Zero-Effort Payments: Design, Deployment, and Lessons

Christopher Smowton*, Jacob R. Lorch†,
David Molnar†, Stefan Saroiu†, Alec Wolman†
†Microsoft Research, *University of Cambridge

ABSTRACT

This paper presents *Zero-Effort Payments* (ZEP), a seamless mobile computing system designed to accept payments with no effort on the customer's part beyond a one-time opt-in. With ZEP, customers need not present cards nor operate smartphones to convey their identities. ZEP uses three complementary identification technologies: face recognition, proximate device detection, and human assistance. We demonstrate that the combination of these technologies enables ZEP to scale to the level needed by our deployments.

We designed and built ZEP, and demonstrated its usefulness across two real-world deployments lasting five months of continuous deployment, and serving 274 customers. The different nature of our deployments stressed different aspects of our system. These challenges led to several system design changes to improve scalability and fault-tolerance.

Author Keywords

mobile payments; Bluetooth; BLE; face recognition; indoor localization; biometrics; latency; scalability; fault tolerance.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation (e.g. HCI): User Interfaces.

INTRODUCTION

Although mobile computing has delivered on the promise of *computing everywhere*, mobile devices are not only not invisible, but they often require users' undivided attention. The alternative is one where *computing everywhere* is done in a seamless manner, requiring zero effort from users. Although still in its infancy, seamless mobile computing is starting to emerge both in industry and the research community. One example is mobile payments, where several companies [25, 19, 27] recently launched "hands-free" payments – to make a purchase, a customer only needs to tell the cashier their name. Another example is "smart" home appliances, such as robotic vacuum cleaners [11] or learning thermostats [16], that infer what schedules will likely minimize human inconvenience.

This paper presents a seamless mobile computing system, called *Zero-Effort Payments (ZEP)*. As the name suggests, ZEP is a payment system in which mobile users pay with "zero-effort". Upon approaching a cashier for payment, the user is identified based on a combination of face recognition, low-power wireless radios (e.g., BLE), and human assistance (i.e., relying on the cashier's input to confirm the user).

To the best of our knowledge, ZEP is the first mobile payments system that is seamless. Earlier efforts advocated using the phone as a wallet (e.g., Google Wallet or NFC-based payments); however, we regard these systems as different than ZEP because they do not attempt to recognize customers seamlessly. Since ZEP's inception, a couple of similar systems have emerged in industry (e.g., Square [25], PayPal Beacon [19]), but none of them use face recognition. Instead, they rely solely on the cashiers sorting through a list of photos to identify the buyer. In contrast, ZEP makes use of face recognition to automatically sort the photos, and thus assists the cashier in their task of finding the buyer's identity. Without such assistance, cashiers are more likely to make errors. We will present the results of a user study we conducted that demonstrates how the lack of cashier assistance can lead to high error rates. Our experience with ZEP has shown that customers appreciate a short video of their transaction. Only very recently, another startup (Uniqul [27]) appears to be in the process of launching technology similar to ZEP; like ZEP, Uniqul also uses face recognition for biometrics. However, Uniqul requires the customer to enter a PIN number when its identification has a low confidence [28]. In contrast, ZEP aims at a completely seamless paying experience.

Over two years, we have designed, implemented, and deployed ZEP from the ground-up. ZEP is the result of three man-years worth of effort leading to a 5 month continuous deployment to serve 274 customers while processing over 30 million video frames. Along the way, ZEP has undergone major modifications as we encountered unforeseen challenges.

This paper's main contribution is to describe the lessons from and the insights into designing, implementing, deploying, and operating a mobile seamless system for payments. We present the three major phases of ZEP, from inception, to making the system highly scalable, and finally to making it fault tolerant. For each phase, we analyze the design choices available. Some choices turned out to be wrong; we present these choices and describe how we revised them. We also describe the lessons learned from ZEP; these lessons cover various operational aspects, such as meeting our IRB's stringent privacy requirements, and which aspects of the system's novelty wore off quickly on our users, and which did not.

PHASE #1: NO SINGLE IDENTIFICATION SCHEME IS SUFFICIENT

During the very early phases of ZEP's design, we envisioned building a payment system capable of identifying customers

UbiComp'14, September 13–17, 2014, Seattle, WA, USA.

ACM 978-1-4503-2968-2/14/09...$15.00.
http://dx.doi.org/10.1145/2632048.2632067

seamlessly based only on biometrics. Such a vision proved to be a chimera because all biometrics schemes we considered have serious shortcomings.

Biometric Schemes

A biometric scheme must meet three requirements to fit the needs of ZEP.

It must be accurate. The biometric scheme must have low false positives and low false negatives. False positives lead to mis-identification, whereas false negatives lead to people not being identified by the system.

It must be non-invasive. The biometric scheme should require little additional effort on the part of the customer. This is essential for meeting the seamless requirement of ZEP.

It must resist attacks. It should be difficult for an adversary to impersonate a particular customer. In ZEP, biometric identification is done in the presence of a cashier. Having a human in the loop makes biometrics much harder to attack. For example, holding a photo in front of a camera can easily fool face recognition, but it is much harder to fool a cashier.

Fingerprints

While fingerprint-based identification has high accuracy, we determined that fingerprints do not meet the seamlessness requirement because they require customers to touch a fingerprint sensor. Cleanliness of the sensor is an additional issue because some people question the hygiene of such a solution. There are also people, such as cooks or people who have survived a fire, who lack easily readable fingerprints.

Basic fingerprint readers are not attack-resistant. It is possible to build "fake fingers" undetectable to casual inspection by a store employee. Although sophisticated mechanisms to combat such fake readings exist, these methods are expensive and make readings less seamless [15, 30].

Voice-based Identification

Unfortunately, the state of the art in voice-based identification requires long voice samples to provide accuracy rates of roughly 80-85% [29]. As the state of the art advances, it may become viable to do zero-effort identification of users by listening to short statements they make as a natural part of conducting a transaction. At this time however, the relatively high error rate combined with the length requirement prevented use from using voice-based identification in ZEP.

Iris Recognition

The human iris contains distinctive patterns that seem unique to each individual, even between identical twins. Almost two decades ago, researchers proposed a way to compute a short iris representation called an *iris code* [9]. The key requirement is that the eye be illuminated with a suitable source of infrared light, then viewed by an infrared-sensitive camera. Iris codes have been computed across populations of tens of thousands of people from different demographics with low false positive rates, and NIST conducts competitions periodically to measure accuracy [17].

Today, multiple companies sell iris scanners that have high accuracy at short range [23, 5]. To use one, a user must look into an eyepiece that combines illumination and a camera, a procedure far from effortless. Furthermore, recent work has shown how to fool iris recognition using eye images synthesized from iris codes [10]. This suggests the need for the "human in the loop" to avoid simple impersonations.

Longer-range systems have started to emerge, being aimed at airport security terminals [24]. They consist of two large pillars, similar to a metal detector. When a person walks through, an infrared light is shined in their eyes, and the system captures the iris images. Although this is more seamless than short-range systems, such systems suffer from deployment barriers because they require placing pillars wherever people must be identified.

Gait Identification

One possible depth-enabled biometric is *gait identification*, which recognizes a person through idiosyncrasies in walking. This biometric is effortless since a customer would only need to walk into a section of the store covered by the depth sensor. Recent work has shown the Kinect and existing machine learning algorithms can reconstruct skeletal data using depth sensors with a 91.0% accuracy rate [21]. Although the security of this biometric is not well understood, it appears difficult to intentionally mimic.

However, we determined that gait identification is not ready for use in ZEP because the technology has yet to be proven in real-life scenarios. Furthermore, the study described above only used seven subjects who all performed the same walk in the same room. Even in such an unrealistic environment, one test subject could not be recognized at all.

Face Recognition

Most face recognition research work is done at the algorithmical level focusing on improving the accuracy rates of the underlying algorithms, and testing them against published benchmarks [20]. Our related work section will describe this work in more detail, but for a survey of recent results, see [34]. The accuracy rates reported by this work vary widely (e.g., 50% accuracy rate in [12], and 92% accuracy rate in [35]) depending on the algorithms used, the quality of the training data, and the conditions under which testing is done, such as the degree of illumination, and the variation in the subjects' pose or expressions.

Upon surveying the work on face recognition, two observations emerged. First, accuracy degrades rapidly as the *gallery* size increases. The gallery refers to the size of the database of identities matched against. Second, face recognition today cannot produce perfect results even under ideal conditions. While the accuracy rate can improve drastically in well-controlled experiments, it can never be guaranteed to be perfect. These two observations led us to conclude that deployments in practice can succeed only when the gallery size is not large and imperfect answers can be tolerated.

The Need for Additional Identification

As Table 1 shows, all biometric schemes we considered fall short of meeting ZEP's needs. Face recognition is the only biometric that shows promise, and that only when the gallery

	Finger	Voice	Iris	Gait	Face
Accurate	✓		✓		✓*
Non-invasive		✓		✓	✓
Secure		✓	✓		✓

Table 1. Shortcomings Biometrics when used for ZEP. Face recognition is accurate only if the gallery size is small; we denote this by ✓*.

size is small. This analysis convinced us of the need for additional identification in ZEP.

Fortunately, we were able to overcome the shortcomings of face recognition by adding two additional identification schemes. First, we use device identification – identifying a person by identifying a device he or she is carrying. The use of device identification can ensure that the gallery is never too large. We carefully considered three different technologies, Passive Radio-Frequency Identification (RFID), Bluetooth, and Bluetooth Low Energy (BLE), and chose BLE due to its low power and low-latency device discovery.

Second, we use cashier assistance – asking the cashier to make the final confirmation of the buyer's identity. The additional human in the loop can correct for face recognition errors. For example, face recognition could provide a set of four choices for customer identification to a store associate.

We felt this last step is necessary because the cost of a single ZEP mis-identification is high – a false payment attributed to a customer. However, ZEP could be used in scenarios that could tolerate some errors, such as (1) providing coupons on-the-fly to customers in a store based on their shopping histories, or (2) dispatching the "right" sales associate trained in the type of merchandise most interesting to the customer. We believe human assistance is not necessary in these scenarios due to their cost of errors being lower than for payments.

ZEP ARCHITECTURE

Armed with the combination of our three identification technologies, we proceeded to design and implement ZEP. This section's goal is to provide a high-level description of the ZEP architecture and hardware. This description serves as the basis of our presentation of the next two phases of ZEP that correspond to our two deployments.

ZEP has three main goals: accuracy, speed, and scalability. High accuracy ensures we do not create frustration for cashiers and customers due to mis-identification. Low latency is important to ensure we provide data in time for it to be useful. Finally, for the system to scale to large populations, the system must reliably identify customers even when hundreds of thousands of potential customers are registered. This section describes the design chosen to satisfy these goals.

An overview of this design is presented in Figure 1. The *detector* is a computer equipped with a camera and with Bluetooth Low Energy (BLE) capability. It uses the camera to view customers' faces and BLE to detect the presence of customers' devices. The detector determines which customers are present and sends this information to a *selector*, which is a tablet computer. The selector presents the customers' names and head shots to the store, so when an employee needs to know a certain customer's identity, he or she can readily

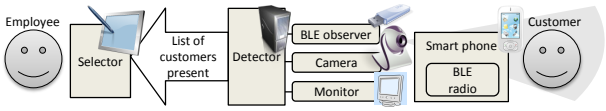

Figure 1. Overview of ZEP design.

deduce it by comparing the customer's appearance with the presented head shots. The system obtains these head shots, along with the customer device information, when the customer registers with the ZEP system.

Face Recognition

To determine which customers are in a certain location (e.g., directly in front of the cash register), we point the detector's camera at that location. By comparing the faces appearing in the camera's video with those of customers whose devices are nearby, the detector decides which customers are not just close to, but actually in, the area of interest.

To identify customers' faces, the detector uses Microsoft FaceSDK[1], a state-of-the-art library for detecting and identifying faces in digital pictures. For each video frame the camera captures, it passes the frame to FaceSDK. It also provides FaceSDK with a corpus of potential customers to choose from, namely those whose devices have recently been observed. It describes this set of customers to FaceSDK as a set of *profiles*, each of which contains a set of face images of a certain customer. These are images we collect from the customer during registration.

FaceSDK provides a ranking of each customer profile for each face detected. These rankings can be combined in many ways to produce face recognition scores for a final identification during payment. During our deployments, we calculate a *score* for each customer using an exponentially-weighted moving average. Our evaluation describes the results of our exploration of the effectiveness of different rankings.

Face recognition is CPU-intensive and can incur high latency. To combat this, the detector offloads tasks to *workers*, i.e., server-class machines on the premises or in the cloud. Face identification can also raise privacy concerns, so we need to inform customers and give them control over their private data. To provide more transparency in what data our system gathers, we installed a CCTV monitor next to the camera that shows customers the captured feed.

Hardware Details

We run the detector on an HP Z210 workstation equipped with 16 GB of RAM and an 8-core Intel Xeon E3-1245 CPU running at 3.3GHz. We use a Microsoft LifeCam Studio camera which retails for US$100, and a CC2540 USB dongle as the BLE scanner (next section will describe more details about our BLE devices).

BLE-based Identification

Each registered customer must carry a device, such as a modern smartphone, equipped with a BLE radio. The customer configures the device to be always discoverable. Because

[1] `http://research.microsoft.com/en-us/projects/facesdk/`

BLE is designed for low-power discovery, such configuration does not significantly impact the device's battery life.

BLE includes various protocols for use in discovery, formalized as a set of *roles* a participant can fill [6]. The ones we use are the *broadcaster* role, which periodically broadcasts an advertisement, and the *observer* role, which watches for advertisements. The customer's device acts as the broadcaster and the detector's device acts as the observer.

Hardware Details

Many mobile devices support BLE, such as the iPhone and many Android smartphones. However, as BLE is fairly new, it is not yet well exposed to developers on these platforms. For instance, the iPhone does not currently allow applications to use the broadcaster role, or to maintain BLE discoverability while the phone is asleep. This, combined with the fact that our customers did not carry phones with BLE at the time of our deployment, made us use an alternative BLE device.

We use Texas Instruments CC2540 BLE Mini Development Kits, shown in Figure 2. Each kit includes a USB dongle and battery-powered key fob, and we use the USB dongle as the detector's BLE device, and we provide each customer with a key fob that simulates a smartphone with better BLE support.

Each BLE fob uses a CR2032 battery, which provides 200 mAh at 3 V. By connecting a power meter to the fob, we determined that it consumes an average of 0.22 mW. This suggests the battery should last slightly under four months, which is consistent with our experience. Furthermore, since consuming 0.22 mW for 24 hours would use only about 0.1% of an iPhone's battery capacity, we expect customers will not mind running BLE continuously.

Human Assistance

ZEP provides guidance to a human employee who makes the ultimate customer identification. It provides this guidance via a tablet facing the cashier (labeled as the "selector" in Figure 1), which operates as follows. Every second, it requests a list of present customers from the detector. The detector returns the list of customers whose devices are present, sorted in decreasing order of face recognition score. The tablet then presents these to the cashier to aid in his or her identifications.

The tablet displays the head shots and names of the customers in order. We experimented with various forms of UI and found that showing four head shots provides an intuitive set of identity choices to the cashier. Since the cashier may want to consider customers with even lower scores than these four, the tablet provides a way to scroll to lower-scored customers.

One-time Registration

A customer must perform a one-time registration with ZEP. During registration, we record the customer's BLE MAC address and a short video of the customer's face. We produce a FaceSDK profile from this video, and we tell the customer to select a single one of the frames as if he were selecting one for a picture ID. We use this single frame as his head shot, i.e., the picture we present to an employee trying to find a match for a physically-present customer's face.

Figure 2. Texas Instruments CC2540 Mini Development Kit includes, from left to right, a debugger for programming BLE devices, a key fob BLE device, and a USB dongle BLE device.

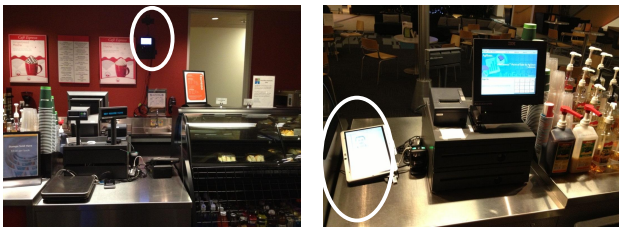

Figure 3. ZEP deployment in our building's cafeteria. On the left, the camera and CCTV-like monitor are placed on the backwall. On the right, the tablet-based selector is placed next to the POS.

ZEP DEPLOYMENTS

We deployed ZEP in two environments. The first was a two-day technology fair with thousands of attendees, and the second was a long-term installation at a coffee stand in our corporate cafeteria shown in Figure 3. Table 2 summarizes high-level statistics of the data gathered in each of the two deployments. During the technology fair, we only gathered data during the second deployment day.

The two deployments were quite different. During the technology fair, many people visited our booth and coffee cart. Thus, most gathered frames contain several faces. In contrast, the frames gathered during the long-term deployment in our corporate cafeteria have many fewer faces on average. On the left, Figure 4 illustrates the distribution of the number of faces per frame during each deployment. On the right, Figure 4 displays the distribution of all faces found within 20 seconds before a transaction occurred. For deployment #2, many frames only had one face; because our camera recorded video at 10 fps, only a few transactions (15%) had more than 200 face images in the 20 preceding seconds. In contrast, during deployment #1, more than 200 face images were discovered in those 20 seconds 95% of the time.

During both deployments, we never learned of any misidentification for any transaction. No customer ever reported not being charged properly, or being charged on behalf of someone else. In both our deployments, the ZEP tablet showed up to four identities as potential matches on its screen. A simple interface allowed the cashier to scroll down for the next four matches. During the technology fair, which was a busy environment with many people nearby, ZEP displayed the correct identity of the paying customer on the first screen, i.e., in the top four, 80% of the time. If one considers the second screen as well, then ZEP was perfect: All customers appeared on the top two screens. In fact, the correct identity was in the top five matches 92% of the time. In our

	Deployment #1	Deployment #2
Duration	2 days (03/07-03/08, 2012)	20 weeks (05/14- 09/28, 2012)
# of registered users	255	19
# of payments made	102	540
# of frames gathered	256,831	30,998,191

Table 2. High-level statistics of our two deployments.

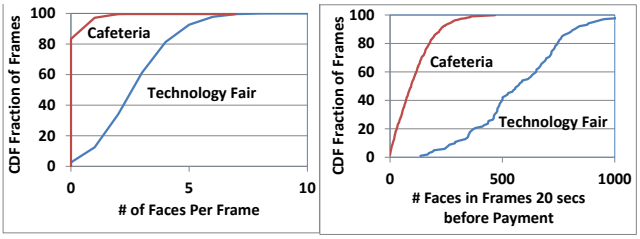

Figure 4. Distribution of number of faces per frame. On the left, the data is gathered from all captured frames, whereas on the right, only from the frames 20 seconds prior to a transaction.

cafeteria deployment, ZEP was always perfect and showed the correct identity on the first UI screen.

PHASE #2: REDUCING FACE RECOGNITION LATENCY

Original Design

Face recognition is CPU-intensive, so unless the detector is highly provisioned it may become overloaded, causing face identification tasks to queue and experience high latency. One strategy to avoid this is to detect when the load is high and have the detector drop frames instead of queueing them; after all, there is a lot of redundancy in consecutive frames. However, a preferred strategy is to offload work to *worker* threads, if they are available on the premises or in the cloud.

Worker threads can offer large amounts of computing power for the process of face identification. To leverage more than one worker, we can readily parallelize the work because the processing of each video frame is independent. We can thus leverage multiple workers by sending different frames to different workers.

We used the following equation to provision the number of workers in ZEP:

$$\#\_of\_workers \geq fr\_latency\_per\_frame * 10 \quad (1)$$

The number of workers should be greater than the latency of processing a single frame times the camera throughput, which is 10 frames per second. In all our lab experiments, a single frame was always processed in less than one second. We thus originally provisioned ZEP with 16 workers for our technology fair deployment (each worker runs on one CPU core).

This assumption turned out to be wrong. The technology fair is a very busy environment with lots of people visiting the fair and our booth. This created very "busy" frames that typically took 2-to-3 seconds to process, and sometimes even more than 5 seconds. Figure 5 shows one frame gathered with 11 faces in it, and this frame takes 5.26 seconds to be processed by one worker. Our original 16 workers often fell behind keeping up with the 10 frames per second rate captured

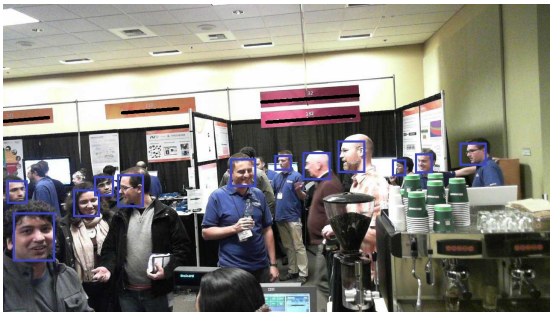

Figure 5. The technology fair deployment had very "busy" frames. This frame has no fewer than 11 faces in it.

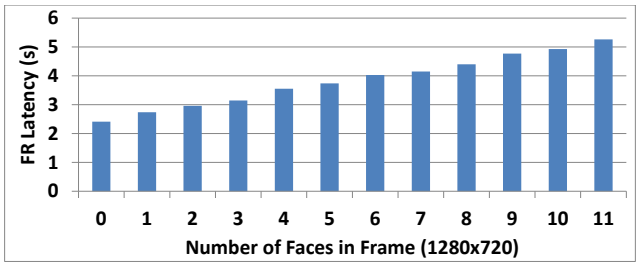

Figure 6. Latency of processing one frame increases linearly with the number of faces found in the frame.

by the camera. To remedy this problem, we immediately deployed an additional 22 workers for a total of 40 workers. This level of provisioning was sufficient to ensure that ZEP could process all the gathered frames without drops.

The Need for Sub-frame Parallelism

While dealing with ZEP's throughput needs was simply a matter of overprovisioning, reducing the latency of face recognition turned out to be much more difficult. Although ZEP's original design used frame-by-frame parallelization to reduce the latency of face recognition, this technique cannot reduce the latency below the time to process a single frame. Unfortunately, as mentioned above, this time can be significant. Figure 6 shows the latency of processing a frame as a function of the number of faces found in the frame.

To solve this problem, we designed another technique to further reduce the latency of face recognition. We divide each frame into sub-frames, and send each sub-frame to a separate worker. To ensure we do not leave out any faces by cutting them into unrecognizable halves, we cut the frames in a way that *guarantees* that each face will appear in full in at least one sub-frame. To do this, we determine the maximum length any human's face could possibly occupy in any dimension from the camera's perspective. We then choose sub-frames so that they overlap in at least this amount. Figure 7 shows an illustrative four-way cut in which any two adjacent sub-frames overlap, but note that our technique can use subframe counts other than four. As the overlapping region's size is equal to the maximum facial length, each face appears in full in at least one sub-frame.

Our technique is effective as long as the typical face size is small relative to the size of a frame. Fortunately, this is the case with ZEP because the camera is placed behind the cashier at about 5-7 meters away from the customer's face.

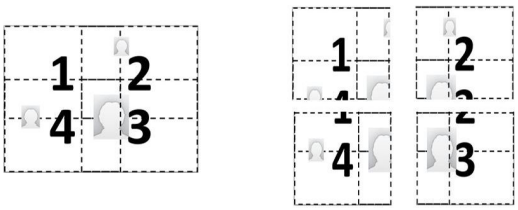

Figure 7. Dividing frames into sub-frames to reduce latency.

We also experimented with bringing the camera closer to the customer's face to improve the accuracy of face recognition. However, some customers felt uncomfortable when seeing a camera closer to their faces. To reduce the degree of discomfort, we decided to place the camera behind the cashiers.

PHASE 3: THE UNSUITABILITY OF FAST CRASH

ZEP was also deployed at a coffee stand in our building for over four months. The long-term nature of this deployment stressed the fault tolerance aspects of our system.

The detector was a single point of failure. For this reason, we originally decided to pursue a design that minimizes the recovery time for the detector and uses a fast crash recovery model: on any error, the detector crashes and quickly restarts afresh. Fast crash recovery is a classic technique to improve availability in complex systems [3, 7, 18]. Note that a detector failure does not immediately affect the UI tablet (i.e., the selector) which continues to function and show the identification matches. Each second, the tablet contacts the detector; if down, the tablet eventually times out (5 seconds) and shows no more identification matches. Until the time-out fires, the tablet remains functional and can be used to conduct transactions. This design makes the detector's restart transparently to employees and customers.

Unfortunately, our hardware's behavior made us abandon the fast crash recovery model for the detector. First, our camera took an average of two seconds to initialize, a behavior consistent with inexpensive Web cameras. We believe a *lower bound* of two seconds of downtime on recovery is unacceptable for the detector's availability needs. Second, the camera driver would sometimes return an error to an initialization request if the camera was recently running. Thus, sometimes it would take 4-5 seconds of repeated tries for the camera to initialize successfully. Third, the camera would sometimes get stuck while initializing; we thus had to write a watchdog that would detect this and reboot again. Finally, sometimes the auto-focus mechanism of the camera would not work properly – the camera would be out of focus on a restart. We found that covering the camera completely for a few seconds and uncovering it would re-trigger the auto-focus mechanism and make the camera focus properly.

Additional problems stemmed from the BLE packet capture software written by TI, the manufacturer of our BLE hardware. This software would capture any incoming BLE packets and relay them over UDP to the detector. As this software was designed for short-term debugging rather than long-term operation, it often froze without reporting an error. So, about two months into our second deployment, we re-wrote the firmware of our BLE sniffer device and eliminated the need to run the TI software.

Both these hardware issues made us abandon a fast crash recovery model for the detector. Instead, we decided to try to make the detector as robust as possible by offloading as much functionality as possible from it. The detector ended up being quite lean; its roles were to capture the frames, write one copy to the disk and feed one copy to the face recognition workers, and offer a live feed of the face recognition scores to the UI tablet. Despite relatively little functionality, making the detector robust turned out to be challenging: at best, our detector could run for a week without crashing. Most errors encountered continued to stem from camera hardware. To overcome this issue, we restarted the detector manually at off hours every couple of days.

EVALUATION

This section characterizes the error rates of each identification scheme in ZEP, when used separately. At large scale, all identification schemes (including ZEP's) have high error rates. Face recognition, BLE identification, and human-based identification are far from perfect when faced with hundreds (or even tens) of potential identities. This observation raises the following question: *What scale can each identification scheme tolerate*?

Performance of Face Recognition Alone

During each ZEP transaction, the cashier selects the person paying at the register to enable payment. This triggers a video receipt to be sent. We consider the cashier's slection to be the "ground truth" because these receipts never generated any complaints about misdirected payments.

Our accuracy evaluation compares the rankings produced by face recognition with the ground truth. In particular, we compute the rank of the paying customer using face recognition alone for identification; a rank of one would be a perfect match. However, in our system deployments, any rank between one and four guarantees that the customer's face immediately appears on the selector UI facing the cashier. For ranks higher than four, the cashier would need to scroll down through the UI to find the identify of the paying customer.

Because face recognition produces a ranking of customers for every face image found in every frame immediately preceding a transaction, these separate rankings must be aggregated together to produce a single full ranking. Many separate schemes and heuristics can be used to aggregate these rankings. Based on our experimentation, we selected the following aggregation schemes, and used them on all frames gathered up to 20 seconds before a purchase was made:

1. Average/Median of all rankings. This scheme computes, for each customer, both the average and the median of all rankings that person attains for every face image in the transaction. Customers are then ranked by this measure.

2. Best ranking. This scheme computes, for each customer, the best ranking achieved for all faces in the transaction. Customers are then ranked by this measure.

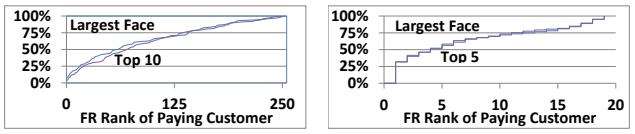

Figure 8. Face recognition accuracy for deployment #1 (left) and #2 (right).

3. Largest face. This scheme select the largest face in a frame only. The intuition behind this heuristic is that the paying customer is likely to be closest to our camera. However, note that this is not perfect because (1) people have different face sizes, and (2) several people can all stand in front of the cashier even though only one is the true paying customer.

4. Top k. This scheme considers the top k matches for each ranking for each face. These rankings are then summed. An additional boost is added to the top-ranked customer for each face image. The intuition behind this top-k filtering is that beyond the first k results, face-recognition results are probably very noisy and should be discarded.

We found the last two schemes to produce the highest accuracy, so we omit presenting results for the other schemes.

Hundreds of Identities

On the left, Figure 8 shows the distribution of accuracy for our technology fair trace having 255 identities. The error rates are high at this scale – the customer's true identity is in the top 10 only 20% of the time. Even worse, the customer is in the bottom half of the ranking 25% of the time. Clearly, face recognition at the scale of hundreds of individuals is not viable.

Tens of Identities

On the right, Figure 8 shows the distribution of accuracy for our second trace having only 19 identities. The error rates improve significantly. While for the previous trace we used $k = 10$ for our top k heuristic, here we used $k = 5$ due to the smaller-sized database. This heuristic alone would find the true customer on the first screen of the tablet, i.e., among the top four matches, more than half the time. However, in some cases, the true identity of the customer is still ranked low.

Conclusion: Face recognition suffers from high error rates at scales beyond tens of individuals.

Performance of BLE Alone

In indoor environments, RF signal strength can vary due to many factors including multipath interference, physical obstructions (including people), and interference from other wireless networks. Nevertheless, previous research has shown that RF signal strength can be used as an approximate measure of the distance between wireless devices [4].

Various factors can affect the performance of device identification, such as where the device's holder is standing, where and how the holder holds the device, and the remaining battery charge level of the device because that affects the voltage supplied by the battery. Unfortunately, we could not control or even measure all these properties during our deployments. Thus, to evaluate their effects, we manually modify these conditions and measure their effects.

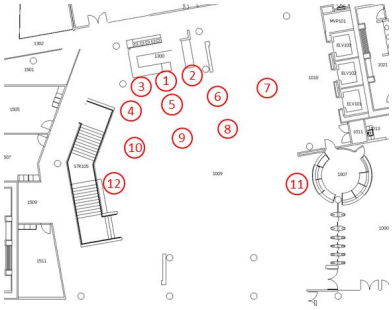

Figure 9. Area where the ZEP long-term deployment took place. 1 through 12 show the locations where BLE signal strength was recorded.

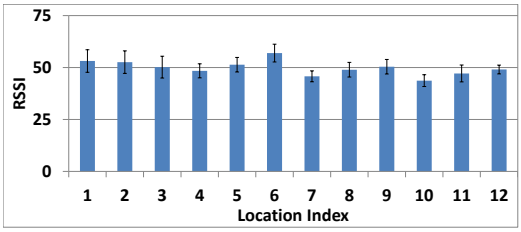

Figure 10. Mean signal strength (RSSI) and standard deviation, at locations 1 through 12.

Experiment Design

Figure 9 shows a diagram of the building lobby and coffee stand where our long-term deployment took place. Locations 1 through 12 in the diagram represent locations where we placed a customer's BLE radio, and then recorded the signal strength as reported by the BLE receiver. Location 1 in the diagram is where a customer at the head of the line would stand when making a purchase, directly across from the cash register. Locations 2 through 12 correspond to other possible locations of people in the lobby area.

We performed experiments at each location. For certain locations we varied the position of the BLE radio on the person performing the experiment, and we varied the battery charge on the BLE radio. Each result shown in the graphs below shows the signal strength data as recorded by the BLE receiver over a two minute period. We show both the mean signal strength and the standard deviation for each result.

Result #1: RSSI does not correlate with distance

In the environment shown in Figure 9, location 1 corresponds to where the current paying customer is likely to stand, which is also the nearest location to our BLE receiver. Figure 10 shows the signal strength from all 12 locations, where each location is identified on the x-axis. In every location the BLE radio was in the front pocket of the person performing the experiment.

There is very little correlation of distance and signal strength, but with significant variation. For example, the mean RSSI at location 6 is larger than that at locations 1, 2, and 5. However, 1, 2, and 5 are all closer to the BLE receiver than location 6.

Result #2: RSSI affected by how device is held

Figure 11 presents the effects on signal strength of where the BLE radio is located on the person carrying it. For each of four positions, we show four bars, which represent 1) in the

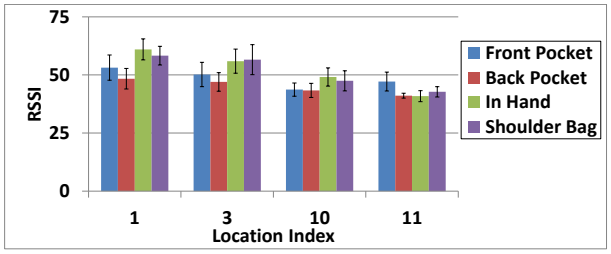

Figure 11. Comparison of BLE signal strength at different positions on a person (front pocket, back pocket, in hand, and in shoulder bag).

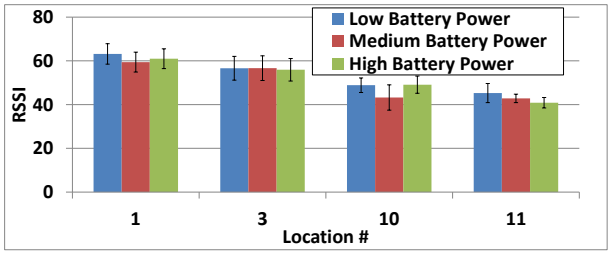

Figure 12. Comparison of battery charges on the BLE transmitter (Low = 1.75V, Medium = 2.10V, High = 2.8V).

person's front pocket, 2) in the person's back pocket, 3) in the person's right hand, and 4) inside a laptop shoulder bag that is zipped shut. From this graph, we see that the transmitter's location on the person carrying it can have a large effect on signal strength: at location 1, the in-hand signal strength is more than 12 dB larger than the back-pocket signal strength.

Result #3: RSSI affected by battery levels

Figure 12 presents the effects of battery charge of the BLE transmitter on RSSI. As with the previous graph, we show three bars for each of four locations. Each bar type corresponds to a different remaining-charge level, and thus a different voltage level, for the battery that powers the BLE radio. We used the same BLE transmitter, but replaced the battery to perform these experiments. In this graph, once again we find that variations in battery charge can lead to measurable differences in the receiver mean signal strength.

Conclusion: BLE alone suffers from high error rates and is affected by a variety of factors, such as how the device is held or the battery level.

Performance of Human-based Identification Alone

We used Amazon's Mechanical Turk to conduct an experiment to determine the error rates of human-based identification when faced with tens of potential identities. We recreated ZEP's tablet UI on a website which we posted on the Mechanical Turk and asked people to identify 15 ZEP customers out of a database of 60 faces. Like ZEP's UI, the experiment showed four faces at a time together with arrows to scroll back and forth. Table 3 shows the accuracy rates of human-based identification for each of the 15 ZEP customers.

On one hand, these results show that humans alone are very good at identifying customers, better than face recognition alone, or BLE alone. On the other hand, human-based identification also does not scale – when faces with tens of faces,

Customer #	1	2	3	4	5	6	7	8
Accuracy (%)	44	76	72	80	84	92	76	80
Customer #	**9**	**10**	**11**	**12**	**13**	**14**	**15**	
Accuracy (%)	44	80	88	84	76	72	72	

Table 3. The accuracy of human-based identification.

Figure 13. Three customers mistaken for each other.

the error rates vary between 8% (at best) and 56% (at worst), which is unacceptable for a payment system.

A deeper investigation of these results revealed a surprising finding. The customer with the least accuracy (index #9) was frequently mistaken for two other customers. When looking at their pictures, we also found these three customers to resemble each other (see Figure 13). In our Mechanical Turk's tests the correct ZEP customer was the one on the left.

One of our experiment's shortcomings is missing the longer-term implications of human cognitive overload after selecting among images during a whole workday. It is possible that accuracy worsens as the cashiers become more tired over time.

Conclusion: Humans are more accurate at identifying customers than either face recognition alone or BLE alone. However, their error rates are still too high at a scale of tens of individuals.

Making Identification Work with ZEP

The combination of BLE and human-based identification could work in low-traffic areas. However, busy areas, such as in our technology fair deployment or inside a mall, could have tens of people wearing BLE-enabled devices, and thus will likely lead to human errors.

Instead ZEP combines face recognition with BLE before asking for human-assistance. Using BLE allows ZEP's face recognition to reduce the size of the database because the candidates for face recognition would only be people discoverable by BLE. We systematically investigated the combination of face recognition and BLE by selecting different samples of people assumed to be "nearby" due to BLE discovery. Figure 14 shows the probability of displaying the correct identity of the paying customer on ZEP's tablet top screen (i.e., top 4 matches) and top two screens (i.e., top 8 matches) as a function of the number of people nearby for the technology fair deployment. Using BLE as a filtering mechanism drastically improves the accuracy of face recognition.

Additionally, ZEP used human-assistance as a final confirmation step. The combination of face recognition, BLE, and human-assistance has identified the correct customer in all 642 payments processed during our deployments.

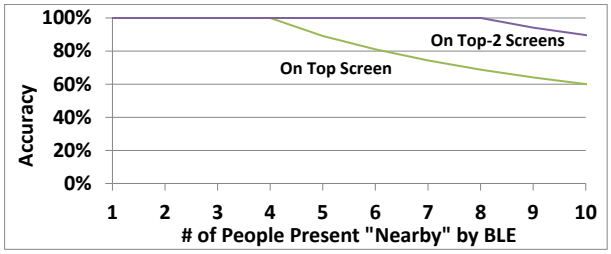

Figure 14. Accuracy of combining face recognition and BLE.

LESSONS FROM REAL WORLD USAGE

Privacy. ZEP raises serious privacy concerns. In particular, some people become uncomfortable knowing their faces are captured and submitted to an automatic face recognizer. Some people are under the impression that technology has reached a point today where it is able to identify and track them even if they are not registered with our system. The discrepancy between what people think face recognition technology can do and what it in fact does is quite large, and may stem from technologically-implausible scenarios prevalent in today's popular movies and TV shows. To alleviate these concerns, we have taken several steps with the goal of making the ZEP system as transparent as possible. These steps were taken after consultations with the IRB, privacy, and legal teams at our institution.

First, in all our ZEP deployments, we made heavy use of signage indicating a face recognition system is deployed in the area. Also, the area covered by our camera was clearly delimited. To guarantee that no identification was possible outside of the delimited area, we carefully oriented the camera position and angle to ensure we would not capture any portion of a person standing outside the area.

The delimited area was sprinkled with signs describing our process and also listing three ways in which customers could opt-out. To ensure that our data would not capture their faces, they can either: avoid stepping into the well-delimited area, ask the operator to turn off the camera, or sending us e-mail specifying a time range and requesting that we manually remove all frames where their faces were captured accidentally. Over the course of our deployments, there were several times when the camera was turned off due to a customer request, and we also received e-mail requesting manual removal twice. Another lesson learned from these deployments is that there is a need for a universally-known signage indicating the presence of system performing face recognition, the same way there is well-understood signage for police investigation areas, or CCTV cameras. Figure 15 shows one such sign possibility.

Video Receipts. Our experience with ZEP has shown a surprising lesson: while customers enjoyed their seamless payments, this enjoyment quickly wore out after the first couple purchases. In the long-term, customers' most useful benefit was that ZEP was automatically e-mailing video-based receipts of their purchases. The simple act of making receipts convenient and video-based carried a longer-term value than the payments alone. When our system crashed, the most common complaint received from customers was the absence of receiving a video-based receipt rather than the lack of a seamless purchase. Customers expected video-based receipts even when our system was down and they had to resort to traditional payment methods.

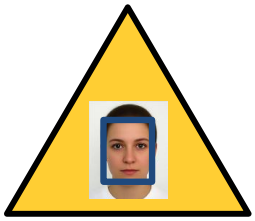

Figure 15. Face recognition privacy sign.

Hardware Problems. We never anticipated the hardware problems encountered during ZEP deployments. From the start, we limited our hardware (i.e., the camera and the BLE dongles) to inexpensive, commodity choices. In addition to the slow restart problems described earlier, we also had to replace the camera several times throughout our long-term deployment because it stopped working. Although it is hard to generalize, we suspect inexpensive cameras are not meant to be run 24/7 for several months at a time.

UI Problem. An earlier iteration of our tablet's user interface posed a problem for employees. Whenever the set of highest-scoring customers changed, our UI immediately refreshed the display. Thus, frequently, an employee attempting to screen-touch a customer's face would be frustrated by the UI quickly changing just before the touch. In some cases, the employee would not even realize the screen had quickly changed before the touch, and would become confused by the confirmation page showing a different customer face. To fix this, we consulted a UI expert, who recommended the sliding-tile motif we now use: The tablet's display is a tile of static information; when it needs to be changed, the tile visibly slides off the screen as another tile with new content slides in to replace it. While obvious in retrospect, the following UI principle guided the design of all our user-facing components: UI refreshes must be made gradually and not instantaneously.

The Scalability Limitations of Today's Mobile Payments Systems. Several startups (Square [25], PayPal Beacon [19]) have launched mobile payments that rely on the cashiers sorting through a list of photos to identify the buyer. Square relies on smartphone geo-fencing to detect when customers are nearby. We believe that the granularity of geo-fencing is too coarse for payments because it is not discriminatory enough. If Square were to become popular, the scalability limitations of humans sorting through pictures are likely to introduce a high degree of errors. In contrast, PayPal relies on BLE to detect nearby customers; BLE's shorter range is likely to make PayPal have better scalability than Square.

Foodservice Management Industry. After hearing about ZEP, the foodservice management industry expressed strong interest in funding and participating in additional ZEP deployments. This industry consists of contract management companies and self-operated facilities in large corporations, colleges and universities, hospitals, nursing homes, lifecare

facilities, and correctional institutions. When we raised the privacy issues encountered during our deployments, they pointed out that such technology could be especially useful in places with lesser privacy requirements, such as elementary schools or correctional institutions.

This industry is already deploying initial pilots of biometrics-based identification inside their cafeterias. Their interest in face recognition for identification is not due to seamlessness or the "coolness" aspects. Instead, they focus solely on service efficiency and scalability. Their main question to us was: "how fast can ZEP get people through a queue?", a question we did not focus on in our deployments.

RELATED WORK

The techniques used for seamless customer identification draw on previous work in wireless localization, face recognition, and mobile systems that make use of computer vision.

Wireless Localization. Over the past decade, there has been much work on using wireless radios for localization. One of the earliest projects was Radar [2], which built an indoor positioning system based on Wi-Fi signal strength. However, research projects have used a variety of types of wireless radio including Wi-Fi, RFIDs, Bluetooth, cellular, and ZigBee, to locate people indoors; excellent overviews can be found in two recent books [13, 33].

Face Recognition. Most research work in face recognition has focused on designing new algorithms and improving their accuracy rates. In contrast, there is much less published work on the challenges facing the deployment of face recognition systems in practice. In the US, the National Institute for Standards and Technology (NIST) has put together a benchmark called the *face recognition grand challenge* (FRGC). While researchers are measuring their algorithms' accuracy against benchmarks like FRGC, these benchmarks are far from the conditions systems experience in practice.

In contrast, our experience with ZEP differs from this research due to the needs of application to practice. For instance, frames do not necessarily capture front shots of people; indeed, some people never look at the camera. Also, the lighting can drastically change over time, e.g., a lightbulb may stop working for a day then get replaced with a newer, much brighter bulb. Additionally, for cost and logistical reasons, we used a non-professional-grade camera.

Nevertheless, the face recognition literature [31, 35, 34] contains several projects focused on evaluating the accuracy of face recognition in more realistic scenarios. One project evaluated the accuracy of recognizing a set of 35 celebrities in videos stored on YouTube; it reported a 60–70% accuracy rate depending on the algorithm used [12]. To achieve this, the identification techniques relied on face tracking, which identifies the same person across multiple consecutive frames. Our ZEP deployment did not use face tracking.

Another related project conducted an accuracy evaluation of several face recognition techniques using footage of lower quality [8]. The accuracy is evaluated using a metric called the *half-error total rate*, which is the average of false positive and false negative rates. While measuring accuracy is similar to measuring the false positive rate, a higher accuracy came at a higher false negative rate, which means that many frames reported no faces detected. Examining the results, most algorithms achieved an 80% accuracy rate only by admitting a 25–50% false-negative rate, i.e., by accepting no faces are found in a quarter to half of all frames.

Finally, a few other projects report high accuracy rates for face recognition in uncontrolled environments, specifically 86.3% [26] and 92% [35]. However, these results are obtained by constraining subject poses to be either front-facing [26] or constant across frames [35].

Mobile Systems and Computer Vision. Recent work has started to use computer vision in mobile systems. One application is localizing distant objects, such as buildings, by looking at them through a smartphone [14]. The combination of GPS-based localization with computer-vision processing of images gathered by a smartphone shows promising results in accurately pinpointing an object's location. Another application is cloud-based face recognition, such as that done by Google Picasa, to automatically tag photos taken by a smartphone [22]. Another project implements an indoor localization scheme based on ambience fingerprinting by observing that most stores have very distinct photo-acoustic signatures [1]. Finally, a recent workshop paper demonstrates the practicality of identifying people based on the patterns and colors of their clothes [32].

CONCLUSIONS

This paper describes ZEP, a mobile payments systems in which customers pay with zero-effort. ZEP uses three complementary identification technologies: face recognition, BLE-based device detection, and human assistance. ZEP has been deployed twice and went through three different phases that affected its design. The paper presents the scalability limitations of each identification technology when used alone. The paper also describes the challenges encountered with making ZEP fault tolerant. Finally, the paper presents the lessons learned from our ZEP deployments.

ACKNOWLEDGMENTS

We thank the UbiComp reviewers for their feedback and suggestions to improve our paper. Thanks to Sreenivas Addagatla, Victor Bahl, Ronnie Chaiken, Weidong Cui, Oliver Foehr, Jitu Padhye, Bryan Parno, and Matthai Philipose for many useful discussions and other contributions to ZEP. We thank AJ Brush, who gave us much-needed advice and feedback on UI design. We thank Mike Freeman, Bill Barfoot, Derrick Aiona, Will McGinnis, Bethany Rizk, and the rest of the Building 99 cafeteria staff for operating our system. We acknowledge Microsoft Research's legal and privacy teams who went beyond the call of duty to point out all the minute privacy aspects of our deployments. We thank Peter Lee who helped facilitate our deployments. Finally, we are deeply grateful to the hundreds of participating customers, especially those who used ZEP for months without receiving any coffee discounts in return.

REFERENCES

1. Azizyan, M., Constandache, I., and Choudhury, R. R. SurroundSense: Mobile Phone Localization Via Ambience Fingerprinting. In *Proc. of MobiCom* (2009).
2. Bahl, P., and Padmanabhan, V. N. RADAR: An In-Building RF-Based User Location and Tracking System. In *Proc. of IEEE INFOCOM* (Mar. 2000).
3. Baker, M., and Sullivan, M. The Recovery Box: Using fast recovery to provide high availability in the UNIX environment. In *Proc. of USENIX ATC* (1992).
4. Banerjee, N., Agarwal, S., Bahl, P., Chandra, R., Wolman, A., and Corner, M. Virtual compass: Relative positioning to sense mobile social interactions. In *Proc. of the 8th International Conf. on Pervasive Computing* (May 2010).
5. Bayometric. Crossmatch Retinal Scan 2 Iris Scanner, 2012. **http://www.bayometric.com/crossmatch-retinal-2-iris-scanner**.
6. Bluetooth SIG. Specification of the Bluetooth System, Core Version 4.0. Available at http://www.bluetooth.org, 2010.
7. Candea, G., Kawamoto, S., Fujiki, Y., Friedman, G., and Fox, A. Microreboot – A Technique for Cheap Recovery. In *Proc. of OSDI* (2004).
8. Chen, S., Mau, S., Harandi, M. T., Sanderson, C., Bigdeli, A., and Lowell, B. C. Face Recognition from Still Images to Video Sequences: A Local-Feature-Based Framework. *EURASIP Journal on Image and Video Processing 2011*, 790598 (2010).
9. Daugman, J. High confidence visual recognition of persons by a test of statistical independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence 15*, 11 (1993), 1148–1161.
10. Galbally, J., Ross, A., Gomez-Barrero, M., Fierrez, J., and Ortega-Garcia, J. From the iriscode to the iris: A new vulnerability of iris recognition systems. In *Black Hat Briefings USA* (2012).
11. iRobot. iRobot Roomba Vacuum Cleaning Robot, 2014. **http://www.irobot.com/us/learn/home/roomba.aspx**.
12. Kim, M., Kumar, S., Pavlovic, V., and Rowley, H. A. Face Tracking and Recognition with Visual Constraints in Real-World Videos. In *IEEE Computer Vision and Pattern Recognition (CVPR)* (2008).
13. LaMarca, A., and de Lara, E. *Location Systems: An Introduction to the Technology behind Location Awareness*. Morgan and Claypool, 2008.
14. Manweiler, J., Jain, P., and Choudhury, R. R. Satellites in Our Pockets: An Object Positioning System Using Smartphones. In *Proc. of MobiSys* (2012).
15. Matsumoto, T., Matsumoto, H., Yamada, K., and Hoshino, S. Impact of Artificial "Gummy" Fingers on Fingerprint Systems. In *SPIE Volume #4677 Optical Security and Counterfeit Deterrence Techniques IV* (2002). **http://www.cryptome.org/gummy.htm**.
16. Nest. Nest Thermostat, 2014. **https://nest.com**.
17. NIST. IREX III evaluation of one-to-many iris identification algorithms, 2012.
18. Ongaro, D., Rumble, S. M., Stutsman, R., Ousterhout, J., and Rosenblum, M. Fast Crash Recovery in RAMCloud. In *Proc. of SOSP* (2011).
19. PayPal. PayPal Beacon, 2014. **https://www.paypal.com/webapps/mpp/beacon**.
20. Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Min, J., and Worek, W. Overview of the Face Recognition Grand Challenge. In *IEEE Computer Vision and Pattern Recognition (CVPR)* (2005), 947–954.
21. Preis, J., Kessel, M., Werner, M., and Linnhoff-Popien, C. Gait Recognition with Kinect. In *First Workshop on Kinect in Pervasive Computing, in conjunction with Pervasive 2012* (2012). **http://noggnogg.com/pervasivekinect**.
22. Qin, C., Bao, X., , Choudhury, R. R., and Nelakuditi, S. TagSense: A Smartphone based Approach to Automatic Image Tagging. In *Proc. of MobiSys* (2011).
23. Sarnoff Corporation. IOM N-Glance, 2012. **http://www.sri.com/engage/products-solutions/iom-n-glance-modular-system**.
24. Sarnoff Corporation. Iris-on-the-move, 2012. **http://www.sri.com/engage/products-solutions/iris-move-biometric-identification-systems**.
25. Square. Square Wallet, 2014. **https://squareup.com/wallet**.
26. Tan, X., and Triggs, B. Enhanced Local Texture Feature Sets for Face Recognition under Difficult Lighting Conditions. *IEEE Transactions on Image Processing 19* (2009), 1635–1650.
27. UNIQUL. Uniqul Payments, 2014. **http://uniqul.com/**.
28. UNIQUL. Uniqul Payments – FAQ, 2014. **http://uniqul.com/faq**.
29. US National Institute of Standards and Technology. Speaker Recognition Evaluation (SRE) Results, 2010. **http://www.nist.gov/itl/iad/mig/sre10results.cfm**.
30. van der Putte, T., and Keuning, J. Biometrical fingerpring recognition: Don't get your fingers burned. In *IFIP TC8/WG8.8 Fourth Working Conference on Smart Card Research and Advanced Applications* (2000).
31. W. Zhao and R. Chellappa and P. J. Phillips and A. Rosenfeld. Face Recognition: A Literature Survey. *ACM Computing Surveys 35*, 4 (2003), 299–458.

32. Wang, H., Bao, X., , Choudhury, R. R., and Nelakuditi, S. Recognizing Humans without Face Recognition. In *Proc. of HotMobile* (2013).

33. Yang, J., Chen, Y., Martin, R., Trappe, W., and Gruteser, M. *On the Performance of Wireless Indoor Localization Using Received Signal Strength – in Handbook of Position Location: Theory, Practice, and Advances*. Wiley, 2012.

34. Zhang, C., and Zhang, Z. A Survey of Recent Advances in Face Detection. Tech. Rep. MSR-TR-2010-66, Microsoft Research Technical Report, 2010.

35. Zhang, X., and Gao, Y. Face recognition across pose: A review. *Pattern Recognition*, 42 (2009), 2876–2896.