y_{gyr}, z_{gyr})$, and magnetometer$(x_{mag}, y_{mag}, z_{mag})$) from all the participants' smartphones. We denote the smartphone input as $I_p \in \mathbb{R}^{M\times 11\times k}$, where M represents the number of smartphone devices communicating with the access point and k represents the length of time window synchronized with camera information.

We adopt LSTM units as feature extractors for the multi-modal input considering they offer significant advantages over other vanilla multi-layer network architectures when extracting features from sequential or time-series data. Each LSTM unit (2-layer, bidirectional) in parallel renders every sequence of measurements into a feature vector $f \in \mathbb{R}^{32\times 1}$ for a visual bounding box or a smartphone device. When the number of detected bounding boxes is N and the number of participating smartphones is M, the output feature maps of two LSTM units are $f_v \in \mathbb{R}^{N\times 32}$ and $f_p \in \mathbb{R}^{M\times 32}$ respectively. Let the maximum number of participants presented simultaneously be $N_m^v (N <= N_m^v)$ and the maximum number of phone holders be $N_M^p (M <= N_m^p)$. We append extra zero row-vectors (representing participants that are not present) to f_v and f_p so that $f_p \in \mathbb{R}^{N_m^p\times 32}$, $f_v \in \mathbb{R}^{N_m^v\times 32}$. Thus each row of the feature map f_v and f_p is a feature vector for a bounding box or a smart phone device, and the padded zero vectors represent the participants that are not present. Since N_m^v bounding boxes and N_m^p smartphone devices can form $N_m^v \times N_m^p$ potential association pairs, we enumerate every possible concatenation of a visual feature vector and a phone feature vector to form a feature cubic $\Phi \in \mathbb{R}^{N_m^p\times N_m^v\times 64}$. Thus each potential association pair is represented by a 64 dimensional vector.

The feature cubic is then compressed to an affinity matrix $\mathcal{M} \in \mathbb{R}^{N_m^p\times N_m^v}$ by a compression network that consists of a sequence of convolution layers with 1×1 kernel sizes. Each element of the affinity matrix, $\mathcal{M}_{i,j}$, denotes the association score for the i-th bounding box and the j-th smartphone device. In order to handle the situations where a bounding box is associated with none of the phone devices or vice versa, we append an extra row and column to the affinity matrix $\mathcal{M}$ to obtain $\mathcal{M}_r \in \mathbb{R}^{(N_m^p+1)\times N_m^v}$ and $\mathcal{M}_c \in \mathbb{R}^{N_m^p\times(N_m^v+1)}$. For $\mathcal{M}_r$, its i-th column associates the i-th bounding box with $N_m^p + 1$ devices, where the "+1" indicates the extra unidentified bounding box that does not associate to any of the existing devices. By applying a row-wise softmax operation over M_r, we obtain a matrix $\mathcal{A}_r \in \mathbb{R}^{(N_m^p+1)\times N_m^v}$ whose rows encodes probabilistic associations between smartphone devices and visually detected bounding boxes. Similarly, we can also apply a column-wise softmax operation to obtain $\mathcal{A}_c \in \mathbb{R}^{N_m^p\times(N_m^v+1)}$, whose columns encode the backward probabilistic associations between bounding boxes and phones.

6.2 Network Loss

During training phases, we compute the network loss for back propagation using ground truth affinity matrices $\mathcal{M}_g \in \mathbb{R}^{(N_m^p+1)\times(N_m^v+1)}$ and the predicted affinity matrices $\mathcal{A}_r \in \mathbb{R}^{(N_m^p+1)\times N_m^v}$ and $\mathcal{A}_c \in \mathbb{R}^{N_m^p\times(N_m^v+1)}$. The total loss L is the average of the following sub-losses: 1) Phone-to-camera association loss L_{pc}, which penalizes incorrect bounding box ID assignments for phone devices. 2) Camera-to-phone association loss L_{cp}, which penalizes incorrect phone device ID assignments for bounding boxes. 3) Consistency loss L_{cons}, which encourages consistency between L_{pc} and L_{cp}. 4) Affinity loss L_{aff}, which suppresses non-maximum entries in the affinity matrix. Detailed definitions of the losses are

$$L_{pc}(\mathcal{M}_g^{\dagger}, \mathcal{A}_r) = \frac{\sum(\mathcal{M}_g^{\dagger} \odot (-\log \mathcal{A}_r))}{\sum(\mathcal{M}_g^{\dagger})}, \tag{1}$$

$$L_{cp}(\mathcal{M}_g^{\ddagger}, \mathcal{A}_c) = \frac{\sum(\mathcal{M}_g^{\ddagger} \odot (-\log \mathcal{A}_c))}{\sum(\mathcal{M}_g^{\ddagger})}, \tag{2}$$

$$L_{cons}(\mathcal{A}_r^{\dagger}, \mathcal{A}_c^{\ddagger}) = \|\mathcal{A}_r^{\dagger} - \mathcal{A}_c^{\ddagger}\|_1, \tag{3}$$

$$L_{aff}(\mathcal{M}_g^{\dagger\ddagger}, \mathcal{A}_r^{\dagger}, \mathcal{A}_c^{\ddagger}) = \frac{\sum(\mathcal{M}_g^{\dagger\ddagger} \odot (-\log(\max(\mathcal{A}_r^{\dagger}, \mathcal{A}_c^{\ddagger}))))}{\sum(\mathcal{M}_g^{\dagger\ddagger})}, \tag{4}$$

$$L = \frac{L_{pc} + L_{cp} + L_{cons} + L_{aff}}{4}, \tag{5}$$

where the notations of "†" and "‡" represent the operations of trimming the last row and the last column of a matrix respectively. "⊙" represents the operation of element-wise product. "∑" represents the operation of taking element-wise sum of a matrix.

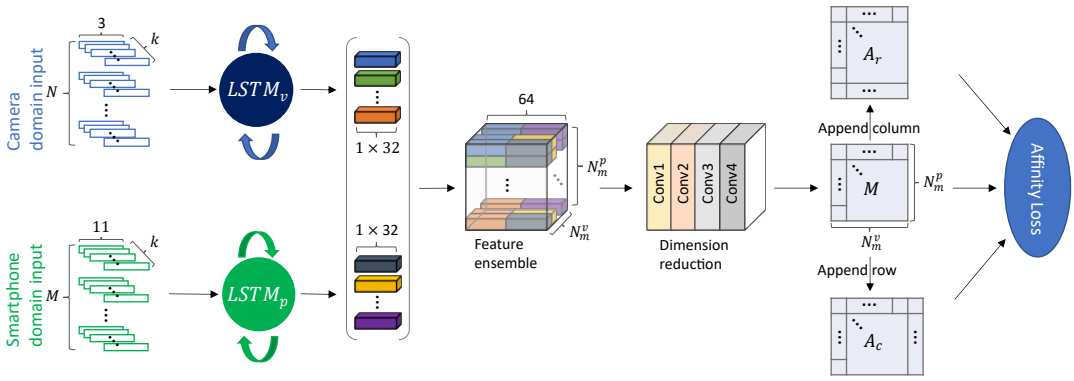


Figure 8: Deep affinity multi-modal architecture. For camera modality, sequential measurements of each detected participants (bounding box coordinates and depth measurements) are fed into a bidirectional LSTM unit. For phone modality, sequential wireless measurements (FTM and IMU measurements) are fed into to another bidirectional LSTM unit. The output feature embeddings are exhaustively combined to form a feature ensemble in which every pair of camera-phone association is encoded by a concatenated embedding vector. The 3D ensemble cubic is squashed to a 2D affinity matrix by fully convolutinoal layers. Each cell of the affinity matrix represents the probability of associating a visually detected person with a smartphone identifier.

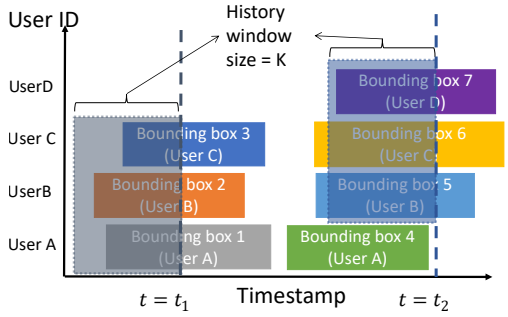


Figure 9: Varying duration of detected tracklet during test phase. The width of a solid rectangle represents a tracklet's duration of presence. User B and C exit and re-enter the scene with different tracklet IDs. We make association decisions for every frame as the time cursor (vertical dashed line) moves from left to right. The association prediction made at timestamp t is based on measurements within a sliding window.

6.3 Network deployment

Online association. After training the network, we conduct online association prediction for every frame of the video. At a specific timestamp t, we aim to associate the current computed visual detections (bounding boxes) with the correct smartphone IDs by using histories of measurements from both camera and smartphone modalities. Due to the random walking pattern of pedestrians and occlusions, every tracklet (bounding boxes with temporarily fixed detection ID) provided by ZED object tracking module has different duration of presence. Unlike the baseline method that utilizes whole motion history to make association decisions (depicted in 6), when making inference with the deep affinity network, we apply a sliding window mechanism so that at most 10 frames' measurements (contains approximately 3 seconds information) are considered.Setting such a limited sliding window whose size is not growing with program's execution is vital for tolerable processing time, which paves the way to real-time performance with marginal latency. Figure 9 illustrates two testing cases. At timestamp $t = t_1$, there are 3 detected pedestrians (User A, B and C) with bounding boxes' duration of presences all being less than 10 frames; At timestamp $t = t_2$, there are 3 detected pedestrians (User B, C and D) with different track IDs. Two of them having measurement histories that are greater than 10 frames. Notice that we evaluate our algorithms under real-world environment with minimal constraint, where users and passer-by pedestrians walk freely in the area. There exist a large portion of situations where users may exit and re-enter the camera's field of view asynchronously.

Offline Association using Consistency voting. Leveraging a series of previous measurements allows us to predict association at every timestamp. However, the association algorithm has no "memory" at this point. For every timestamp, the association is made only for the current time and previous association decisions have no influence on the current frame's association. In order to improve consistency of association predictions among consecutive timestamps, we introduce a sliding window voting scheme on top of the association predicted at every timestamp. The voting scheme maintains a buffer that contains a certain number of most recent association decisions. For a specific bounding box at a timestamp, the final decision is the majority vote among the history of decisions in the sliding window.

7 EVALUATION

We evaluate the performance Vi-Fi based on the correctness, quantified using *accuracy* metric, of association for various modality combinations. We conduct micro-benchmark analysis to study the impact of different factors and parameters on association accuracy.

Data and Processing Preparation. We note that a participant is a Vi-Fi opt in user and a passerby is a randomly occurring pedestrian in the scene. Dataset A contains 15 sequences with 5 participants

Input	$I_c \in \mathbb{R}^{N\times3\times10}$	$I_p \in \mathbb{R}^{M\times11\times10}$
Feature extraction	LSTM_v (3, 32)	LSTM_p (11, 32)
Dimension reduction	Conv1 (64, 128) BatchNorm2D, ReLU Dropout2D	
	Conv2 (128, 64) BatchNorm2D, ReLU	
	Conv3 (64, 32) ReLU	
	Conv4 (32, 1) ReLU	
Output	$\mathcal{M} \in \mathbb{R}^{5\times15}$	

Table 2: Detailed configuration of the Affinity Matrix Learning Network. All convolutional layers have 1×1 kernels.

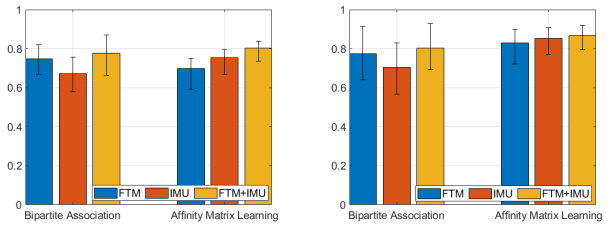
and 0 passerby pedestrians. Dataset B contains 75 sequences collected from 5 different outdoor locations where the number of participants varies between 2 and 3, and the number of passerby pedestrians varies from 1 to 9. To construct training/testing samples, we extract windows of length 10 samples from vision and wireless data of all the participants and passerby pedestrians. For dataset A, the total number of training samples is 540,000, including images, FTM and IMU data. We train and evaluate the deep neural network's performance under the paradigm of cross validation. In order to fully exploit the dataset while avoiding information leakage during training phases, we conduct leave-one-out cross validation for the 15 sequences. For dataset B, the total number of data samples constructed is 1,070,000. Given the larger dataset, we conduct a 5-fold cross validation on 70 sequences and test on 5 unseen sequences that are randomly picked from different scenes. We implement the network architecture using PyTorch [44] – the detailed configurations of the network layers are listed in Table 2. We train the network with an NVIDIA 1080-Ti GPU for 200 epochs with batch size of 32, learning rate 0.001 (0.0001 after 100 epochs).

Evaluation Methodology. We evaluate the two proposed methods for association under online and offline modes. For the bipartite matching algorithm, the online mode takes as input a history of measurements from the beginning of the video to the current timestamp, and then make association decisions for the current timestamp frame by frame. In the offline mode, the ZED tracker is first applied to the video and tracklets (contiguous image frame sets) with different tracking IDs are obtained. Then the bipartite matching algorithm is invoked so that every tracklet is assigned to a smartphone ID. For the affinity matrix learning, in the online mode the pre-trained network *predicts* the bounding boxes' identities for each frame of the ZED camera video feed using a history (up to 10 timestamps) of most recent measurement. In its offline mode we employ the consistency voting scheme that updates each frame's prediction over set of 30 frames, equivalent to 10 seconds (practically reasonable for offline processing).

Evaluation Metric. We use identification precision (*IDP*) [46] and define association accuracy as

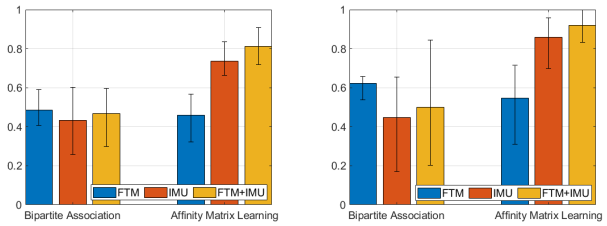
$$IDP = \frac{IDTP}{IDTP + IDFP}$$

where IDTP (IDFP) represents the number of correctly (incorrectly) associated bounding boxes. IDP essentially computes the fraction



(a) Online Processing (b) Offline Processing

Figure 10: Vi-Fi association accuracy for Dataset A. We compare the performance of affinity matrix deep learning approach with the bipartite motion baseline method.



(a) Online Processing (b) Offline Processing

Figure 11: Vi-Fi association for Dataset B. We compare the performance of affinity matrix deep learning approach with the bipartite motion baseline method.

of detected bounding boxes that are correctly associated with their wireless devices. Other metrics including identification recall (IDR) and identification F1 score (IDF1) also take into account false detection or miss detection which are dependent on the tracker's performance. In our evaluation, we focus on IDP instead of IDR or IDF1 because our methods take as input the tracker's outputs, and the associations are computed only for the detected bounding boxes that are present in the scene.

7.1 Vi-Fi Accuracy on Dataset A

We summarize and present the overall association accuracy for dataset A for affinity matrix and bipartite motion (baseline) methods, and under online and offline modes in Figure 10. To better understand how each sub-modality from the phone perspective contributes to the task of association, we conduct an ablation study on both methods. We show and compare the association accuracy for the bipartite matching algorithm and the neural network when the phone modality only comprises the FTM or IMU data. As can be observed from Figure 10, the best association performance results from leveraging features from both WiFi FTM measurements and IMU sensor data. We observe sub-optimal performances when we use only one modality to associate camera data with phone data. It is worth noticing that FTM and IMU sensors encode different aspects of spatial information of pedestrians and compensate each other. More specifically, FTM measurements only capture users' relative distances to the access point. Relying solely on FTM measurements may not distinguish situations where participants at the same distance from the access point have different walking

patterns. On the other hand, IMU data only encodes users' moving patterns or gait information. Thus associating only IMU data can fail in cases where participants have similar walking patterns. We also observe that, though the difference is small, the data-driven deep learning approach outperforms the baseline model-driven approach, showing the importance and encouraging results of applying deep learning and affinity matrix concepts for multi-modal association problem.

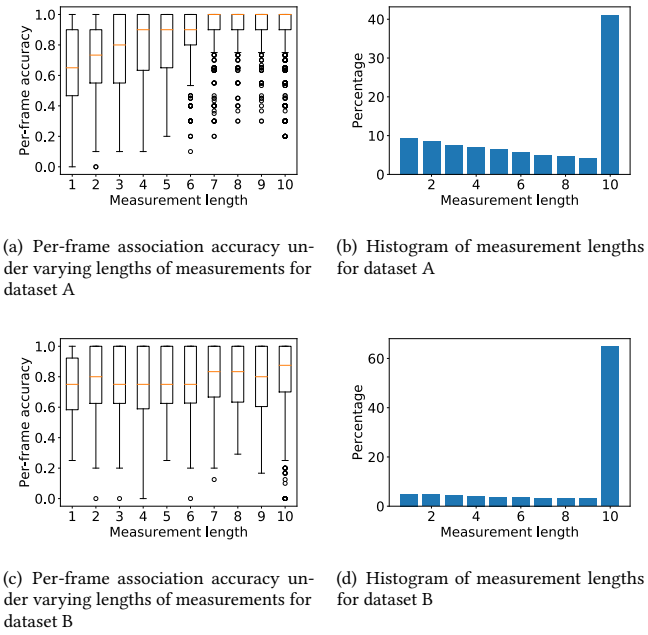
7.2 Vi-Fi Accuracy on Dataset B

We present the average online and offline association accuracy for dataset B in Figure 13(b). While the performance of bipartite algorithm significant degrades, the affinity matrix learning presents consistent association accuracy (81.1% online and 90.2% offline) compared with results of dataset A. Moreover, we observe larger performances gaps in the ablation study for dataset B. Training the network with only FTM sub-modality results in 45.8% online accuracy and 50.1% offline accuracy. The 8% gap between IMU only and FTM+IMU indicate that, although features learned from FTM alone are not representative to deliver good association accuracy, combining FTM with IMU measurements indeed help the network to learn a more distinctive feature representation compared to learning using IMU measurements only.

The performance drop for bipartite matching algorithm suggests the limitation of *hand-crafted feature embedding*, that it is limited to more constraint environments like dataset A, where there is no passerby pedestrians in the camera's view so that a person detected by the camera must always be one of the participants. In comparison, the deep learning model is capable to learn a more complex feature embedding for both modalities such that it can handle situations not limited to dataset A, but other challenging scenarios in dataset B where the number of passerby pedestrians varies and a camera detected pedestrian might not belong to any of the (opt in) participants.

7.3 Microbenchmarks for Affinity Matrix Learning

7.3.1 Varying Lengths of Measurement Sequences. When making association predictions based on different lengths of measurements, it is natural to speculate that longer time series of measurements are more helpful in solving the association problem, since they encode more information of users' motion. In order to evaluate the influence of measurement series lengths on association accuracy, we compute per-frame association accuracy with different minimum measurement lengths at a timestamp. From Figure 12(a) and 12(c), we observe higher association accuracy when we have longer measurement series lengths. Especially, we achieve 89% median accuracy for dataset B when we have measurement lengths that are great or equal to 10 (3 seconds). Figure 12(b) and 12(d) further show the histogram of measurement lengths in our dataset. From the histogram we observe a large portion of testing cases where we have measurement series that are long enough ($\geq$ 3 seconds) for us to make reliable association decisions. This makes the affinity learning a favorable approach especially for online processing, where we only rely on 3 seconds instead of the whole history to make associations with high accuracy.



(a) Per-frame association accuracy under varying lengths of measurements for dataset A

(b) Histogram of measurement lengths for dataset A

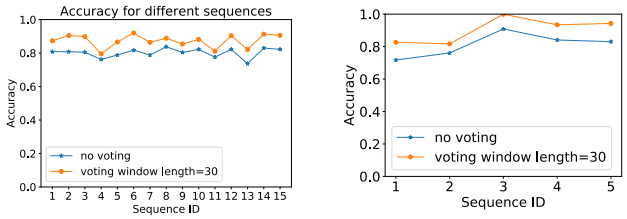
(c) Per-frame association accuracy under varying lengths of measurements for dataset B

(d) Histogram of measurement lengths for dataset B

Figure 12: Length of measurement series affects association accuracy for dataset A and B. Longer series of measurements result in better association.

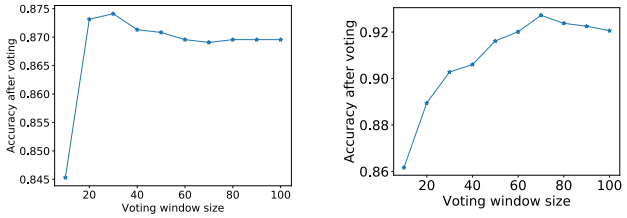
7.3.2 Effect of Consistency Voting. We applied the voting scheme with voting window size 30 to every video sequence of our dataset and compare the per-frame and voted association accuracy for each recorded video. As Figure 13 shows, the voting scheme is capable of correcting inconsistent predictions within a series of per-frame predictions and improve the over-all accuracy by up to 12%. Applying the voting scheme on top of the per-frame prediction means we need extra computational resources to store the prediction history. In order to find the optimal history length, we further explore how different voting window sizes might affect the overall accuracy. Fig. 14 presents the relationship between majority vote window length and overall association accuracy. When the size of the voting window reaches 30 frames ($\approx$ 10 seconds), the voted association accuracy starts to plateau. Considering that the associations for the first 30 timestamps (10 seconds) be conducted *offline*, the insight from this result suggests the feasibility to maintain a history buffer of per-frame predictions for the most recent 10 seconds, thus allowing for real-time voting and prediction updates in real-time, on the fly.

7.3.3 Real-time Performance. Unlike methods such as [27, 31, 56, 59] that post-processed all vision tracklets with wireless measurements over a long period, our proposed affinity matrix prediction supports real-time execution. It requires no future information and only a limited range of sequential information from up to 10 most recent frames (3 seconds) to make accurate association prediction for the current frame. Our current real-time prototype system achieves the association at an average processing rate of 2.8 fps, which is equivalent to an end-to-end processing time of 360ms.



(a) Dataset A (b) Dataset B

Figure 13: Association accuracy on testing set



(a) Dataset A (b) Dataset B

Figure 14: Voted association accuracy v.s. different voting window size

8 DISCUSSION

We will discuss some of the limitations of the current version of the Vi-Fi system:

Scaling the number of pedestrians: The maximum number of pedestrians and phones that Vi-Fi can handle is defined by the dimension of the affinity matrix in the network architecture. This number is pre-defined for the network architecture and cannot be changed *after* the training process. If we are required to make associations for more crowded scenes, we will need to modify the dimension of the affinity matrix and fine-tune the network with the updated architecture. The fine-tuning process does not require a complete re-training of the network from scratch and thus can be conducted in the form of a in-frequent network architecture update.

Privacy: Although Vi-Fi will not run automatically in the background and users need to give consent to opt-in, we are aware that associating camera monitored pedestrians with their smartphone ID may still impose privacy concerns. For example, malicious apps might obtain access to the camera and wireless channel so that users' information is breached without consent. Addressing the privacy concerns would require separate experiment designs, systematic studies and evaluations, which are outside the scope of the current work and considered as future work.

9 CONCLUSION

In this work, we addressed the fundamental problem of associating subjects observed in camera views and messages transmitted from their wireless devices. We designed a recurrent deep affinity matrix learning architecture that learns more distinctive representations from raw sequential sensor data measurements of vision and wireless modalities. In particular, our approach builds a latent space representation and uses the notation of affinity matrix to compute correlation between phone ID available through WiFi meta-data, and images, WiFi FTM depth and IMU sensor values. We developed a bipartite matching based baseline approach that constructs hand crafted motion features from IMU and FTM measurements and builds similarity correlation checks using bipartite graph matching. To facilitate design and evaluation, we collected and evaluated a large scale multi-modal dataset at different real-world environments with varying number of participants and passerby pedestrians. Our evaluation results show that the learned features from our proposed network architecture are more distinctive for challenging crowded environments where varying number of passerby pedestrians exist. The proposed network architecture achieves an overall association accuracy between 81% (real-time) to 91% (offline) across diverse real-world environments.

10 ACKNOWLEDGEMENTS

This research has been supported by the National Science Foundation (NSF) under Grant No. CNS-1901355, CNS-1910170, CNS-1901133, CNS-2055520. Thanks to Rashed Rahman, Shardul Avinash, Abbaas Alif, Bhagirath Tallapragada and Kausik Amancherla for their help with data labeling.

REFERENCES

[1] [n.d.]. https://goo.gl/BSUCdG. Wi-Fi CERTIFIED Location.
[2] [n.d.]. https://goabode.com/product/abode-cam-2. Adobe Security Camera.
[3] [n.d.]. https://store.google.com/us/product/nest_cam_battery?hl=en-US. Google Nest Camera.
[4] [n.d.]. https://ring.com/security-cameras. Ring Security Camera.
[5] [n.d.]. https://sites.google.com/winlab.rutgers.edu/vi-fidataset.
[6] [n.d.]. https://github.com/vifi2021/Vi-Fi.
[7] [n.d.]. https://github.com/ORB-HD/deface.
[8] [n.d.]. https://data.bris.ac.uk/data/dataset/1gt0wgkqgljn21jjgqoq8enprr. SPHERE-Calorie dataset.
[9] [n.d.]. https://github.com/ale152/video-accelerometer-matching.
[10] [n.d.]. :https://www.tnt.uni-hannover.de/en/project/VIMPT2019.
[11] [n.d.]. https://www.stereolabs.com/docs/object-detection/.
[12] [n.d.]. https://github.com/microsoft/VoTT.
[13] [n.d.]. https://web.archive.org/web/20120105112913/http://www.math.harvard.edu/archive/20_spring_05/handouts/assignment_overheads.pdf. The Assignment Problem and the Hungarian Method.
[14] [n.d.]. https://bit.ly/2Wn3FvS. Wi-Fi location: ranging with RTT.
[15] 2016. "IEEE Standard for Information technology–Telecommunications and information exchange between systems Local and metropolitan area networks–Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications". *"IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)"* (Dec 2016), 1–3534. https://doi.org/10.1109/IEEESTD.2016.7786995
[16] Alexandre Alahi, Albert Haque, and Li Fei-Fei. 2015. RGB-W: When vision meets wireless. In *Proceedings of the IEEE International Conference on Computer Vision*. 3289–3297.
[17] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.
[18] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. 2019. Tracking without bells and whistles. In *Proceedings of the IEEE international conference on computer vision*. 941–951.
[19] Siyuan Cao and He Wang. 2018. Enabling Public Cameras to Talk to the Public. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2 (July 2018), 1–20.
[20] Siyuan Cao and He Wang. 2018. Enabling public cameras to talk to the public. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–20.
[21] Yi Cao. 2020. *Munkres Assignment Algorithm*. https://www.mathworks.com/matlabcentral/fileexchange/20328-munkres-assignment-algorithm
[22] Shiwei Fang, Tamzeed Islam, Sirajum Munir, and Shahriar Nirjon. 2020. EyeFi: Fast Human Identification Through Vision and WiFi-based Trajectory Matching. In *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 59–68.

[23] Shahina Ferdous, Kapil Vyas, and Fillia Makedon. 2012. A survey on multi person identification and localization. In *Proceedings of the 5th International Conference on Pervasive Technologies Related to Assistive Environments*. 1–3.

[24] Beatriz Quintino Ferreira, Joao Gomes, Cláudia Soares, and Joao P Costeira. 2018. FLORIS and CLORIS: Hybrid source and network localization based on ranges and video. *Signal Processing* 153 (2018), 355–367.

[25] Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. *IEEE Access* 7 (2019), 63373–63394.

[26] Roberto Henschel, Timo von Marcard, and Bodo Rosenhahn. 2019. Simultaneous identification and tracking of multiple people using video and IMUs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[27] R. Henschel, T. Von Marcard, and B. Rosenhahn. 2020. Accurate Long-Term Multiple People Tracking Using Video and Body-Worn IMUs. *IEEE Transactions on Image Processing* 29 (2020), 8476–8489. https://doi.org/10.1109/TIP.2020.3013801

[28] Panwen Hu, Zizheng Yan, Rui Huang, and Feng Yin. 2019. How Effectively can Indoor Wireless Positioning Relieve Visual Tracking Pains: A Cramer-Rao Bound Viewpoi. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3083–3087.

[29] Mohamed Ibrahim, Hansi Liu, Minitha Jawahar, Viet Nguyen, Marco Gruteser, Richard Howard, Bo Yu, and Fan Bai. 2018. Verification: Accuracy Evaluation of WiFi Fine Time Measurements on an Open Platform. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 417–427.

[30] Tatsuya Ishihara, Kris M Kitani, Chieko Asakawa, and Michitaka Hirose. 2018. Deep Radio-Visual Localization. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 596–605.

[31] Deokwoo Jung, Thiago Teixeira, and Andreas Savvides. 2010. Towards cooperative localization of wearable sensors using accelerometers and cameras. In *2010 Proceedings IEEE INFOCOM*. IEEE, 1–9.

[32] Deokwoo Jung, Thiago Teixeira, and Andreas Savvides. 2010. Towards Cooperative Localization of Wearable Sensors using Accelerometers and Cameras. In *2010 Proceedings IEEE INFOCOM*. 1–9. http://dx.doi.org/10.1109/INFCOM.2010.5462059

[33] 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*. 1–15.

[34] Hansi Liu, Pengfei Ren, Shubham Jain, Mohannad Murad, Marco Gruteser, and Fan Bai. 2019. Fusioneye: Perception sharing for connected vehicles and its bandwidth-accuracy trade-offs. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[35] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. When Video meets Inertial Sensors: Zero-shot Domain Adaptation for Finger Motion Analytics with Inertial Sensors. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. 182–194.

[36] Chris Xiaoxuan Lu, Hongkai Wen, Sen Wang, Andrew Markham, and Niki Trigoni. 2017. SCAN: learning speaker identity from noisy sensor data. In *2017 16th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 67–78.

[37] Alessandro Masullo, Tilo Burghardt, Dima Damen, Toby Perrett, and Majid Mirmehdi. 2019. Who Goes There? Exploiting Silhouettes and Wearable Signals for Subject Identification in Multi-Person Environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

[38] Alessandro Masullo, Tilo Burghardt, Dima Damen, Toby Perrett, and Majid Mirmehdi. 2019. Who goes there? exploiting silhouettes and wearable signals for subject identification in multi-person environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

[39] Filippo LM Milotta, Antonino Furnari, Sebastiano Battiato, Maria De Salvo, Giovanni Signorello, and Giovanni M Farinella. 2018. Visitors localization in natural sites exploiting egovision and gps. *Eye* (2018).

[40] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. 2018. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8427–8436.

[41] Le T Nguyen, Yu Seung Kim, Patrick Tague, and Joy Zhang. 2014. IdentityLink: user-device linking through visual and RF-signal cues. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 529–539.

[42] Takayuki Nishio and Ashwin Ashok. 2016. High-speed mobile networking through hybrid mmWave-camera communications. In *Proceedings of the 3rd Workshop on Visible Light Communication Systems*. 37–42.

[43] Shijia Pan, Tong Yu, Mostafa Mirshekari, Jonathon Fagert, Amelie Bonde, Ole J Mengshoel, Hae Young Noh, and Pei Zhang. 2017. Footprintid: Indoor pedestrian identification through ambient structural vibration sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–31.

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[45] Thuy Thi Thanh Pham, Thi-Lan Le, and Trung-Kien Dao. 2016. Fusion of wifi and visual signals for person tracking. In *Proceedings of the Seventh Symposium on Information and Communication Technology*. 345–351.

[46] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. *CoRR* abs/1609.01775 (2016). arXiv:1609.01775 http://arxiv.org/abs/1609.01775

[47] Carlos Ruiz, Shijia Pan, Adeola Bannis, Ming-Po Chang, Hae Young Noh, and Pei Zhang. 2020. IDIoT: Towards ubiquitous identification of iot devices through visual and inertial orientation matching during human activity. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 40–52.

[48] Carlos Ruiz, Shijia Pan, Adeola Bannis, Xinlei Chen, Carlee Joe-Wong, Hae Young Noh, and Pei Zhang. 2018. IDrone: Robust drone identification through motion actuation feedback. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–22.

[49] HA Baier Saip and Claudio Leonárdo Lucchesi. 1993. Matching algorithms for bipartite graph. *Relatorio Tecnico* 700, 03 (1993).

[50] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, and Mubarak Shah. 2019. Deep affinity network for multiple object tracking. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 104–119.

[51] Thiago Teixeira, Gershon Dublon, and Andreas Savvides. 2010. A survey of human-sensing: Methods for detecting presence, count, location, track, and identity. *Comput. Surveys* 5, 1 (2010), 59–69.

[52] Thiago Teixeira, Deokwoo Jung, and Andreas Savvides. 2010. Tasking networked cctv cameras and mobile phones to identify and localize multiple people. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 213–222.

[53] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John P Collomosse. 2017. Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors.. In *BMVC*, Vol. 2. 1–13.

[54] Richard Yi-Chia Tsai, Hans Ting-Yuan Ke, Kate Ching-Ju Lin, and Yu-Chee Tseng. 2019. Enabling identity-aware tracking via fusion of visual and inertial features. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2260–2266.

[55] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. 2019. MOTS: Multi-object tracking and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7942–7951.

[56] He Wang, Xuan Bao, Romit Roy Choudhury, and Srihari Nelakuditi. 2015. Visually fingerprinting humans without face recognition. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. 345–358.

[57] Zhujun Xiao, Yanzi Zhu, Yuxin Chen, Ben Y Zhao, Junchen Jiang, and Haitao Zheng. 2018. Addressing Training Bias via Automated Image Annotation. *arXiv preprint arXiv:1809.10242* (2018).

[58] Jingao Xu, Hengjie Chen, Kun Qian, Erqun Dong, Min Sun, Chenshu Wu, Li Zhang, and Zheng Yang. 2019. ivr: Integrated vision and radio localization with zero human effort. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–22.

[59] Juxiang Zeng, Pinghui Wang, Qiqi Zhao, Jianye Pang, Jing Tao, and Xiaohong Guan. 2019. Effectively Linking Persons on Cameras and Mobile Devices on Networks. *IEEE Internet Computing* 23, 4 (2019), 18–26.

[60] Peijun Zhao, Chris Xiaoxuan Lu, Jianan Wang, Changhao Chen, Wei Wang, Niki Trigoni, and Andrew Markham. 2019. mID: Tracking and identifying people with millimeter wave radar. In *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 33–40.

[61] Dali Zhu, Hongju Sun, and Di Wu. 2021. Fusion of Wireless Signal and Computer Vision for Identification and Tracking. In *2021 28th International Conference on Telecommunications (ICT)*. IEEE, 1–7.

[62] Han Zou, Yuxun Zhou, Jianfei Yang, and Costas J Spanos. 2018. Unsupervised WiFi-enabled IoT device-user association for personalized location-based service. *IEEE Internet of Things Journal* 6, 1 (2018), 1238–1245.