IdentityLink: User-Device Linking through Visual and RF-Signal Cues

Le T. Nguyen, Yu Seung Kim, Patrick Tague, Joy Zhang
Carnegie Mellon University
{le.nguyen, yuseung.kim, patrick.tague, joy.zhang}@sv.cmu.edu

ABSTRACT

Mobile devices have become people's indispensable companion, since they allow each individual to be constantly connected with the outside world. In order to keep connected, the devices periodically send out data, which reveal some information about the device owner. Data sent by these devices can be captured by any external observer. Since the observer can observe only the wireless data, the actual person using the device is unknown. In this work, we propose *IdentityLink*, an approach leveraging the captured wireless data and computer vision to infer the user-device links, i.e., inferring which device is carried by which user. Knowing the user-device links opens up new opportunities for applications such as identifying unauthorized personnel in enterprises or finding criminals by law enforcement. By conducting experiments in a realistic scenario, we demonstrate how IdentityLink can be effectively applied to real practice.

Author Keywords

Identity Linking, Computer Vision, RF Sensing

ACM Classification Keywords

C.3 Special-Purpose and Application-Based Systems: Miscellaneous

INTRODUCTION

Mobile devices have become an indispensable companion for our everyday lives. People use them to check email, chat with friends and play games. Many applications running on mobile devices generate traffic even when the device user does not interact with the devices [23]. Applications such as Gmail, Facebook or Skype periodically send and receive background data to synchronize with the cloud. Even the operating system itself generates traffic without user initiation (e.g., to proactively find available Wi-Fi access points).

Owing to the shared nature of the wireless medium, data sent over the air can be captured by any observer. Even though the content of data packets can be encrypted, many approaches have been developed to infer users' personal information such as applications running on the mobile device [6], visited websites [17] or even social links between device owners [5, 3].

Ubicomp '14, September 13 - 17 2014, Seattle, WA, USA

ACM 978-1-4503-2968-2/14/09$15.00.
http://dx.doi.org/10.1145/2632048.2636072

In this work, we propose *IdentityLink*, an approach leveraging the captured wireless data and computer vision to infer user-device links, i.e., inferring which device belongs to which person. The proposed approach identifies user-device links based on users' activities, which can be observed both visually through a camera and wirelessly through a RF (Radio Frequency) signal receiver. Suppose a person carrying a phone walks away from an observer as shown in Figure 1. The observer's camera can detect the increasing distance between the user and the observer, and a wireless receiver can detect the decreasing received signal strength (RSS) of wireless signal from the user's device. Thus, in an environment with multiple users and devices, IdentityLink analyzes the visual and RF-signal patterns to infer user-device pairs.

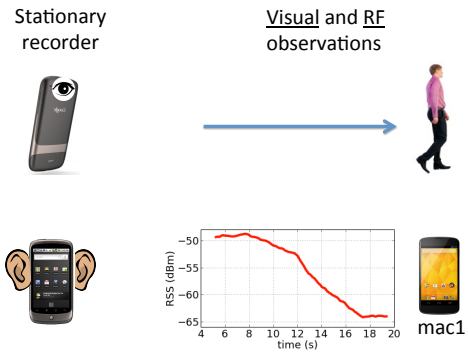


Figure 1: When a person walks away from the recorder, the camera observes an increased distance, and the RF receiver observes a decreased RSS.

Knowing which device belongs to which user opens up new opportunities for applications such as identifying unauthorized personnel in enterprises or tracking criminals by law enforcement. In the aforementioned scenarios, a person's visual identity (captured through a camera) and a device's network identity (captured through a RF receiver) can be combined to infer additional information about a person or a group of interest (e.g., finding people who are socially connected, but come to a certain place at different time of a day).

In such applications, it is essential to enable the linking capability without user's intervention or even recognition. Therefore, the proposed approach infers links by only passively observing people and devices, i.e., we can neither put any additional sensors on people, nor install or modify any application

on their devices. We show the feasibility of IdentityLink by using a single passive observer, which is equipped with video recording and Wi-Fi monitoring capabilities. These capabilities are available even in a single smartphone, allowing easy deployment in an arbitrary environment. Our approach can leverage any available surveillance and Wi-Fi infrastructure.

The goal of this work is to understand how visual and RF signals can be used to infer links between people and devices and what factors influence the linking performance. We summarize the key contributions of this work as follows:

1. **User-Device Linking:** We formalize the user-device linking problem and propose IdentityLink using cameras and RF signals to infer links without user participation.

2. **Prediction Models:** We propose two prediction models for IdentifyLink to match visual and RF patterns using a matching likelihood score. We further show that the two predictors combine to improve overall linking accuracy.

3. **Evaluation in Real-World Settings:** We evaluate the accuracy and limitations of IdentityLink through real-world experimentation. We analyze how factors such as the number of users or amount of RF traffic affect linking accuracy.

USER-DEVICE LINKING

In this section, we define the problem of user-device linking. We then discuss its usefulness in the real world and related solutions for this problem.

Linking Users and Devices

As shown in Figure 2, a person can be identified by a visual identity (e.g., a face captured through a camera) or a device identity (e.g., MAC address captured by a RF receiver). The goal in this work is to infer links between the identities belonging to the same person, i.e., inferring which device identity belongs to which visual identity.

In the following, we assume that there is a one-to-one mapping between a person and a visual identity, i.e., a face we can uniquely identify a person, while a device identity may map to zero or more people. We thus refer to the visual-device identity linking problem as *user-device linking* problem. Moreover, we will sometimes use "person" while referring to his or her visual identity.

Applications

To motivate the value of user-device linking, we describe several applications which can greatly benefit from our approach.

Re-identification: One of the challenges of the vision-based tracking systems is the re-identification problem [7, 25]. The goal of re-identification is to detect whether visual identities appearing on multiple video feeds belong to the same person (e.g., is the person A appearing at 12pm on Monday the same as person B appearing at 1pm on Friday?). Typically, vision-based features such as a face, body shapes or clothing are used to re-identify a person [7]. However, these features are often occluded (e.g., by sunglasses or a cap) or modified (e.g., a person growing a beard), making re-identification

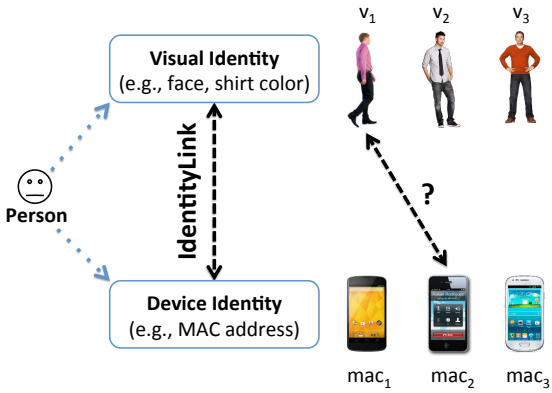


Figure 2: IdentityLink links visual and device identities belong to a person, respectively represented by a visual object v_i captured through a camera and by a MAC address mac_j captured through an RF receiver.

challenging. Instead of relying only on the identifiers extracted through computer vision, we can use the unique identifiers of the mobile device carried by the user such as the device MAC address (as shown in Figure 3). First, we use IdentityLink to infer the device identities of the human subjects visible on the camera. A matching device identity is a good indicator that the two visual identities belong to the same person.

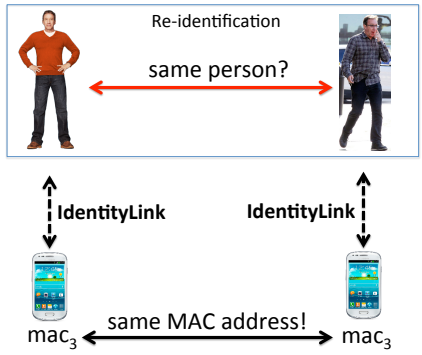


Figure 3: Face of a person can be covered (e.g., through sunglasses) making re-identification challenging. We can use IdentityLink to infer the device identities of the visual subjects and used the inferred identities to identify whether the visual subjects belong to the same person.

Context-aware applications: Camera-based systems can be used to infer a user's context information such as mood, whether the user is alone or with family, what the user is looking at in a store, etc. Using IdentityLink, this contextual information can be delivered to the mobile device of the user. Context-aware applications such as product search, promotions discovery or restaurant recommendations can leverage such contextual information to deliver more accurate results.

Enterprise security: Enterprise networks are often well-protected from the outside but are relatively vulnerable to unauthorized access by insiders [10]. While existing techniques can identify which device is used for unauthorized access, IdentityLink can further identify the person operating

the device, instead of blaming the device owner. Moreover, in cases of device theft, IdentityLink can be used to identify the visual identity of the person who stole the device.

Law enforcement: Modern public safety systems use widely deployed surveillance cameras to detect criminal activities such as vandalism and theft. However, criminals often cover their faces to avoid identification. Mobile devices carried by the criminals may expose a significant amount of information about them such as their affiliation (e.g., school, work place), places they frequently visit (e.g., restaurants, hotels), and their social relationships [18, 5]. IdentityLink can be used to identify the device carried by a criminal and provide law enforcement agents with this additional information.

Related Work

By using multiple sensors one can combine the advantages of each sensing modality. Vision-based sensing has many advantages, since it allows passively tracking users' fine-grained location, activities and interactions in the environment [14, 19, 4]. Due to the challenges with re-identification of human subjects across video feeds, Schulz et al. [22] proposed using infrared ID-batches worn by users and deployment of infrared receivers in the environment. The identity of visual subjects is inferred by tracking a person in a certain region both visually and through the infrared sensing.

Since deployment of additional infrastructure is not practical, Teixeira et al. [24, 25] proposed using the accelerometer and magnetometer sensors on the phone for re-identification. They correlate the user's movements captured through both a camera and mobile sensors to identify the user-device links. This approach assumes that the user has installed an application on the mobile device to report the sensor readings.

In this work, we relax the mentioned assumptions by opportunistically leveraging wireless signals sent out from the user's mobile devices to link the users with their devices. Thus, our approach can address the re-identification problem *without having users carry any special hardware or requiring them to install an application on their mobile devices*. In so doing, our approach significantly reduces the user's effort (e.g., a user can use the context-aware search without installation of any additional application). Additionally, our approach can be applied to application scenarios where one cannot assume cooperativeness of the mobile device users (e.g., the mentioned law enforcement use cases).

IDENTITYLINK

In this section, we formally define the concept of user-device linking and then give an overview of the IdentityLink system, including the details of each system component.

User-Device Linking Problem

The user-device linking problem is based on the visual identities $v_1, \cdots, v_n$ and device identities (MAC addresses) $mac_1, \cdots, mac_m$ observed by the camera and RF receiver.

We let $P(v_i)$ and $P(mac_j)$ denote the person associated with the visual and device identities. If $P(v_i) = P(mac_j)$, then v_i and mac_j belong to the same person.

Let $S(v_i, mac_j)$ be a score function, which computes a likelihood of v_i and mac_j belonging to the same person. We frame the user-device association problem as follows: given a visual identity v_i, we want to find a device identity mac^* so that $P(v_i) = P(mac^*)$. This corresponds to finding mac^*, which has the highest score for a given v_i:

$$mac^* = \arg\max_{mac_j} S(v_i, mac_j). \quad (1)$$

Therefore, for each v_i we go through each mac_j and compute a score. Then we choose mac_j with the highest score and assign it to v_i, as illustrated in Figure 4.

The above problem statement assumes that each v_i is associated with exactly one mac_j. However, there are situations when v_i is not associated with any device identity (i.e., the person does not carry a phone). To address this case, we can use a thresholding approach, i.e., we assign a mac^* to v_i only if $S(v_i, mac^*)$ is greater than a certain threshold. We thus eliminate situations of linking a visual and a device identity which are not likely to belong to the same person.

Moreover, there are cases when v_i is associated with more than one device identity (i.e., the person carries more than one phone). To address this case, we can leverage existing techniques for detecting co-moving devices [3]. We first link the device mac^* with the highest score for v_i, then link all devices co-moving with mac^* to v_i as well.

The key to solving the user-device linking problem is finding an appropriate score function to compute how likely v_i and mac_j belong to the same person. In the following section, we propose two score functions to infer the user-device links.

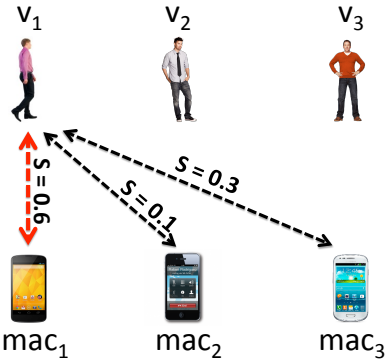


Figure 4: Given a visual identity, we compute a score for each device identity and select the one with the highest score.

System Overview

Figure 5 depicts the overview of the proposed system, and Figure 6 shows how the two different signal sources (video and RF) are processed in each stage. The system starts by recording a video and capturing wireless data. The recorded video is then processed to infer the movement trajectory of each person. In parallel, the recorder captures wireless data packets and converts them into RSS timeseries. User trajectories and RSS timeseries are then input into a predictor, making sure to keep the video and RF timestamps synchronized.

In this work, we propose two predictors: a motion-based and a distance-based predictor. Each predictor is composed of two stages: feature extraction and score computation. In the first stage a predictor transforms the input into *visual features* and *RF features*. For example, the motion-based predictor extracts the motion information from each trajectory by inferring when a person moved and when he or she was stationary. Thus, the visual feature corresponds to the binary timeseries where 0 indicates no movement and 1 indicates non-trivial movement. Similarly, the predictor infers the RF motion features from the RSS timeseries. In the second stage, the predictor computes a score for each pair of visual and RF features, yielding a score matrix. The score matrices of both predictors are passed into the link inference component, which determines the user-device links.

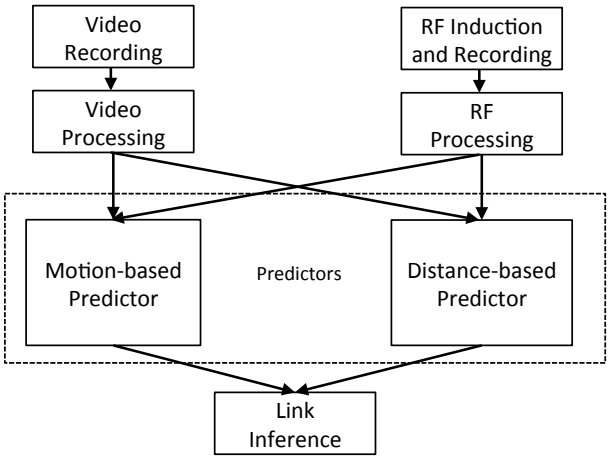


Figure 5: The system components of IdentityLink: two signal sources (video and RF) are processed to infer links between people and devices.

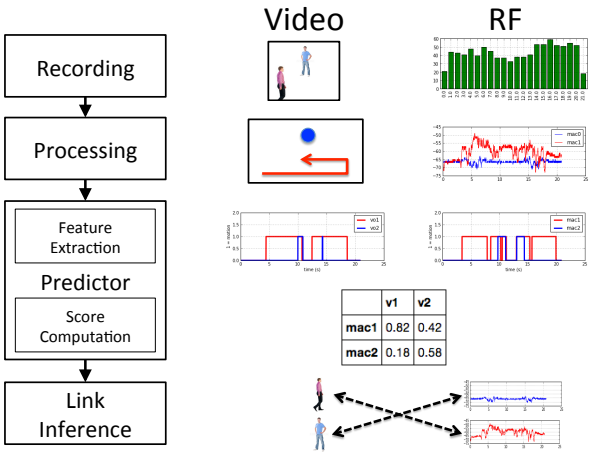


Figure 6: Video and RF signals are processed at each state of IdentityLink.

Video Recording and Processing

Video is recorded using a stationary camera. The goal of the video processing component is to infer the location of human objects detected on the video feed. Video processing is divided into three steps: *human segmentation*, *tracking* and *trajectory inference* as shown in the Figure 7.

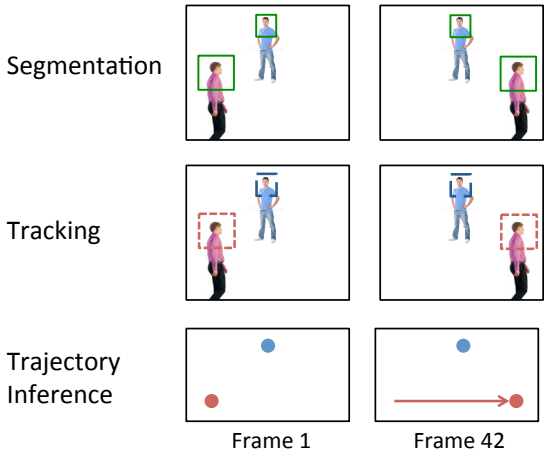


Figure 7: The three steps of video processing include human segmentation, tracking and trajectory inference.

In the first step, we identify *human objects* (shown as a green solid rectangle) in each video frame. We use the approach of Kruppa et al. [15] to detect human objects based on upper body shapes such as head and shoulders. This approach even detects partially hidden people or those not facing the camera.

In the second step, we identify human objects, which appear on multiple frames and belong to the same person. For simplicity, we use an appearance model introduced by Bird et al. [2], which assigns human objects with the same clothing color to the same visual identity. For example, whenever we see a person in pink shirt, we assume that is the same person and assign the human object to the visual identity v_1. More robust techniques have been developed to identify human identities such as using human face or body shapes [12]. Due to the complexity of the implementation we decided to integrate their functionality in future work, as the specific computer vision techniques are not key to our contributions.

In the last video processing stage, we infer the human subject trajectories. Many advanced techniques have been developed to infer a 3D trajectory from image sequences captured from an ordinary camera [20, 21, 16]. For simplicity, we leverage an output of the 3D camera (Kinect) to estimate people's trajectories [27], expecting that advanced 3D trajectory inference techniques would yield similar results.

As mentioned, our approach would benefit from more sophisticated and robust computer vision algorithms [12]. However, our main focus is the proposed user-device linking approach. Therefore, the mentioned computer vision improvements are out of the scope and will be explored in the future work.

RF Induction, Recording and Processing

Any Wi-Fi device, such as an AP, laptop or smartphone, can capture Wi-Fi traffic sent over the air. In this work, we show the feasibility of IdentityLink using a Nexus One smartphone

as an RF receiver. We configure the device to capture all Wi-Fi packets [8]. Using *tcpdump* [11], we record the time of arrival, sender MAC address and the RSS value for each captured packet. By grouping data packets from the same sender, we obtain one RSS timeseries per sender.

It is necessary to capture a sufficient amount of RSS samples (around 10 samples per second) in order to infer user-device links. However, it is not guaranteed that the tracked devices will always generate this amount of wireless data. Thus we highlight two traffic induction techniques to increase the amount of data generated by the tracked device.

The first technique assumes that the tracked devices are connected to a certain Wi-Fi network and the recorder has access to this network (e.g., enterprise Wi-Fi network). First, we obtain the MAC and IP address of surrounding devices by simple eavesdropping or by broadcasting ICMP Echo Request (ping) messages [1] and listening for replies, each providing useful measurement data for our inference problem. Depending on how much information is available for certain device IP addresses, we can send more or fewer requests to specific addresses as needed. Empirically, we were able to persuade user devices to provide useful measurements at a rate of over 50 samples per second using this technique.

The second approach extends on the first to include cases where a target device is either not connected to any network or connected to a network the recorder cannot gain access to. In this case, the recorder can take a more aggressive approach by forcing the target devices to connect to its own network, relying on an approach known as a Karma attack [26]. Wi-Fi client software is configured to actively search for previously used APs, using control messages known as probe requests containing the SSIDs of preferred APs. After passively observing probe requests from target devices, the recorder can advertise a fake AP copying one of the target's SSIDs. As long as the fake AP transmits with a strong signal, the target devices would automatically connect to the fake AP, even if the SSID is the same as another nearby AP. Once the target device is connected, we can use the previously described method. Note that this method may be prohibited by law in certain circumstances. We do not discuss further details on the related regulations in various situations.

Motion-based Predictor

We first introduce a simple scenario to illustrate the motion-based predictor. Figure 8 shows the Scenario L-R (Left-Right) with three people carrying mobile devices, two (P_2 and P_3) stationary and one (P_1) walking from left to right, pausing for a few seconds, then walking back to the left.

For clarity in the following, we rearrange the subscripts so v_i and mac_i are the visual and device identities of person P_i. For an easier understanding, we use the same color and style to plot timeseries belonging to the same person.

The motion-based predictor is built on the idea of detecting users' movements from the video and the RSS stream. Consider the moving person in the Scenario L-R. At any given time, if the person moves, the movement can be detected by observing the changes in location from the video. At the same time, the RF receiver will observe a significant RSS fluctuation from the mobile device carried by the moving person. This fluctuation is caused by the device changing its location and angle with respect to the receiver [3, 13]. Since the video reveals when the person started and stopped moving, we try to find a device with RSS fluctuating at the corresponding time period which we refer to as the *motion period*.

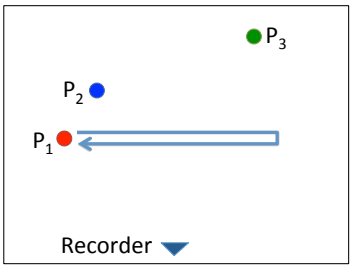


Figure 8: Scenario L-R: Trajectory of three people, two stationary. One walks in front from left to right, pauses for a few seconds, then walks back to the left.

The prediction process consists of 1) feature extraction and 2) similarity computation. Feature extraction is further devided into two parts: 1a) motion observed in the video and 1b) motion inferred from the RSS streams. The feature extraction outputs are visual and RF features, which are then input into the similarity computation component to compute a score for each visual-RF feature pair.

Motion Detection from User Trajectories

To detect whether a user is moving or not, the system uses trajectories inferred by a video processing component. First, the system computes a user's speed s_t at time t as

$$s_t = \|(x_{t-1}, y_{t-1}) - (x_t, y_t)\|_2 \quad (2)$$

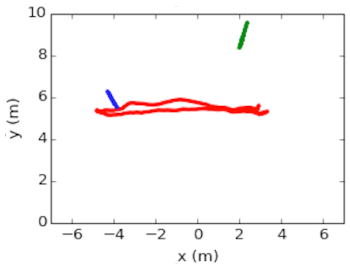
where $\|\cdot\|_2$ is Euclidean distance, x and y are coordinates of the video object measured in meters and t is time in seconds.

Figure 9a shows trajectories of each visual identity v_i extracted by the video processing component. The trajectories are used to compute speed shown in Figure 9b. A user is moving if the speed crosses a certain threshold. Figure 9c shows the detected motion using an empirical threshold of 0.5.

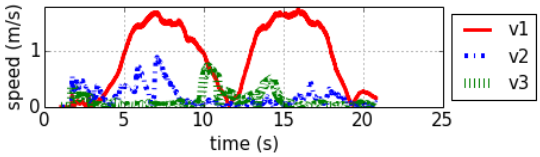
From the figure, we observe that the motion period of v_1 starts at around the 4th second; the person moves for 6 seconds, pauses for a while and then moves for another 6 seconds. We also observe false positive of the motion detection for v_2 and v_3, which are in reality stationary. This is caused by the noise and inaccuracies of video processing.

Motion Detection from RSS Stream

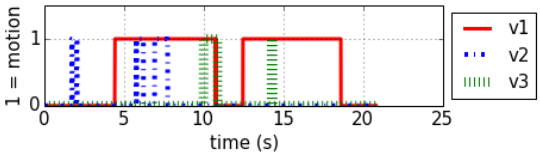
We frame motion detection from the observed RSS stream as a machine learning problem, specifically as a binary classification problem, where we try to predict one of the two classes, "moving" or "not-moving". To train the model, we use common statistical features for motion classification [9] such as RSS variance, minimum, maximum, range (i.e., maximum - mininmum) and coefficient of variation extracted



(a) **Users' trajectories** are estimated from the video feed.



(b) **Speed** is inferred from users' trajectories.



(c) The **visual motion feature** is inferred by thresholding the user's speed.

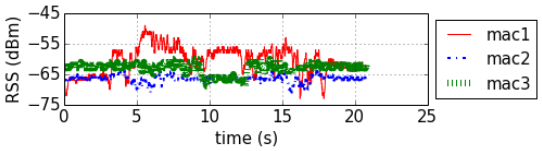
Figure 9: The speed and motion features are extracted from user trajectories. Due to the noise in video processing, v_2 and v_3 are estimated to be moving from time to time, even through they are stationary.

from a 2 second sliding window. Moreover, we extract frequency domain features by computing the spectral density of the signal and then averaging over bands of interest [9].

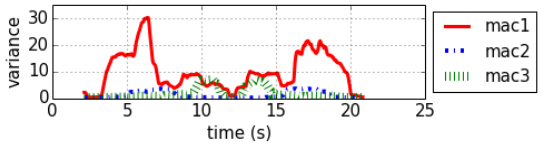
Figure 10a shows the RSS timeseries collected for each of the three devices. Figure 10b shows the RSS variance feature computed using a sliding window with a size of 2 seconds. Variance of mac_1 increases as the person starts moving and decreases when the person pauses. This observation is consistent with findings from the previous work, indicating the RSS fluctuation caused by the human motion [13].

The output of the prediction is shown in Figure 10c, where 1 indicates that a motion was detected for a certain device at a given time. Similar to the visual case, we observe that the motion period of device mac_1 starts at around the 4th second; the device moved for approximately 6 seconds, paused and then moved for another 6 seconds.

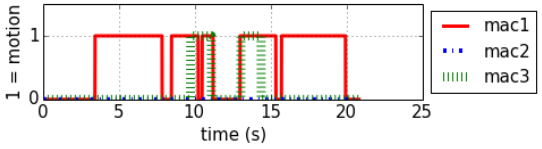
False positives and false negatives of the prediction shown in the figure are caused by the fact that when the person moves, the wireless properties of the environment are changed. This causes RSS fluctuation not only for the moving person's device, but also for devices nearby. In the Scenario L-R, P_1 crosses the line-of-sight (LOS) between P_3 and the recorder twice. Whenever P_1 crosses the LOS, we observe an increase of RSS variance for the device mac_3 (shown in Figure 10b). This increase can be falsely interpreted as motion, even though the device is stationary at all times.



(a) **RSS** measurements are observed from the three devices.



(b) **RSS Variance** is computed using a sliding window with a size of 2 s.



(c) **RF motion features** are inferred by using the machine learning model.

Figure 10: We compute various features such as variance from the raw RSS values and use these features to infer device motion.

Score Computation

To infer links, we compute a score for each pair of visual and RF motion features using the score function

$$S_M(v_i, mac_j) = \frac{1}{T}\sum_{t=0}^{T} F_{M,v}(v_i)_t \cdot F_{M,m}(mac_i)_t, \quad (3)$$

where $F_{M,v}(v_i)$ and $F_{M,m}(mac_j)$ are visual and RF-based motion features. The score reflects the time-averaged inner product of the timeseries, capturing the correlation between motion features. Table 1 shows the score matrix for the Scenario L-R. For each visual identity v_i we compute a score for each device identity mac_j. From the matrix we can observe that the pair (mac_1, v_1) has a high score since their motion feature timeseries have a high amount of overlap.

	v_1	v_2	v_3
mac_1	**0.51**	0.06	0.04
mac_2	0.00	0.00	0.00
mac_3	0.12	0.00	0.05

Table 1: Pairwise similarity of the motion indicators is used for user-device linking.

Distance-based Predictor

The distance-based predictor is based on the inverse relationship between distance and RSS illustrated in Figure 1. Intuitively, motion toward and away from the recorder will result in respective increase or decrease of observed RSS on average. In the case where one person walks in random directions, Figure 11 shows the inverse proportionality of the distance measure compared to the observed RSS. From this figure, we

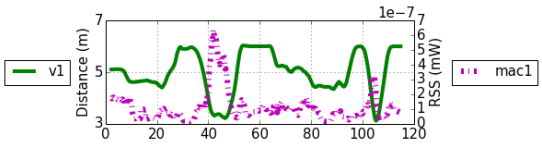


Figure 11: When a person walks toward or away from the recorder, the estimated distance and observed RSS vary accordingly.

observe measurements consistent with the expected inverse relationship between distance and RSS. We leverage this inverse proportionality to define a window-based score function corresponding to the covariance

$$S'(v_i, mac_j)_k = \frac{1}{T}\sum_{t=k}^{k+w}(F_{D,v}(v_i)_t - \overline{F_{D,v}(v_i)}) \cdot (F_{D,m}(mac_j)_t - \overline{F_{D,m}(mac_j)}) \quad (4)$$

where $F_{D,v}(v_i)$ and $F_{D,m}(mac_j)$ are the distance and square root of RSS values over time and $\overline{F_{D,v}(v_i)}$ and $\overline{F_{D,m}(mac_j)}$ are their mean values. The size w of a sliding window is empirically set to 2 seconds. The final score is computed by summing all the negative covariance values over the sliding windows and then negating the sum to obtain a positive score

$$S_D(v_i, mac_j) = -\sum_k min(S'(v_i, mac_j)_k, 0). \quad (5)$$

Note that instead of summing over all the values, we sum up only the negative covariance values (and ignore the positive values). We have empirically observed that the negative covariance values are good indicators that the RSS stream and the distance stream belong to the same person moving towards/away from recorder. However, the positive values are typically caused by the fluctuations of the RSS.

User-Device Link Inference

The link inference component uses the output score matrices of the motion- and distance-based predictors to infer user-device links. Each column of the matrix corresponds to a score vector of one visual identity v_i, and the entries in this vector are similarity scores with each device identity mac_j. Link inference makes a joint prediction using both matrices.

Given the motion-based score matrix S_M and distance-based score matrix S_D, we create a combined score matrix S through normalization and linear combination. Each column S^i_M and S^i_D of the matrices S_M and S_D is corresponding to visual identity v_i, is normalized to have unit sum, and the normalized columns are combined as

$$S^i = \alpha S^i_M + (1-\alpha)S^i_D \quad (6)$$

where $0 \leq \alpha \leq 1$ is a weighting factor between the predictors. We discuss a strategy for selection of α later in the evaluation section. Table 2 provides an example of predictor combination for a visual identity v_1.

	S^1_M	S^1_D	S^1
mac_1	0.37	**0.51**	**0.44**
mac_2	**0.39**	0.21	0.30
mac_3	0.24	0.28	0.26

Table 2: S^1 is obtained by combining the motion-based score vector S^1_M and distance-based score vector S^1_D with $\alpha = 0.5$.

The device identity mac^* with the highest score is then linked with the visual identity v_i:

$$mac^* = \arg\max_{mac_j} S^{i,mac_j}, \quad (7)$$

where S^{i,mac_j} is the value at column v_i and row mac_j of the combined score matrix S.

From Table 2, we observe that the predictors individually come to different conclusions. Motion-based predictor would link v_1 with mac_2, whereas the distance-based predictor would link it to mac_1. In combining the scores, we can consider the confidence of the individual predictors.

EVALUATION

Linking performance depends on many factors such as the number of users, motion patterns, and the amount of RF signals observed. In the following, we conduct experiments to analyze how these factors influence the linking accuracy.

We conduct two sets of experiments: small-scale controlled experiments and larger scale experiments with random movements. In the controlled experiments, we analyze the linking performance when users perform elementary movement patterns. This provides an understanding of the strengths and weaknesses of our approach under different conditions. In the second set of experiments, we show how IdentityLink performs in a real-world scenario, with unscripted movements.

Figure 12 shows the layout of the evaluation environment, which is comprised of an 11.5 x 5 meter room and a 1.7 meter wide hallway. The field of view of the camera covers an area of around 5 x 5 meters. This corresponds to a real-world use case in which the camera can capture the motion only in an limited area, whereas the RF receiver can capture the signals from all devices in range, including those not in direct line-of-sight (e.g., in the hallway).

Our experiments involve 10 subjects, each carrying a mobile device in a pocket and initially sitting in one of the 10 chairs in the room. In the following scenarios, a subset of users is asked to stand up and perform different motion patterns. We then evaluate the linking accuracy for the moving subjects.

Experiments with Elementary Motion Patterns

We first consider scenarios with only two or three moving subjects. These subjects are asked to perform elementary movement patterns, which are building blocks for more complex movement patterns in real-world use cases. Our results show which movement patterns provide sufficient information to perform user-device linking. We evaluate the linking performance of each predictor individually to understand its strengths and weaknesses.

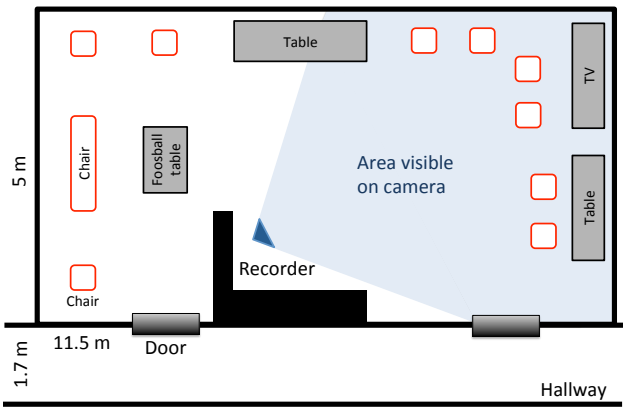


Figure 12: Layout of the test environment.

Motion-based Prediction

The motion-based predictor performs linking based on the motion periods observed visually and through RF sensing. Intuitively, when motion periods of visual and device identities match (both identities moved between time t_1 and t_2), the two identities likely belong to the same person. However, with multiple moving subjects, the linking process becomes more challenging since motion periods of multiple subjects and devices can be time-overlapping. In the following, we analyze basic scenarios of multiple subjects moving with a certain degree of time-overlap.

Figure 13 shows four scenarios of users moving in front of the recorder at different time periods. Scenario 1A describes the case when subjects move at different times (e.g., one person leave, then another enters later). Scenario 1B shows a partial overlap of the motion periods. Scenario 1C shows a complete time-overlap (e.g., two people walking side-by-side). Finally, scenario 1D shows three subjects with at least two moving at any time.

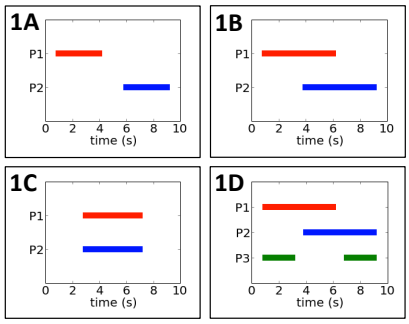


Figure 13: Four scenarios of users moving at different times. The bar indicates the motion period of each person.

We collect video and RF signal data for the scenarios in Figure 13 from the 10 devices, repeating each scenario 50 times.

For each moving visual identity detected in a video, we infer its device identity using Equation (7). For example, we detect two moving visual identities in each video of scenario 1A, noting the remaining people are stationary or out of view, and infer a total of 100 links. The linking accuracy corresponds to the percentage of correctly inferred links out of the total attempts. Since the RF receiver can always observe 10 devices, each a candidate match to the visual identity, the expected linking accuracy of a random guess is 10%.

	1A	1B	1C	1D
Motion-based	95%	95%	52%	87%

Table 3: Linking accuracy of the motion-based predictor. The predictor achieves high accuracy if the motion period of one user is distinguishable from that of others.

The linking accuracy of the motion-based predictor is shown in Table 3. The predictor achieves high accuracy for scenarios 1A and 1B, since the motion periods for the two subjects are distinguishable, i.e., there are times when only one person moves. These time periods are key for correct inference.

In scenario 1C, the two subjects are moving at exactly the same time. Therefore, the motion-based predictor has insufficient information to differentiate. In such cases, a moving subject will be linked with a moving device, but since there are two moving devices with similar patterns, the system will randomly choose one. Note, however, that the motion-based predictor chooses randomly from a pool of only two moving devices, eliminating the stationary devices as candidates.

The results for scenario 1D show that the motion-based predictor performs well even when two people move simultaneously at any time. The key requirement for the predictor is that the motion period of one person has to be differentiable from others, providing sufficient evidence for the correct linking. However, we also notice a decrease in linking accuracy in 1D as compared to 1B. These two scenarios are similar with the only difference of having one additional moving subject in 1D. This additional moving subject causes additional RSS fluctuations in the environment, increasing false positives and thus decreasing linking accuracy. We later evaluate how a greater number of moving subjects impacts the linking accuracy.

Distance-based Prediction

As shown in the previous experiment, the motion-based predictor poorly handles cases with indistinguishable motion. In the following, we present multiple scenarios with simultaneously moving users and analyze how the distance-based predictors can handle such cases.

The effectiveness of the distance-based predictor depends highly on the motion trajectory of a user. Figure 14 shows four scenarios with two users moving simultaneously in different trajectories. In scenario 2A the users walk in opposite directions toward or away from the recorder, whereas in 2B the users walk in the same direction. In 2C one of the users moves horizontally with respect to the recorder, keeping roughly the same distance to the recorder. The scenario 2D is a more complicated trajectory, essentially two combined instances of 2C.

The linking accuracy of all four scenarios is shown in Table 3. Since in each scenario both subjects move simultaneously (as

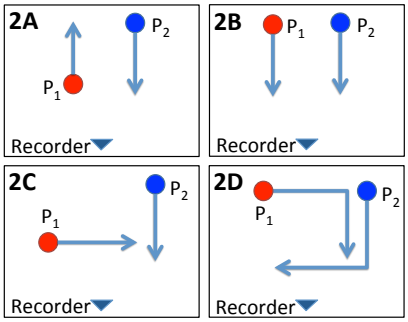


Figure 14: Four scenarios of two users moving simultaneously in different trajectory shapes.

in scenario 1C), the motion-based predictor has insufficient information for the correct inference, thus achieving an average accuracy of 50%.

The distance-based predictor performs well in scenarios where users' distance profiles are distinct, such as in scenario 2A. However, if both users have similar RSS patterns such as in 2B, the distance-based predictor will not have sufficient information to differentiate between them. Thus, neither the motion- nor distance-based predictors can distinguish between the two users in this case, resulting in a random matching between the two devices.

In scenario 2C, the accuracy is different for the two subjects. The predictor performs well for the person P_2, who moves towards the recorder, but person P_1 often cannot be distinguished from the remaining stationary subjects. Scenario 2D, however, leveraged distance information for both subjects to infer links with reasonable accuracy.

	2A	2B	2C	2D
Motion-based	48%	52%	51%	50%
Distance-based	81%	48%	P_1:17%, P_2:84%	75%

Table 4: Linking accuracy of the distance-based predictor.

We observed that the accuracy of the distance-based predictor depends highly on the users' motion patterns. The predictor can correctly link a person to a device if the person walks toward or away from the recorder for at least part of the motion period. However, if the person walks around the recorder keeping a near-constant distance, the predictor will not have sufficient information. Hence, the placement of the recorder is important for practical consideration (e.g., above a door where people enter and exit). Optimal recorder placement is beyond our scope of work and will not be discussed further.

We observe scenarios in which multiple users cannot be distinguished using IdentityLink, including when two users walk side by side. However, any unique movements (e.g., two users walk side by side, then one turns) provide subtle differences for IdentityLink to make the correct decision.

Experiments with Random Motion Patterns

Our second set of experiments focuses on how strengths and weaknesses of linking elementary movements combine in real-world scenarios. In the following, we analyze the performance of IdentityLink under real-world conditions with a variable number of users moving randomly.

We again use the room shown in Figure 12 with 10 subjects carrying devices and initially sitting in chairs in the room. We ask five of the subjects to remain seated while five move around the room and hallway. Our primary instruction to users was to occasionally exit the room and re-enter through the door on the right. We refer to the interval between entering and exiting the room as a user's *session*. Within a session, which had no specified duration, subjects could sit on chairs, get food from the table, socialize with others, or wander around the room. We asked each subject to participate in 25 sessions, waiting a random amount of time in the hallway between each session.

In the collected dataset, we observe high variance of session duration ranging from 5 seconds (subject enters, takes food from the table, then exits) up to 2 minutes (subject enters, sits on a chair and socializes with other subjects). The average session duration was approximately 23 seconds. Users are not always visible to the camera during sessions, due to incomplete coverage and possible occlusion.

We first consider the case where visual identities are not explicitly linked across sessions, i.e., each person creates 25 independent sessions. With five moving subjects, the data represents 125 user sessions with at most 10 people in the room at any time. In this case, we infer links using only the motion- and distance-based predictors within sessions. Table 5 shows the results of this prediction, noting a 10% baseline with 10 people in the room.

	Motion-based	Distance-based
Accuracy	53.9%	47.6%

Table 5: We highlight the accuracy achieved when using only the motion-based or distance-based predictor.

Figure 15 shows the accuracy achieved when combining the results of both predictors using Equations (6) and (7). Setting $\alpha = 0$ corresponds to using only the distance-based predictor, whereas $\alpha = 1$ corresponds to using only the motion-based predictor. We achieve the best results when setting $\alpha = 0.8$, which we employ in the following experiments.

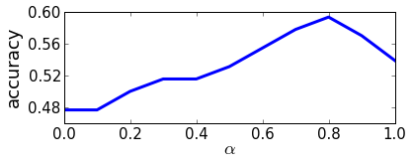


Figure 15: We illustrate the accuracy achieved by combining the motion-based and distance-based predictors through a linear combination of their scores.

Number of Moving Subjects

We next study how the number of moving subjects influences the linking accuracy. We previously considered five stationary and five moving subjects. In the following experiment, we reduce the number of moving subjects and recompute the

linking accuracy. For five stationary devices and a variable number of moving devices, the linking accuracy is shown in Figure 16.

Even though the movement of one person results in RSS fluctuations in all the devices in the vicinity, the motion-based predictor achieves 99% linking accuracy. On the other hand, the distance-based predictor achieves a lower accuracy, since its performance highly depends on the motion pattern. Since in some sessions a test subject walks through the area visible on camera while keeping a constant distance to the recorder, the distance-based predictor does not always have sufficient information for the linking.

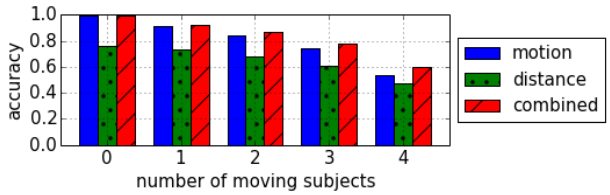


Figure 16: The linking accuracy decreases with increasing number of moving subjects.

We observe that linking accuracy decreases with the number of moving subjects, as expected. Moreover, we empirically observed that varying the number of stationary subjects has no impact on the linking accuracy, i.e., IdentityLink achieved the same accuracy when removing all 5 stationary subjects from the dataset. Thus, the linking accuracy depends only on the number of devices carried by moving subjects, independent of the number of stationary mobile phones, laptops or desktops in the environment.

Observing One Subject in Multiple Sessions

We next consider the case where visual identities are linked across sessions, i.e., the same person is observed repeatedly over time. This corresponds to a real-world scenario of capturing the same person using multiple recorders across rooms or using one recorder across multiple sessions. Figure 17 shows the accuracy achieved when we link visual identities across sessions. Repeated observation of the same person significantly increases the linking accuracy, e.g., achieving accuracy of 85% from three sessions.Moreover, in a real-world scenario, it is likely that different devices would be present in different sessions, where as our results correspond to the same 10 devices across sessions. Hence, higher accuracy would be expected by eliminating devices that do not appear in all datasets containing the person of interest.

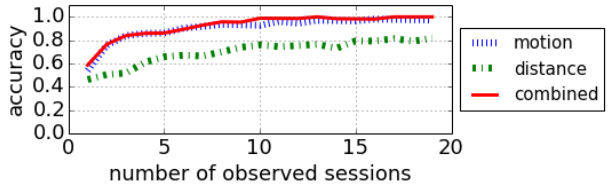


Figure 17: The linking accuracy increases significantly when the same person is observed multiple times.

Sufficiency of Wireless Data

To infer user-device links we assume that the tracked devices generate sufficient RF signal data over time. In this work, we described two techniques to induce additional network traffic. In what follows, we analyze how the RF data rate affects linking accuracy.

In our base experiment, each device to generated approximately 50 packets per second. To study the effect of less traffic, we downsample the measurements and re-compute the linking accuracy for different data rates. From the results shown in Figure 18, we observe that data rates in excess of about 10 packets per second are sufficient to achieve similar performance. However, with fewer than 3 packets per second, the linking accuracy decreases significantly. Based on this observation, the recorder can scale how aggressively it must induce traffic to maintain the desired accuracy.

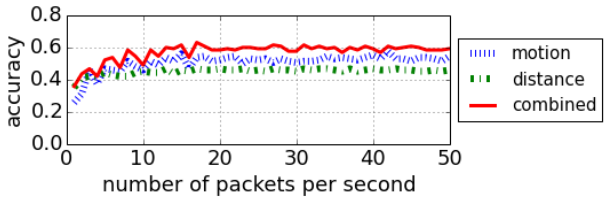


Figure 18: The linking accuracy is similar for anything more than 10 packets per second. Fewer than 3 packets per second appears insufficient for linking.

CONCLUSIONS AND FUTURE WORK

In this work, we proposed IdentityLink to link users to mobile devices using video and RF signal data with no explicit participation or app installation required by the observed users. We studied the feasibility and the accuracy of our IdentityLink approach in structured and randomized user scenarios with a variety of system parameters, showing IdentityLink can achieive high accuracy with repeated user observations even with partial camera coverage. Through our experiments, we demonstrated the feasibility of IdentityLink with up to 10 users, primarily limited by the need for more robust computer vision techniques beyond the scope of this work. IdentityLink relies on relative changes of the visual and RF signals caused by user mobility. Therefore, user-device links can be inferred for moving subjects. However, we have empirically observed that movement of one person affects signals from nearby devices. Future work will explore how IdentityLink could exploit this dependence to additionally link stationary users. Moreover, explicit location information could be integrated as a third component of IdentityLink to further increase accuracy. IdentityLink succeeds by leveraging distinguishable movement patterns of tracked subjects. In some cases, however, movement patterns of multiple people are indistinguishable, e.g., people moving as a group. In such cases, IdentityLink is limited to inferring links between groups of people and groups of devices. Moreover, the placement of the recorder in a room plays an important role, so future work could further investigate optimal recorder placement.

ACKNOWLEDGEMENT

This research was supported in part by National Science Foundation award IIS-1344768.

REFERENCES

1. Almquist, P. Type of service in the internet protocol suite.
2. Bird, N. D., Masoud, O., Papanikolopoulos, N. P., and Isaacs, A. Detection of loitering individuals in public transportation areas. *IEEE Trans. Intelligent Transportation Systems 6*, 2 (2005), 167–177.
3. Chandrasekaran, G., Ergin, M. A., Gruteser, M., Martin, R., Yang, J., and Chen, Y. Decode: Detecting co-moving wireless devices. In *5th Intl. Conference on Mobile Ad Hoc and Sensor Systems*, IEEE (2008), 315–320.
4. Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., and Murino, V. Social interaction discovery by statistical analysis of f-formations. In *BMVC* (2011), 1–12.
5. Cunche, M., Kaafar, M.-A., and Boreli, R. Linking wireless devices using information contained in wi-fi probe requests. *Pervasive and Mobile Computing* (2013).
6. Dai, S., Tongaonkar, A., Wang, X., Nucci, A., and Song, D. Networkprofiler: Towards automatic fingerprinting of android apps. In *32nd IEEE Intl. Conference on Computer Communications (INFOCOM)* (2013).
7. Doretto, G., Sebastian, T., Tu, P., and Rittscher, J. Appearance-based person reidentification in camera networks: problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing 2*, 2 (2011), 127–151.
8. Feinstein, R. Monitor mode for broadcom wifi chipsets, Sept. 2012. `http://bcmon.blogspot.com/2012/09/working-monitor-mode-on-nexus-one.html`.
9. Huynh, T., and Schiele, B. Analyzing features for activity recognition. In *Joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*, ACM (2005), 159–163.
10. IBM Software. Avoiding insider threats to enterprise security, Oct. 2012. `http://public.dhe.ibm.com/common/ssi/ecm/en/wgw03016usen/WGW03016USEN.PDF`.
11. Jacobson, V., Leres, C., and McCanne, S. The tcpdump manual page. *Lawrence Berkeley Laboratory, Berkeley, CA* (1989).
12. Ke, S.-R., Thuc, H. L. U., Lee, Y.-J., Hwang, J.-N., Yoo, J.-H., and Choi, K.-H. A review on video-based human activity recognition. *Computers 2*, 2 (2013), 88–131.
13. Kleisouris, K., Firner, B., Howard, R., Zhang, Y., and Martin, R. P. Detecting intra-room mobility with signal strength descriptors. In *11th ACM international symposium on Mobile ad hoc networking and computing*, ACM (2010), 71–80.
14. Koyuncu, H., and Yang, S. H. A survey of indoor positioning and object locating systems. *IJCSNS International Journal of Computer Science and Network Security 10*, 5 (2010), 121–128.
15. Kruppa, H., Castrillon-Santana, M., and Schiele, B. Fast and robust face finding via local context. In *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)* (2003), 157–164.
16. Lepetit, V., and Fua, P. *Monocular model-based 3D tracking of rigid objects*. Now Publishers Inc, 2005.
17. Liberatore, M., and Levine, B. N. Inferring the source of encrypted http connections. In *13th Conference on Computer and Communications Security*, ACM (2006), 255–263.
18. Lindqvist, J., Aura, T., Danezis, G., Koponen, T., Myllyniemi, A., Mäki, J., and Roe, M. Privacy-preserving 802.11 access-point discovery. In *2nd ACM conference on Wireless network security*, ACM (2009), 123–130.
19. Poppe, R. A survey on vision-based human action recognition. *Image and vision computing 28*, 6 (2010), 976–990.
20. Rosales, R., and Sclaroff, S. Improved tracking of multiple humans with trajectory prediction and occlusion modeling. Tech. rep., Boston University Computer Science Department, 1998.
21. Rosales, R., and Sclaroff, S. 3d trajectory recovery for tracking multiple objects and trajectory guided recognition of actions. In *Conference on Computer Vision and Pattern Recognition*, vol. 2, IEEE (1999).
22. Schulz, D., Fox, D., and Hightower, J. People tracking with anonymous and id-sensors using rao-blackwellised particle filters. In *IJCAI* (2003), 921–928.
23. Stöber, T., Frank, M., Schmitt, J., and Martinovic, I. Who do you sync you are?: smartphone fingerprinting via application behaviour. In *6th ACM conference on Security and privacy in wireless and mobile networks*, ACM (2013), 7–12.
24. Teixeira, T., Jung, D., Dublon, G., and Savvides, A. Pem-id: Identifying people by gait-matching using cameras and wearable accelerometers. In *Third ACM/IEEE International Conference on Distributed Smart Cameras*, IEEE (2009), 1–8.
25. Teixeira, T., Jung, D., and Savvides, A. Tasking networked cctv cameras and mobile phones to identify and localize multiple people. In *12th international conference on Ubiquitous computing*, ACM (2010), 213–222.
26. Wirelessdefence.org. Karma attack. `http://www.wirelessdefence.org/Contents/KARMAMain.htm`.
27. Xia, L., Chen, C.-C., and Aggarwal, J. Human detection using depth information by kinect. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE (2011), 15–22.