Demo: *Lost and Found!* Associating Target Persons in Camera Surveillance Footage with Smartphone Identifiers

Hansi Liu, Abrar Alali, Mohamed Ibrahim, Hongyu Li, Marco Gruteser, Shubham Jain, Kristin Dana, Ashwin Ashok, Bin Cheng, Hongsheng Lu
Rutgers University, Old Dominion University, Stony Brook University, Georgia State University, InfoTech Labs, Toyota Motor North America R&D
{hansiiii,mibrahim,hongyuli,gruteser}@winlab.rutgers.edu,kristin.dana@rutgers.edu,
jain@.stonybrook.edu,aalal003@odu.edu,aashok@gsu.edu,{bin.cheng,hongsheng.lu}.@Toyota.com

ABSTRACT

We demonstrate an application of finding target persons on a surveillance video. Each visually detected participant is tagged with a smartphone ID and the target person with the query ID is highlighted. This work is motivated by the fact that establishing associations between subjects observed in camera images and messages transmitted from their wireless devices can enable fast and reliable tagging. This is particularly helpful when target pedestrians need to be found on public surveillance footage, without the reliance on facial recognition. The underlying system uses a multi-modal approach that leverages WiFi Fine Timing Measurements (FTM) and inertial sensor (IMU) data to associate each visually detected individual with a corresponding smartphone identifier. These smartphone measurements are combined strategically with RGB-D information from the camera, to learn affinity matrices using a multi-modal deep learning network.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Human-centered computing** → *Ubiquitous and mobile computing systems and tools.*

KEYWORDS

Person identification; WiFi FTM ranging; Machine learning; Multi-modal learning

ACM Reference Format:
Hansi Liu, Abrar Alali, Mohamed Ibrahim, Hongyu Li, Marco Gruteser, Shubham Jain, Kristin Dana, Ashwin Ashok, Bin Cheng, Hongsheng Lu. 2021. Demo: *Lost and Found!* Associating Target Persons in Camera Surveillance Footage with Smartphone Identifiers. In *The 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '21), June 24-July 2, 2021, Virtual, WI, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3458864.3466904

1 INTRODUCTION

Association of cross-domain sensor data is a fundamental need in applications and systems that exploit multi-modal sensor data.

MobiSys '21, June 24-July 2, 2021, Virtual, WI, USA

ACM ISBN 978-1-4503-8443-8/21/06.
https://doi.org/10.1145/3458864.3466904

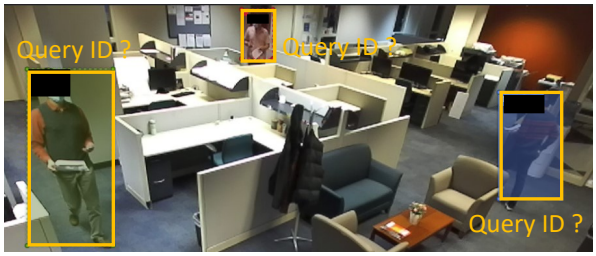

Figure 1: Motivation: A person in a public surveillance footage needs to be matched to a query ID.

With the pervasive use of cameras and wireless devices, one key instance of this problem is the association between persons detected in camera video and wireless data originating from transmitters of these persons.

We propose 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). A key insight is that both cameras and wireless receivers are now becoming capable of improved ranging—cameras through RGB-D technology and WiFi through the Fine Timing Measurement (FTM) standard [1, 3]. Since cameras are also increasingly equipped with WiFi transceivers, 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. We explore a supervised method that leverages 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 multi-modal affinity matrix learning network that learns a latent similarity metric and predicts an affinity matrix for multiple camera-phone pairs.

Solving the association problem for camera and smartphone modalities enables faster and reliable identity tagging for video analysis. As Figure 1 illustrates, when a person in a public area needs to be tagged in the surveillance video, our approach, assuming that person's phone had opted-in to our technology, allows faster tagging compared to manual labeling. Our approach doesn't rely on facial recognition to perform re-identification, which is vulnerable to occlusions or challenging lighting conditions. The method exploits similarities across multi-modal data and assigns smartphone ID to visually detected pedestrians, thus providing

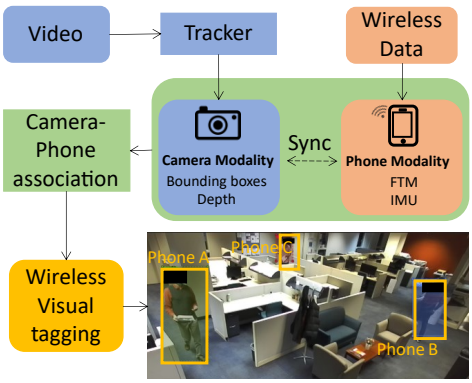

Figure 2: Method overview. The approach takes as input a video sequence and synchronized smartphone data. It associates multi-modal data for the same identities. ID tags are generated for each visually detected participant.

unique and constant tags when pedestrians are occluded or they exit and reappear in the camera view. Usage of this application can be in malls or other public areas where surveillance camera and WiFi access points are accessible.

2 DESIGN OVERVIEW

Figure 2 presents an overview of our approach. The camera detects and tracks moving subjects in its field of view and estimates each participant's depth. Meanwhile, each phone exchanges WiFi FTM messages with the access point while gathering motion sensor measurements including accelerometer, gyroscope, and magnetometer. The access point and the camera is placed close to each other so that FTM measurements are approximately the distances between pedestrians and the camera. For privacy reasons, our method is designed to only track and associate consenting users. Phones of consenting users actively share their measurements with a server that computes associations.

We propose a network architecture that extracts feature embeddings from raw measurement sequences of camera modality (bounding boxes, depth) and smartphone modality (FTM, IMU). the cross-modal feature vectors of the same identity have more similarity since one participant's spatial-temporal information (moving pattern, heading information, etc.) is encoded in both modalities' measurements. The network utilizes the learned features to compute affinity matrices where association probabilities are encoded for every camera-phone pair. An affinity matrix $\mathcal{M} \in \mathbb{R}^{(M+1)\times(N+1)}$ 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). N denotes the maximum number of participants in the camera view and M denotes the maximum number of phones communicating with the access point. The extra row and extra column of the affinity matrix handles the situations where none of the camera (phone) modality is associate to the phone (camera) modality. For example, a pedestrian's bounding box is detected in the camera view but her phone is not connected to the access point. As a result there is no associated identity assigned to the bounding box.

The network takes two branches of inputs from the multi-modal dataset. From camera modality, sequences of measurements for each detected participants (bounding box coordinates, bounding box depth measurements) are fed into an LSTM. From smartphone modality, sequences of smartphone data (FTM and IMU measurements) are fed into another LSTM. The output feature embeddings are then exhaustively combined to form a feature ensemble in which every pair of camera-phone association is represented by a combined embedding vector. By using a sequence of 1×1 convolutional layers, the 3D ensemble cubic is squashed to a 2D affinity matrix, where every cell represents the probability of associating a visually detected participant to a smartphone.

3 DEMONSTRATION

Demo setup. To train the multi-modal network, we will collect a multi-modal dataset that comprises RGBD, WiFi, and IMU information. The dataset will contains videos of indoor scenarios where 5 participants randomly walk around the venue. Each participant will be holding a smart phone that is exchanging FTM messages with an access point (located besides the wall mounted camera) while logging its IMU sensor data. The dataset collection processs conforms to IRB policies of the authors' institutions as well as COVID safety protocol. All participating users will wear masks and will stay 6ft apart during data collection. We will capture video footage in our lab environment and show the workings of our system through annotations and outputs being displayed. We will create a demo video of our approach applied to a recorded footage. We do not need any specific infrastructure for our demonstration, and the ability to stream our video with audio output should be sufficient.

Online association. We adopt the ZED object tracking module [2] to obtain visual detections of the participants and their pair-wise distances with the help of RGBD information. Then we associate the current computed visual detections (bounding boxes) with the correct smartphone Querry IDs by using histories of measurements from both camera (bounding box coordinates and depth measurements) and smartphone modalities (FTM and IMU data). Each entry of the computed affinity matrix $\mathcal{M}(i, j)$ represents the association probability for the i-th smartphone and the j-th bounding box. Applying the 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. We labeled each bounding box by the associated smartphone ID. We highlight the bounding box if the assigned smartphone ID equals to the target query ID.

ACKNOWLEDGEMENT

This research has been supported by the National Science Foundation (NSF) under Grant No. CNS-1901355, CNS-1910170, CNS-1901133, CNS-2055520.

REFERENCES

[1] [n.d.]. https://goo.gl/BSUCdG. Wi-Fi CERTIFIED Location.
[2] [n.d.]. ZED API. https://www.stereolabs.com/docs/object-detection/.
[3] 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