Beyond Video Surveillance: Exploiting Sleep-Talk of Apps to See Smartphone's ID

Zhuochen Fan[†§], Tian Liu[¶], Jun Huang[‡] and Tong Yang[†§]
[†]School of Computer Science, and National Engineering Laboratory for Big Data Analysis Technology and Application, Peking University, China [§]Peng Cheng Laboratory, China
[‡]Department of Computer Science, City University of Hong Kong, China
[¶]Center for Energy-Efficient Computing and Applications, Peking University, China
Emails: {fanzc, liutian}@pku.edu.cn, jun.huang@cityu.edu.hk, yangtongemail@gmail.com

2022 IEEE Wireless Communications and Networking Conference (WCNC) | 978-1-6654-4266-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/WCNC51071.2022.9771759

***Abstract*—Video cameras have been widely deployed at city-scale for security surveillance. However, under poor light condition and limited video resolution, it is often impossible to determine the identity of a pedestrian by using computer vision alone. Motivated by the fast and continuous penetration of smartphones, in this paper, we explore the feasibility of augmenting video cameras to 'see' the identities of smartphones (*e.g.*, MAC addresses, cellular IDs, or even phone numbers) carried by pedestrians. We develop IDCam – a system that integrates a video camera with a smart antenna array, which leverages spatial-domain sensing of smartphone's sleep-talk (*i.e.*, the packets transmitted by apps while the phone's screen is off) to match the angles of smartphone's packets and pedestrians in the video, enabling passive identity linking without the cooperation of target. Experiment results show that IDCam accurately links visual and wireless identities in a complex deployment environment with tens of pedestrians and intensive multipath signal propagation.**

***Index Terms*—array signal processing, computer vision, object detection, sensor fusion**

I. Introduction

On March 7th, 2013, a 5-year old boy was abducted at a square of Nanchang, the capital city of Jiangxi province, China. A video footage (as shown in Fig. 1) of the surveillance camera recorded the whole process. Unfortunately, due to poor image quality, computer vision cannot identify the suspect except giving a coarse profile (sex, height, physique, etc.), which provided little help in tracking down the criminal. However, with the rapid and continuous penetration of smartphones[1], it is very likely that the suspect in Fig. 1 may carry a smartphone. This paper asks the question: can we 'see' the wireless identity (such as MAC address, cellular network ID, or even phone number) of the suspect's smartphone to augment security surveillance? Even if the suspect did not carry a phone or had the phone turned off, knowing the smartphone identities of who witnessed the incident would be greatly helpful, as they may provide a much more detailed profile of the suspect as well as other critical clues.

On the opposite side, a system capable of seeing smartphone identities can be rather privacy-intrusive. The operator of a such augmented surveillance system – either a government agency, a small business, or a personal camera user – could use that system to collect and track pedestrians' identities without their notice and permission. Such link of visual and wireless network identities can be used in a variety of malicious and selfish ways.

Fig. 1. Snapshot of a video footage recording a real incident of child abduction.

Aside of the good and ugly aspects of a such system, this paper explores the technical feasibility of augmenting video surveillance to link pedestrians with wireless identities. Intuitively, a straightforward solution is to first measure the angle-of-arrival (AoA) of a wireless packet, extracting its identity[2], and then matches the AoA with the directions of all pedestrians to determine whose device is the packet sender. This seems to be technically simple. Measuring the angle of a visual object is trivial. High-precision AoA measurement algorithms have been studied for decades.

While the underlying computer vision and signal processing algorithms are readily available, a key question remains open – how can we *passively* link a pedestrian with a wireless device without asking the device to actively transmit packets? Fortunately, smartphones 'talk' while sleeping (*i.e.*, in deep power saving mode when the screen is off) as apps intermittently interact with their cloud servers to upload and pull data. However, a key challenge in exploiting such sleep-talks arises from the fact that, due to signal reflections in practical environments, wireless packets may arrive from multiple directions, making it difficult to identify the line-of-sight (LoS) AoA. In

Corresponding author: Jun Huang (jun.huang@cityu.edu.hk).

[1]According to Newzoo's Global Mobile Market Report, the number of smartphone users in China has reached 782,848,000, with a penetration rate of 55.3%.

[2]In wireless networks, the MAC address of a sender can be obtained from MAC header, which is typically encryption-free.

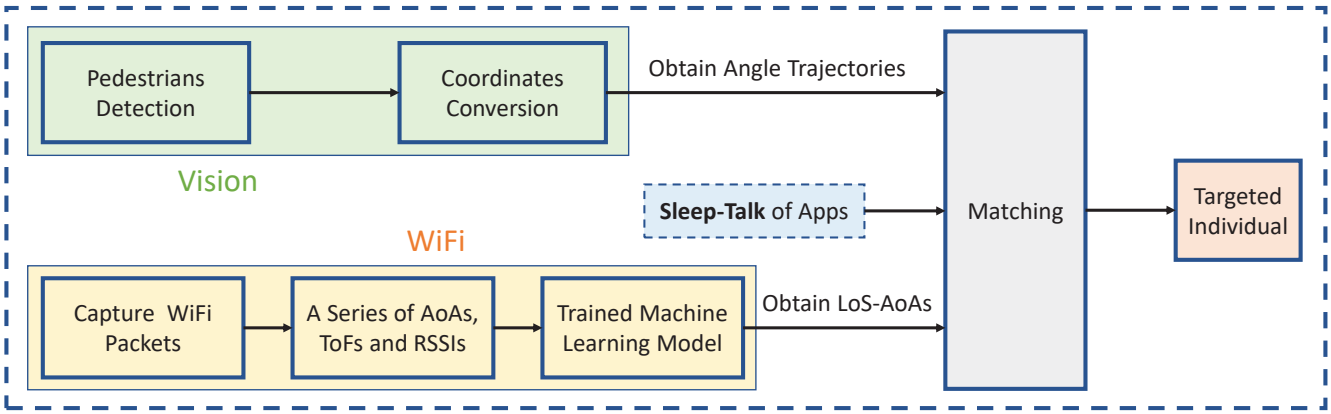


Fig. 2. An overview of IDCam architecture.

particular, in order to precisely measure AoAs and identify the LoS, it typically requires a large number of packets to smooth measurements and suppress noise. While we cannot rely on explicit cooperation from smartphone users (like what previous wireless localization systems do), we ask – do the sleep-talks of smartphone apps sufficiently abundant to support passive identity linking? If not, how do we design a system to tackle this challenge?

In this paper, we characterize smartphone's sleep-talks in terms of talk frequency and the burst length of each talk for two popular apps. We then develop IDCam – a system that fuse a video camera with an antenna array to see smartphone identities. IDCam leverages machine learning to enable accurate identity linking using only a small amount of sleep-talk packets in the presence of intensive multipath and noise. Experiments conducted in a real-world deployment scenario show that IDCam accurately links visual and wireless identities in a complex deployment environment with tens of pedestrians and intensive multipath signal propagation.

II. System Overview

The architecture of IDCam is shown in Fig. 2. The computer vision module detects pedestrians appeared in the video and then tracks their angle trajectories relative to the camera. The antenna array sniffs sleep-talk packets transmitted by nearby devices and extracts device identities from packet headers ($i.e.$, sender MAC addresses). To determine the LoS direction of a smartphone using only a small number of its sleep-talk packets, IDCam applies machine learning to analyze noisy signal AoA, Time-of-Flight (ToF), and received signal strength indicator (RSSI). Finally, the LoS AoA trace of transmitter is compared with pedestrians' angle trajectories to compute a matching score for each pair of pedestrian and transmitter, which characterizes the likelihood of identity link.

III. Technical Background

A. Computer Vision-based Pedestrian Tracking

IDCam employs YOLOv3 [1] trained using the MS COCO dataset [2] for real-time object detection. To obtain angle trajectories, IDCam first transforms the camera coordinate system to the physical coordinate system [3], and then uses Kalman filter [4] to track the angle of detected object relative to the camera. To further improve accuracy, we apply spline interpolation to handle trajectory discontinuities caused by low video quality.

B. Signal AoA Measurement

IDCam sniffs WiFi packets transmitted by pedestrians' smartphones and then measures their AoAs to enable identity linking. Although IDCam targets WiFi, its method is applicable to other wireless technologies, such as 4G/5G/LTE. Specifically, to measure the AoA of received WiFi packets, IDCam employs a uniform linear antenna array and uses a Bartlett beamformer [5] to scan signal power coming from all directions. Compared with other AoA measurement algorithms such as MUSIC [6] and ESPRIT [7] widely used in wireless indoor localization systems, Bartlett [5] is more computationally efficient as it does not require complex matrix computations like eigen-decomposition, thus allowing for real-time AoA analysis for WiFi packets transmitted from the devices of a crowd of pedestrians. Fig. 3 shows the received signal power of one WiFi packet measured using a Bartlett beamformer. We observe five significant power peaks corresponding to the signals coming from five directions, including one LoS path ($i.e.$, LoS AoA) and four reflections. To link identities of wireless device and pedestrians, IDCam must identify the LoS path before performing the match between AoA and the pedestrian's angle trajectory.

Previous studies address this problem by measuring the ToF of signals. For instance, SpotFi [8] measures signal ToF based on the phase shift between two adjacent subcarriers. Specifically, the ToF τ_k of the k^{th} path can be computed as $2\pi\delta f\tau_k$, where δf denotes the frequency interval between subcarriers. Alternatively, based on the fact that the LoS path signals experience the least attenuation due to free of obstruction and reflection, one can compare the signal powers of different paths to identify the LoS signal that has the highest power. However, accurately measuring ToF and signal power typically requires the target device to cooperate in

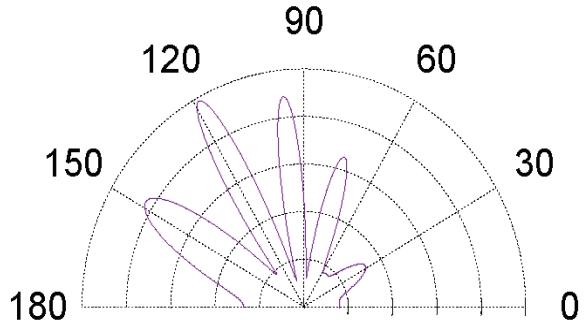


Fig. 3. An example of AoA spectrum measured in a multipath environment.

suppressing noise by sending a large number of packets, which is impractical in passive scenarios. To address this limitation, IDCam applies machine learning on a small volume of sleep-talk traffic to accurately identify the LoS AoA, as detailed in Section V.

IV. Characterizing Smartphones' Sleep-Talks

Smartphones 'talk' while sleeping ($i.e.$, in deep power saving mode when the screen is off) as apps intermittently interact with their cloud servers to upload and pull data. IDCam exploits smartphones' sleep-talk for passive identity linking. In this section, we present measurement results to characterize sleep-talk traffic of smartphones.

We record traffics of smartphones transmitted during sleep mode using Wireshark [9], and then separate the traffics of different apps based on destination IP addresses. We then measure the response time interval between two bursts of sleep-talk packets, and then count the number of packets in each burst.

Fig. 4 shows measurement results for two popular smartphone apps in the categories of instant messaging and video sharing, anonymously named as `APP_1` and `APP_2`, respectively. We observe that both apps transmit sleep-talk packets every tens of seconds in more than 50% and 25% cases. Each burst of sleep-talk contains a couple to tens of packets. Although both apps transmit sleep-talk packets consistently over time, the limited volume of traffic presents a significant challenge for accurately measuring signal AoAs in noisy environments.

V. Design of IDCam

A. Identifying LoS AoA

We propose a machine learning-based method that uses a small number of packets to identify the LoS AoA of the packet transmitter in noisy and multipath-intensive environments. For each received packet, we steer a Bartlett beamformer towards all directions to measure signal powers, and then use a peak detector to extract angles along which signal powers are higher than the noise floor. The extracted angles and the corresponding received powers constitute a multipath AoA spectrum, as shown in Fig. 3. For each path, we further estimate the ToF based on the phase shift between adjacent subcarriers. The obtained AoA, ToF, and signal power are used as features for each path. Intuitively, the LoS path should exhibit the highest signal power, the shortest ToF, and the smallest variance in AoA due to its resilience against reflections from moving objects. Although the measurements of AoA, ToF, and signal powers can be polluted by noise, we intend to leverage the power of machine learning to identify LoS path using only a limited number of noisy measurements.

Table I: 5-fold Cross-validation of model performances

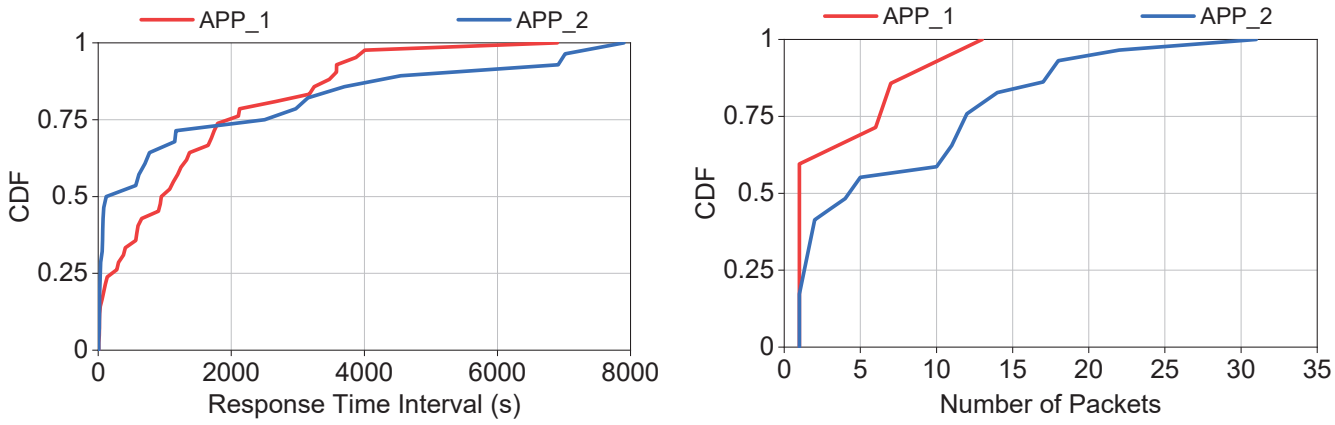
	Precision Rate	Recall Rate	F1 Score
Fine Tree	0.888	0.892	0.890
Kernal Naive Bayes	0.783	0.722	0.751
Fine Gaussian SVM	0.887	0.871	0.879
Weighted KNN	0.903	0.904	0.903
Bagged Tree	0.931	0.910	0.921
RUSBoosted Tree	0.780	0.950	0.857

To achieve this, we train a machine learning model as follows. First, we deploy WiFi transmitters at different locations of the surveillance environment and record the transmitter's angle relative to the antenna array as the groundtruth. We then receive signals from deployed transmitters and process the CSI array to obtain AoA, ToF, and signal powers of each path. Based on the groundtruth of transmitter angle, multipath signal AoAs are labelled as LoS and NLoS before feeding a supervised binary classifier for training. We choose six widely used machine learning algorithms, as listed in Table. I. The collected AoA traces are divided into training and testing sets. A 5-fold cross-validation is then performed to evaluate the performance of selected machine learning algorithms. We compare these algorithms based on three metrics, including precision rate (PR), recall rate (RR), and F1 Score.

The results are reported in Table I. We observe that Bagged Tree [10] achieves the highest F1 score of 0.931 and a good trade-off between PR and RR, which are 0.910 and 0.921, respectively. Fig. 5(a) shows an example of the input and output of Bagged tree. The multipath AoA trace is measured at one location in our deployment site where the transmitter is deployed at 120°. The LoS AoA trace identified by the Bagged Tree is shown in Fig. 5(b), which accurately match the grondtruth.

B. Identity Linking

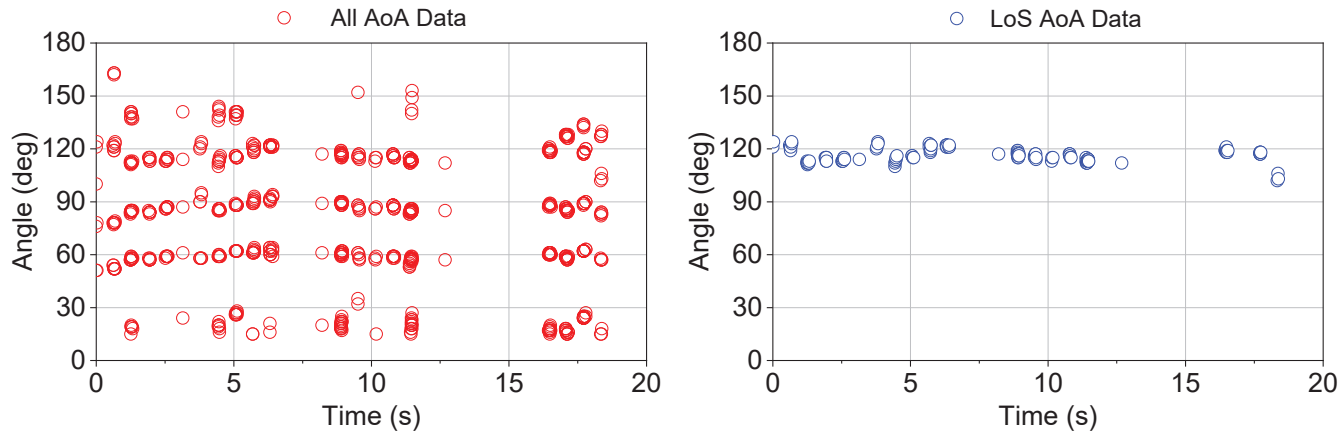
To link the identities of a wireless transmitter to a pedestrian in the video, we match LoS AoAs and pedestrians' angle trajectories to calculate a matching score for each pedestrian. We divide the sequence of LoS AoAs into M equal-sized time windows. Denote the number of LoS AoAs in the k^{th} window as $m_k\,(1 \leq k \leq M, k \in \mathbb{Z}^+)$. Denote the video angle trajectory of a pedestrian as θ, and the angle at the time $t_j\,(1 \leq j \leq T, j \in \mathbb{Z}^+)$ as ϕ_j. Then, we compute the matching score at t_j as $score_j^t = (1 - |\frac{\theta(t_j)}{180} - \frac{\phi_j}{180}|)^3$. Note that the calculation of $score_j^t$ can also use Euclidean distance, Cosine



(a) Response time interval.

(b) Number of packets.

Fig. 4. CDF of the response time interval and the number of packets of APP_1 and APP_2.



(a) Multipath AoA trace.

(b) LoS AoA identified by the Bagged Tree.

Fig. 5. An example of LoS AoA identification using Bagged Tree.

Similarity or their deformation [11], [12], *etc*. The evaluation of different matching score computation methods is left for our future work.

VI. Evaluation

In this section, we report the experiment setup and evaluation results of IDCam.

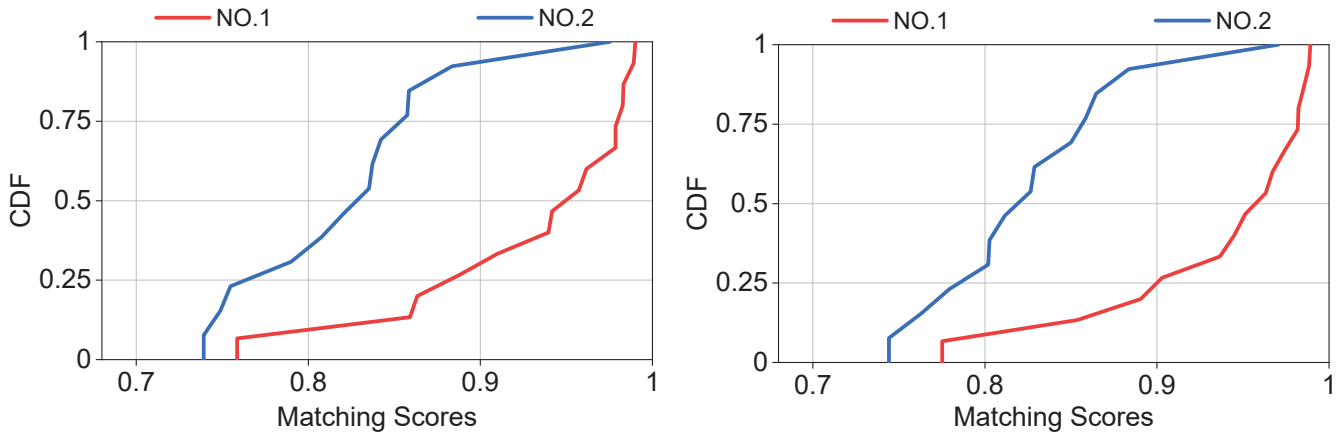
A. Experiment Setup

IDCam is installed on the roof of a five-story building, which allows the camera to monitor a wide area of 2,500 m^2. The surveillance environment is a busy campus site where typically about 50 to 80 pedestrians appear in the video scene during working hours. IDCam consists of an off-the-shelf video camera and a linear 8-element antenna array composed of 8 software radios. Each software radio has two receive chains, in which one chain is used for calibration and synchronization, while the other is used for sniffing. During evaluation, a target carrying a sleep-mode WiFi smartphone walks around in the surveillance site in working hours. The MAC address of the target device is known, which provides the groundtruth for evaluating IDCam performance.

B. Experimental Results

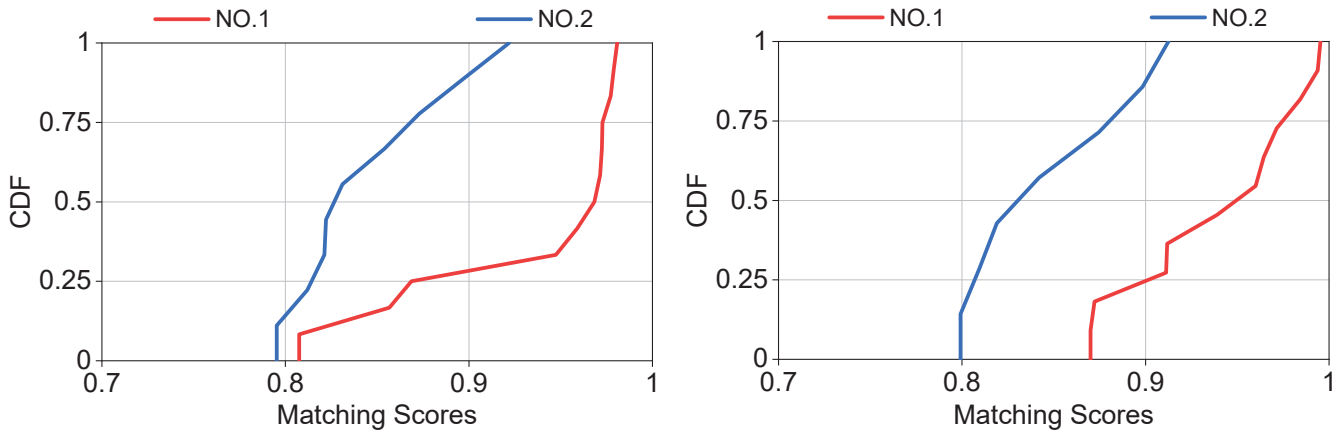
To evaluate the performance of IDCam, we compare the matching scores of top-2 candidates identified by IDCam. We find that the true target has the highest matching score throughout the experiments. The difference of the target's score and the second highest score reflects IDCam's confidence of identity link.

Fig. 6 and Fig. 7 show the matching results after observing one and two bursts of sleep-talk packets of APP_1 and APP_2, respectively. We observe that in both cases the matching score of the true target (the solid red line) is significantly higher than the second highest matching score. Observing two bursts allows IDCam to further improve identification accuracy by about 10%. We note that the performance of IDCam can be further improved if the target's smartphone is installed with



(a) APP_1 (b) APP_2

Fig. 6. CDFs of matching score after one-burst of sleep-talk.



(a) APP_1 (b) APP_2

Fig. 7. CDFs of mathcing score after two-bursts of sleep-talk.

more apps, which will lead to an increased amount of sleep-talk traffic.

VII. Related Work

Pure wireless-based human identification systems like WiFi-ID [13], WFID [14], and Wii [15] analyze WiFi CSI to sense human behaviors like walking style, and then apply machine learning for human identification. However, these systems are sensitive to intensive signal reflections and variations in dynamic environments, therefore suffering poor performance in differentiating multiple pedestrians in the same environment.

Previously, the idea of fusing wireless sensing and computer vision to establish an identity link has been studied in only cooperative settings. For example, IdentityLink [16] and RGB-W [17] fuse computer vision and WiFi packet RSSI to associate a wireless device with people appeared in the video. They require the target device to actively transmit a large number of packets in order to suppress the noise in RSSI measurements. Similar to IDCam, EyeFi [18] fuses computer vision with AoA and uses a neural network to improve AoA estimation accuracy. Different from EyeFi that targets cooperative scenarios and bases on SpotFi [8] for AoA estimation, IDCam exploits only the sleep-talk of smartphones, uses a fast Bartlett-based algorithm for real-time AoA analysis of pedestrians' packets, and engineers features accordingly to enable the application of machine learning.

VIII. Conclusion

This paper presents IDCam, a surveillance system that integrates computer vision and WiFi signal sensing for linking visual and wireless identities. Unlike conventional systems that require target wireless devices to cooperate by actively transmitting packets, IDCam exploits the sleep-talk of smartphone apps and therefore is completely passive. Experiment results show that IDCam achieves accurate identity linking in a noisy multipath wireless environment where 50 to 80 pedestrians

are featured in the video scene. Our results demonstrate the feasibility of augmented security surveillance as well as potential privacy threats that call for further studies.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their thoughtful suggestions. This work is supported by Key-Area Research and Development Program of Guangdong Province 2020B0101390001, National Natural Science Foundation of China (NSFC) (No. U20A20179, 61832001).

REFERENCES

[1] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, Apr. 2018. [Online]. Available: http://arxiv.org/abs/1804.02767
[2] "MS COCO Dataset." [Online]. Available: https://cocodataset.org/
[3] L. Guo, K. Li, Y. Ma, J. Wang, and X. Lian, "Inverse perspective transform based on directional 2-d interpolation," *J. Tsinghua Univ.*, vol. 46, no. 5, pp. 712–715, May. 2006.
[4] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004.
[5] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.
[6] R. Schmidt and R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.
[7] R. Roy and T. Kailath, "Esprit - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 7, pp. 984–995, Jul. 1989.
[8] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "Spotfi: Decimeter level localization using wifi," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 269–282, Aug. 2015.
[9] "Wireshark." [Online]. Available: https://www.wireshark.org/
[10] L. Bbeiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
[11] H. Breu, J. Gil, D. Kirkpatrick, and M. Werman, "Linear time euclidean distance transform algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 5, pp. 529–533, May. 1995.
[12] K. Mikawa, T. Ishida, and M. Goto, "A proposal of extended cosine measure for distance metric learning in text classification," in *Proc. IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Anchorage, United States, Oct. 2011, pp. 1741–1746.
[13] J. Zhang, B. Wei, W. Hu, and S. S. Kanhere, "Wifi-id: Human identification using wifi signal," in *Proc. 2016 International Conference on Distributed Computing in Sensor Systems (DCOSS)*, Washington, DC, United States, May. 2016, pp. 75–82.
[14] F. Hong, X. Wang, Y. Yang, Y. Zong, Y. Zhang, and Z. Guo, "Wfid: Passive device-free human identification using wifi signal," in *Proc. 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous)*, Hiroshima, Japan, Nov. 2016, pp. 47–56.
[15] J. Lv, W. Yang, D. Man, X. Du, M. Yu, and M. Guizani, "Wii: Device-free passive identity identification via wifi signals," in *Proc. 2017 IEEE Global Communications Conference (GLOBECOM)*, Singapore, Singapore, Dec. 2017, pp. 1–6.
[16] L. T. Nguyen, Y. S. Kim, P. Tague, and J. Zhang, "Identitylink: User-device linking through visual and rf-signal cues," in *Proc. 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, Seattle, United States, Sept. 2014, pp. 529–539.
[17] A. Alahi, A. Haque, and L. Fei-Fei, "Rgb-w: When vision meets wireless," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, Dec. 2015, pp. 3289–3297.
[18] S. Fang, T. Islam, S. Munir, and S. Nirjon, "Eyefi: Fast human identification through vision and wifi-based trajectory matching," in *Proc. 2020 International Conference on Distributed Computing in Sensor Systems (DCOSS)*, Virtual Conference, Online, Jun. 2020, pp. 59–68.