PEM-ID: Identifying People by Gait-Matching using Cameras and Wearable Accelerometers

Thiago Teixeira, Deokwoo Jung, Gershon Dublon and Andreas Savvides
Yale University, New Haven, CT 06511 — e-mail: firstname.lastname@yale.edu

***Abstract*—The ability to localize and identify multiple people is paramount to the inference of high-level activities for informed decision-making. In this paper, we describe the PEM-ID system, which uniquely identifies people tagged with accelerometer nodes in the video output of preinstalled infrastructure cameras. For this, we introduce a new distance measure between signals comprised of timestamps of gait landmarks, and utilize it to identify each tracked person from the video by pairing them with a wearable accelerometer node.**

I. Introduction

In this paper we introduce PEM-ID (short for Proprio-Extero Matching IDentification), a system that identifies and localizes multiple people in a scene by fusing data from wearable accelerometers with tracks of people detected by a camera network. We segment each person in the scene and extract a motion signature describing landmark features of the person's gait. The same motion properties are extracted from the accelerometer node worn by each person of interest. The problem of identifying the people in the scene is then reduced to clustering the signals to obtain the matching accelerometer-to-track pairs, and using the unique ID of each accelerometer node to identify each person.

The PEM-ID system is geared toward assisted living applications, corporate environments and security. In assisted living, the identification and localization of individual people opens the doors to higher-level inference systems in multi-person homes. In corporate environments, smart badges with accelerometers can be used to track employees and visitors. This system can also be used to find security personnel moving in the field-of-view of a camera. More importantly, the PEM-ID system gives us the ability to collect experimental traces and scenarios for our research in more macroscopic behavior analysis. We designed the PEM-ID system to leverage the existing CCTV camera infrastructure, where images are typically taken from an oblique view using ceiling-mounted cameras, a configuration that has traditionally been used to maximize camera coverage. Rather than require a network of tightly-packed cameras, as most appearance-based solutions do, the PEM-ID system can identify people even if camera-nodes are far and few.

One of main the challenges tackled by PEM-ID is the extraction of uniquely-identifying information from motion paths. It is known that people move in paths that minimize the energy spent from the source point to the destination [1]. This results in short, smooth paths [2] as well as per-person preferred speeds [3] that are dependent on the person's fitness and anatomy. As a function of this, in typical planar scenarios, people walk mostly in straight lines and with nearly constant speeds. Changes in direction and speed occur primarily near source and destination points, as well as in the presence of obstacles, such as doors, walls or winding hallways. This allows us to divide all walking paths into two types of motion segments:

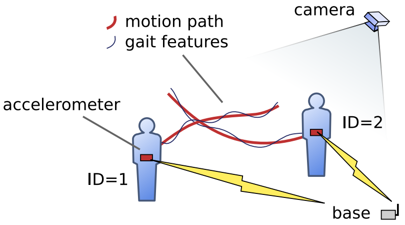

Fig. 1. Overview of the approach described in this paper: a motion signature consisting of landmark features of the person's gait are extracted from cameras and accelerometers. This is in contrast to our previous approach [4] which employed motion path features, i.e. accelerating, decelerating, turning. The gait signatures are matched between accelerometers and cameras-detected tracks. Then, the ID of the accelerometer uniquely identifies the person.

1) *Cruising segments* — long time-span motion segments with near-constant velocity.
2) *Transition segments* — short time-span motion segments where there are changes in velocity, whether in the form of tangential or centripetal acceleration.

The transition segments are characterized by changes in acceleration which, due to their low probability of occurrence in typical scenarios, often contain enough information to discriminate between different people. The greater sensing task here is to obtain such highly-discriminating features from the linear motion that is found in the more prevalent cruising segments. Furthermore, these features must be measurable from both wearable accelerometers and infrastructure cameras in realistic, uncontrolled scenarios. Finally, due to bandwidth considerations, it is desirable to utilize a representation that is compact.

To this end, we use as a personal "motion signature" the timestamps of two landmark events of each person's gait cycle: the *heel-strike* and *midswing* events. As opposed to our previous work [4], which only identified people during segments of transition motion, the motion signature used in this paper allows the identification of people during both types of motion segments (transition and cruising). In this paper, we describe a method to extract the heel-strike and midswing features from both cameras and accelerometers, and introduce a new distance metric to compare them in the presence of uncertainties in detection timestamps and time-synchronization offsets. Finally, a prototype system is used to validate the person-identification procedure. Figure 1 summarizes our system. Our assumptions in this work are that people move on a two-dimensional ground plane perpendicular to gravity, that they are walking (not running), and that they move transversal to the camera. Additionally, we assume all cameras and sensor nodes are synchronized to a few milliseconds, as our distance metric is capable of handling small time offsets in synchronization.

The main contributions of this paper are:

- The formulation of the identification problem as a minimization of the global distance between timestamps of gait landmarks from different sensing modalities.
- The development and characterization of a new distance metric to correlate sparse sequences of timestamps, and which is robust to small synchronization errors and noise.

II. Related Work

Personal identification at a distance has been traditionally done through biometrics, such as face and gait recognition. The biggest obstacles with these approaches is that they rely on the existence of a database of pre-acquired biometrics, which is not feasible to obtain in many situations such as visitor-tracking. Of the two, only gait recognition is viable at long distances for multiple people at a time. But gait recognition can be variant with clothing [5], and it has been shown that gait is susceptible to impersonation attacks [6]. A different approach to identify people is tag them with wearable nodes and utilize one of the localization approaches from the sensor network literature, typically through RF [7] and/or ultra-sound [8]. These require the installation of anchor nodes, and typically consume relatively large amounts of energy, making their deployment cumbersome and costly. Advantages of utilizing cameras instead are their relatively high localization accuracy, and pre-installed infrastructure.

Perhaps more similar to our approach are other systems that fuse anonymous sensors (in our case, cameras) with ID-bearing ones (accelerometer nodes). In [9], laser ranging and RFID-like wearable nodes were used within a particle filter to simultaneously estimate locations and identities of people in a building. In [10], an accelerometer and gyroscope were used for dead-reckoning, while a body-mounted camera corrected the location-estimation errors by recognizing geo-tagged images from a database. We ruled out a dead-reckoning approach to identifying accelerometer-carrying people in a video as it would fail in situations where people simultaneously walk in the same direction, which is common in entrances and exits. Other augmented reality and robot navigation systems have also fused cameras and inertial sensors, but typically in single-user configurations, which preclude the necessity for ID assignment that is critical in multi-user scenarios. A similar type of identification based on video and motion was reported in [11], where a robot learned to recognize itself in a mirror by correlating knowledge regarding the motion of its limbs, to the motion of objects it could see in the mirror. At a high level, our system tackles the identification of multiple people in a scene by treating it as multiple inter-connected self-recognition problems. We fuse information from exteroceptive sensors (cameras) with proprioceptive ones (accelerometers) to identify people, in a process that we call proprio-extero matching (PEM). In early experiments, we tackled the identification problem using the correlation of acceleration measurements from cameras and accelerometers. From this we found that such an approach can successfully identify people during transition segments of motion, but fails during the more common cruising segments. In this paper, we describe a gait-based signature that is present during both types of motion, and derive a new distance metric to disambiguate matching pairs of this motion signature. The literature related to the distance metric is vast, including string matching [12] as well as time-series approaches [13]. String-matching approaches such as longest common subsequence (LCS) do not consider timestamps, only sequences of symbols. Without timestamps, all biped walking sequences are equal (heel-strike, mid-swing, heel-strike, etc.) and cannot be disambiguated. Time-series approaches generally consider a densely populated sequence of uniformly sampled values, which are overkill for lightweight sensor nodes, with regards to processing demands, radio transmissions and power consumption. One possible exception is described in [14], which has similarities to our approach in its use of sparse "landmarks" of the original signals. However, all gait signals used here would be considered similar given their notion of signal similarity.

III. Problem Formulation

Given that people must go through gait cycles whenever they walk, we obtain a motion signature consisting of timestamps of specific landmark events of the gait cycle: the heel-strike and midswing instants. We acquire sequences $H = (h_1, h_2, ...)$ and $M = (m_1, m_2, ...)$ of timestamps of heel-strike and midswing events for both cameras and accelerometers. The cameras signals are referred to with a superscript C, and the accelerometer signals with an A. We denote the motion signatures for a given tracked person ℓ as $S_\ell^C = \{H_\ell^C, M_\ell^C\}$, the signature for a given accelerometer k as $S_k^A = \{H_k^A, M_k^A\}$.

The problem of identifying people from their motion signatures can, then, be defined as finding the most similar pairs of camera and accelerometer signatures. For this we must obtain the matching of S_k^A to S_ℓ^C that minimizes a global distance metric d over all k and ℓ:

$$\arg\min_{\Lambda} \sum_{k} \sum_{\ell}^{n} d(S_k^A, S_\ell^C)\Lambda_{k\ell} \tag{1}$$

where Λ is a matrix such that $\Lambda_{kl} = 1$ if accelerometer k matches track ℓ, otherwise 0.

In Section V we define and characterize a new distance metric d, which compares two sequences of sparse timestamps according to their average shift and jitter. In the next section, we begin our discussion by describing the extraction of gait landmarks from our two sensing modalities.

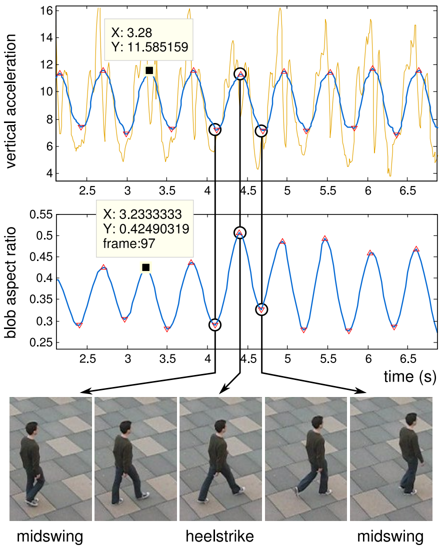

Fig. 2. Detection of gait cycle markers from both cameras and accelerometers. The upward triangles correspond to heel-strike detections, while downward triangles are used to mark the midswing detections. The selected peak (with the black square marker) displays an offset of $\sim 47ms$. Such offsets are dependent on time synchronization errors, communication delays, and quantization and sensor noise. As described in Section V, our distance metric is invariant to these offsets in typical scenarios. The distance metric automatically matches corresponding gait features from accelerometers and tracks.

IV. Extraction of Motion Signature

It is known that different people have different preferred stepping periods and step lengths — and therefore, individual preferred walking speeds ($step\ length/period$). Given the relatively low sampling rate of cameras, the stepping period by itself is not sufficient to disambiguate multiple people, while step lengths cannot be obtained from accelerometers. The stepping timestamps H and M, however, carry all the information of the stepping period biometric plus an additive stochastic component due to the uniform distribution of the instant of the person's initial step. This stochastic component makes it unlikely that two people in a scene would independently exhibit the same sequences of stepping timestamps.

The event that is typically used to mark the beginning of the gait cycle is the *heel-strike* of the swing leg. At that moment, the person's feet are both on the floor (double support phase) and at their farthest distance from one another. Also, the vertical acceleration of the impact can be clearly observed as a double-peak pattern in the person's upward acceleration, as can be seen in the unfiltered accelerometer signal in Figure 2. In addition to the heel-strike instant, we further characterize the person's gait patter by extracting the timestamp of the moment when both legs are at their closest point, or the *midswing* instant.

A. Stepping Pattern from Cameras

The heel-strike and midswing instants can be obtained from the cameras by searching for the moment when the person's feet are, respectively, at their farthest and their closest. These events can be observed as the maxima and minima of the standard deviation of the person's foreground pixels in the ground-plane direction.

To detect this, we first approximate the person's antero-posterior axis by the principal component of the distribution of foreground pixels in the blob. Then the horizontal-plane deviation is taken as the standard deviation of all pixels perpendicular to that axis.

Finally, to accommodate for people at different distances from the camera, we normalize the result by the standard deviation of the foreground pixels in the vertical direction, obtaining a type of aspect ratio measurement. Figure 2 shows the detection of midswing and heel-strike events from this signal.

B. Stepping Pattern from Accelerometers

It is known that during the gait cycle, the vertical position z of the body's center of mass reaches its maximum at the midswing, and its minimum at the heel-strike. If we approximate the vertical bobbing motion by a sinusoid $z = A\sin(\omega t)$ (where $\omega \in [1.6, 2.4]Hz$ approximately and $A \approx 5cm$ [15]), then the second derivative of z is a sinusoid that is off-phase by a half-period: $\ddot{z} = -\omega^2 A \sin(\omega t)$. Hence, we can detect the maxima (midswings) and minima (heel-strikes) in vertical position by observing the minima and maxima of the vertical acceleration, respectively.

This is shown in Figure 2, where heel-strike and midswing events are obtained from both cameras and accelerometers. Due to synchronization offsets, communication delays, and quantization and sensor noise, the detections are often displaced by $\sim 0 - 80ms$. In typical scenarios (as described in Section V) our system's results are invariant to such offsets.

V. Distance Measure

Pearson's correlation coefficient ρ is a popular measure for signal similarity. The correlation coefficient measures the covariance between two signals, normalizing it by their standard deviations:

$$\rho(A, B) = \frac{1}{N-1} \frac{\sum_{k=1}^{N}(a_k - \bar{a})(b_k - \bar{b})}{\sigma_a \sigma_b} \tag{2}$$

where $A = (a_1, ..., a_N)$ and $B = (b_1, ..., b_N)$ are time series carrying uniformly sampled data. This formulation gives ρ a well-defined range ($\rho \in [-1, 1]$, with 0 meaning no correlation), invariance to scale and offset ($\rho(mA+n, B) = \rho(A, B)$), and strong dependence on time since the signals are compared sample-wise. This last property implies that if they measure the same range of time then the input signals A, B must be sampled at the same rate — or interpolated.

In our initial experiments [4] we investigated solely the transition segments of motion and found that the correlation of two time-series of acceleration magnitudes (one from the accelerometer and one from the camera) was an effective measure of whether they originated from the same person. However, this method required there to be frequent changes in acceleration, which are not present during the more common cruising segments of motion.

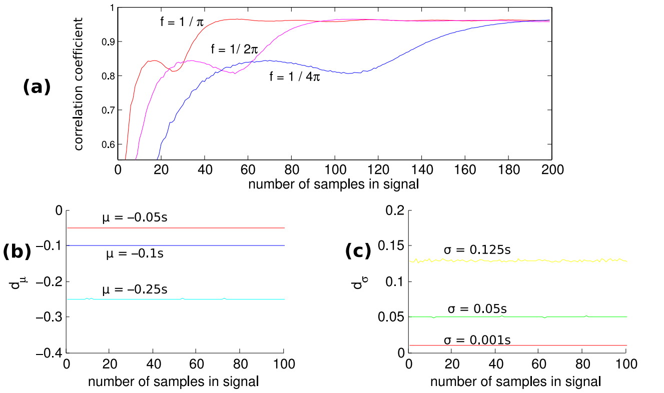

Fig. 3. (a) Correlation coefficient for two same-frequency sinusoids, where one signal was added 20% white Gaussian noise. The results were averaged over 100 instances. (b) Our distance metrics d_μ and (c) d_σ for signals consisting of timestamps at a $1Hz$ frequency but with normally distributed time offsets, averaged over 100 instances. The correlation takes tens of samples to converge, while our metrics do so instantly.

Furthermore, we found that the correlation coefficient requires a large number of samples before converging to a steady-state value, delaying the correct identification of each person in the scene. Figure 3(a) shows the correlation of two synthetic signals, as a function of the number of samples received. The two correlated signals are identical sine waves sampled at $100Hz$, except for the addition of white Gaussian noise to one of them. As can be seen in the figure, the convergence time of the correlation is dependent on the underlying signal's frequency characteristics. Higher-frequency signals converge faster than low-frequency ones. The convergence time is also negatively affected by an increase in the standard deviation of the added noise, as well as a decrease in sampling rate of the signal. All of this indicates that the correlation coefficient requires the transmission of densely-sampled signals over the wireless channel, which causes congestion and limits the total supported number wearable nodes in the scene. Instead, it is desirable to transmit only the smallest amount of samples necessary to accurately match any two signals. In our system we require only the values of the midswing and heel-strike timestamps for each step cycle, which limits the amount of transmitted information to two datapoints per cycle. This, however, makes the Pearson correlation coefficient unsuitable for our motion signatures, since ρ operates on signals consisting of sampled values rather than timestamps.

A. Designing a Distance Metric

In this section we design a distance metric that operates on sequences of timestamps such as H and M. We intend to use this metric to compare each H_k^A to each H_ℓ^C, and each M_k^A to each M_ℓ^C, for all k and ℓ, in order to infer which pairs of signals originate from the same source. This is different from classical substring-matching approaches from the data mining literature given that in our case we must compare sequences that contain *absolute time* information, as opposed to the typical string's *logical time* (i.e. ordering). In this discussion, we abstract the distance metric from its intended use with midswing and heel-strike signals, by building our arguments in terms of two generic lists of timestamps A and B. Then we are looking for a distance $d(A, B)$ which presents the following properties:

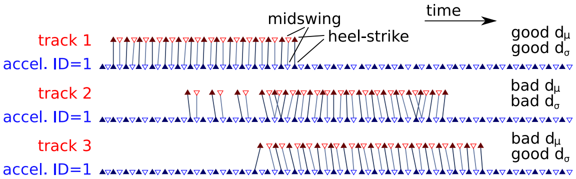

Fig. 4. Automatically matching heel-strikes and midswings to compute d_μ and d_σ in experiment from Figure 10. The matchings used to compute $D(H^C, H^A)$, $D(M^C, M^A)$ from track to accelerometer are shown as lines.

1) **Operates on sparse data**, since the timestamps are expected to be produced in the order of a couple of Hz.
2) **Converges within a small number of samples**, in order to be usable in real-world scenarios where people may be unoccluded in the FOV for short periods.
3) **Rewards coincidences** — If a camera and an accelerometer simultaneously detect a heel-strike or midswing event, then the computed similarity between the two should increase.
4) **Punishes mismatches** — If one signal measures a heel-strike and the other does not, then their similarity should decrease as a function of the distance to the nearest heel-strike.

Note that since the motion signature does not explicitly carry sampled measurement values, such a metric is inherently scale and offset-invariant. In the remainder of this section we describe two distance metrics d_μ and d_σ which measure the average distance and average deviation between two signals A and B of timestamps.

To satisfy the requirements listed above we compute the distance between two lists of timestamps $A = (a_1, ..., a_{|A|})$ and $B = (a_1, ..., a_{|B|})$ in the following manner. If we define the offset between two timestamps as $d(a_i, b_j) = a_i - b_j$ (and not the other way around), we can define the offset between a timestamp a_i and a list of timestamps B as the difference between a_i and the $b_j \in B$ that is closest to it (in terms of absolute value):

$$d(a_i, B) = a_i - \arg\min_{b_j \in B}(|a_i - b_j|) \tag{3}$$

Note that the offset can be (and often is) negative, and that the offset between a timestamp and a list that contains it is zero: $d(a_i, A) = 0 \ \forall \ a_i \in A$.

We, then, define the list of element-wise offsets between the timestamps in A to the list B:

$$D(A, B) = (\, d(a_1, B),\, d(a_2, B),\, \ldots,\, d(a_{|A|}, B)\,) \tag{4}$$

where the capital letter D is used instead of d to underline that this function produces a list, rather than a scalar. Figure 4 shows the matching between timestamps for experimentally-acquired heel-strike signals ($D(H_k^A, H_\ell^C)$) and midswing signals ($D(M_k^A, M_\ell^C)$). The offset between each pair of timestamps is their horizontal distance.

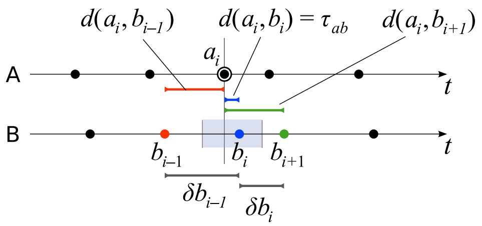

Fig. 5. The distance $d(a_i, B)$ between a point a_i in A and the list of points B is defined as the closest distance between a_i and any point in B. In the case where $B = A + \tau_{ab}$ and $\tau_{ab} < \frac{1}{2}\min\{\delta b_1, \delta b_2, \ldots\}$, then $d(a_i, B)$ will always be equal to τ_{ab}.

Note that $|D(A,B)| = |A|$ and that $D(\cdot,\cdot)$ is not symmetric ($D(A,B) \neq D(B,A)$). This can be clearly seen when $|A| \neq |B|$, which implies $|D(A,B)| \neq |D(B,A)|$, and so $D(A,B) \neq D(B,A)$. This asymmetry is undesirable as it allows mismatching timestamps to be entirely ignored by D in some situations, violating our requirement to punish mismatches. We curb this by defining

$$D'(A,B) = D(A,B) \cup -D(B,A) \tag{5}$$

where $\cup$ represents list concatenation followed by re-sorting.

Finally, we define the distance metrics d_μ and d_σ between A and B as the mean and standard deviation of D':

$$\boxed{d_\mu(A,B) = \text{mean}(D'(A,B)) = \sum_{v \in D'(A,B)} \frac{v}{|D'(A,B)|} \quad (6)}$$

$$\boxed{d_\sigma(A,B) = \text{std}(D'(A,B)) = \sqrt{\frac{\sum_{v \in D'(A,B)} (v - \bar{v})^2}{|D'(A,B)| - 1}} \quad (7)}$$

1) Behavior of Metrics given Time Offsets: Consider the case where A and B are signals obtained from different sensing modalities, but regarding the same person (i.e. they are matching signals). Let B constitute a time-shifted clone of A such that $B = A + \tau_{ab}$, where τ_{ab} is a scalar representing the time synchronization offset between A and B, and where there is no clock skew.

We claim that the $d_\mu(A,B)$ is guaranteed to perfectly measure time offset (τ_{ab}) between A and B, as long as the following is true:

$$\tau_{ab} < \min(\Delta A)/2 = \min(\Delta B)/2 \tag{8}$$

where ΔA is the list of inter-timestamp intervals of signal A (and similarly for ΔB) as defined by:

$$\begin{aligned} \Delta A &= (\, d(a_1,a_2),\, d(a_2,a_3),\, \ldots,\, d(a_{|A|-1},a_{|A|})\,) \\ &= (\, \delta a_1,\, \delta a_2,\, \ldots,\, \delta a_{|A|-1}\,) \end{aligned} \tag{9}$$

If (8) is true, then for any $a_i \in A$ the closest point in B is guaranteed to be b_i (same subscript index i) denoted by $b_i = a_i + \tau_{ab}$, with $d(a_i, b_i) = \tau_{ab}$. This is because any other point $b_j \in B$ is either at the right of b_i (and so $b_j \geq a_i + \tau_{ab} + \min(\Delta B)$) or at the left ($b_j \leq a_i + \tau_{ab} - \min(\Delta B)$). See Figure 5. If the former, then (assuming $\tau_{ab} > 0$ with no loss of generality) $|a_i - b_i| = \tau_{ab} < \tau_{ab} + \min(\Delta B) = |a_i - b_j|$. If the latter, then since $\tau_{ab} < \min(\Delta A)/2$ we have $|a_i - b_i| = \tau_{ab} < \min(\Delta B)/2 < |\tau_{ab} - \min(\Delta B)| = |a_i - b_j|$. Similarly, a_j will be the closest point in A to any point b_j.

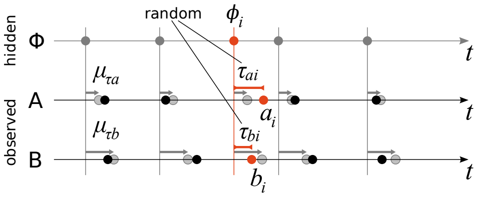

Fig. 6. Definition of symbols for Section V-A2. The time offsets τ_{ai} and τ_{bj} $\forall i, j$ are Normal random variables with mean $\mu_{\tau a}$ and $\mu_{\tau b}$ (shown with gray arrows and gray circles) and standard deviation $\sigma_{\tau a}$ and $\sigma_{\tau b}$.

Thus, (8) guarantees that the distances $d(a_i, B)$ and $d(b_j, A)$ all equal τ_{ab}, and $d_\mu(A,B) = \text{mean}(D'(A,B)) = \text{mean}(\tau_{ab}, \tau_{ab}, \ldots) = \tau_{ab}$, which proves our claim. Similarly, since $\text{std}(\tau_{ab}, \tau_{ab}, \ldots) = 0$, the value of d_σ in the noiseless case is 0 — indicating that A and B are perfectly-shifted versions of one another with no jitter. In the next section we analyse the behavior of d_μ and d_σ under the presence of additive timestamp noise (jitter).

2) Behavior of Metrics given Random Timestamp Noise: Instead of the noiseless relation $B = A + \tau_{ab}$ used in the previous section, where the time shift τ_{ab} was a constant, in this section we consider the time shift to be a random variable. The rationale for this is that in reality the time shift varies with factors such as signal propagation delays, sensor noise and sampling/quantization noise. We show, however, that the expected values of d_μ and d_σ approach the idealized values from Section V-A1, under realistic assumptions (similar to assumption (8)).

A complete description of the relation between two matching signals A and B must consider: (1) the true (hidden) state Φ of which A and B are observations, (2) the time offset of A and B from the true time in Φ, and (3) false positive and false negative timestamps. In this paper we consider a simplification by assuming the probability of false positives and false negatives is approximately 0. Our experiments show that this approximation does not negatively impact the results from this section, since the experimental number of FPs and FNs is low. We to future work leave an analysis of these effects, as well as an in-depth discussion of the sources of noise.

Taking these factors into consideration, relation between A and B can then be written as:

$$A = \Phi + T_A \qquad B = \Phi + T_B \tag{10}$$

where T_A, T_B are lists of independent random variables representing Normally-distributed timestamp offsets:

$$\begin{aligned} T_A &= (\,\tau_{a1}, \tau_{a2}, \ldots\,) \quad \tau_{ai} \sim \mathcal{N}(\mu_{\tau a}, \sigma^2_{\tau a}) \\ T_B &= (\,\tau_{b1}, \tau_{b2}, \ldots\,) \quad \tau_{bj} \sim \mathcal{N}(\mu_{\tau b}, \sigma^2_{\tau b}) \end{aligned} \tag{11}$$

As depicted in Figure 6, τ_{ai} and τ_{bj} are the individual time shifts from ϕ_i to a_i and b_j respectively. The time shifts jitters around their mean values $\mu_{\tau a}$ and $\mu_{\tau b}$ (shown as gray circles).

Equation (10) implies that all $a_i \in A$ and $b_i \in B$ are randomly-shifted versions of some $\phi_i \in \Phi$. Given that there are no false positives or false negatives, whenever an a_i and b_i originated from the same ϕ_i (i.e. they are *true matches*), they share the same subscript index. That is, $a_i = \phi_i + \tau_{ai}$ and $b_i = \phi_i + \tau_{bi}$.

Then consider the following condition on the statistical distributions of T_A, T_B and Φ:

$$\mathcal{P}(|\tau_{ai} - \tau_{bj}| < |\delta\phi|/2) \approx 1 \tag{12}$$

Then, equation (12) guarantees that for all $a_i \in A$ the closest point in B will always be that between a_i and its the true match b_i (and vice versa). That is:

$$\begin{aligned} d(a_i, B) &= a_i - \arg\min_{b_j \in B}(|a_i - b_j|) = \\ &= \phi_i + \tau_{ai} - \phi_i - \tau_{bi} = \tau_{ai} - \tau_{bi} \end{aligned} \tag{13}$$

Regarding the applicability of assumption (12) in our real-world scenario, for stepping signals $\delta\phi \approx 1/(1.2\, to\, 2.8Hz)$. This implies jitter $|\tau_{ai} - \tau_{bi}|$ of up to around $178.6ms$ satisfy these requirements, which is an order of magnitude larger than some of the noisiest systems.

Similar to the noiseless case, the condition in (12) assures that $D'(A, B)$ is composed solely of distances between true matches:

$$D'(A, B) = (\ \tau_{a1} - \tau_{b1}, \tau_{a2} - \tau_{b2}, \dots\ ) \tag{14}$$

Given that all the τ_{ai} and τ_{bi} are independent normal random variables, their difference $\tau_i^\star = \tau_{ai} - \tau_{bi}$ is also Normally distributed. Therefore, $d_\mu(A, B)$ and $d_\sigma(A, B)$ are the mean $\mu^\star$ and standard deviation $\mu^\star$ of random time offset $\tau_i^\star$ between the event timestamps in A and those in B:

$$d_\mu(A, B) = \text{mean}(\{\tau_i^\star\}_{\forall i}) = \mu^\star = \mu_{\tau a} - \mu_{\tau b} \tag{15}$$

$$d_\sigma(A, B) = \text{std}(\{\tau_i^\star\}_{\forall i}) = \mu^\star = \sqrt{\sigma_{\tau a}^2 + \sigma_{\tau b}^2} \tag{16}$$

This can be seen in Figure 7, which portrays the behavior of our distance metrics when faced with random time shifts. In the figure, we consider the case where Φ is a $1Hz$ signal and $T_A = 0$ (i.e. $\Phi = A$) and vary the parameters $\mu_{\tau b}$ and $\sigma_{\tau b}$ governing the distribution of T_B. This simulates the effect of varying $\mu^\star = 0 + \mu_{\tau b}$ and $\sigma^\star = 0 + \sigma_{\tau b}$. As predicted, for small magnitude of noise (offset and jitter) d_μ and d_σ are good approximations of $\mu^\star$ and $\sigma^\star$.

In addition, d_μ and d_σ satisfy our requirements for fast convergence, as as shown in Figure 3(b) and (c). As a function of the sample size, the values of d_μ and d_σ converge to their expected values instantly, while other metrics such as the correlation coefficient require larger intervals.

B. A Note Regarding Complexity

An important aspect of the distance metrics d_μ and d_σ is that they can be computed in linear time by taking into the account the natural ordering of the incoming data. Since there are well-known recursive methods of computing the mean and standard deviation of signals online [16], the only possible computational bottleneck is the calculation of the minimal distance in equation (3), which must be minimized over all combinations of two items from the sets $D(A, B)$ and $D(B, A)$. However, it can be shown that it is possible to compute $D(A, B)$ with complexity $O(|A| + |B|)$ by traversing the input signals A and B in chronological order. Below is the sketch of an algorithm that achieves this, as it traverses A and B at most once:

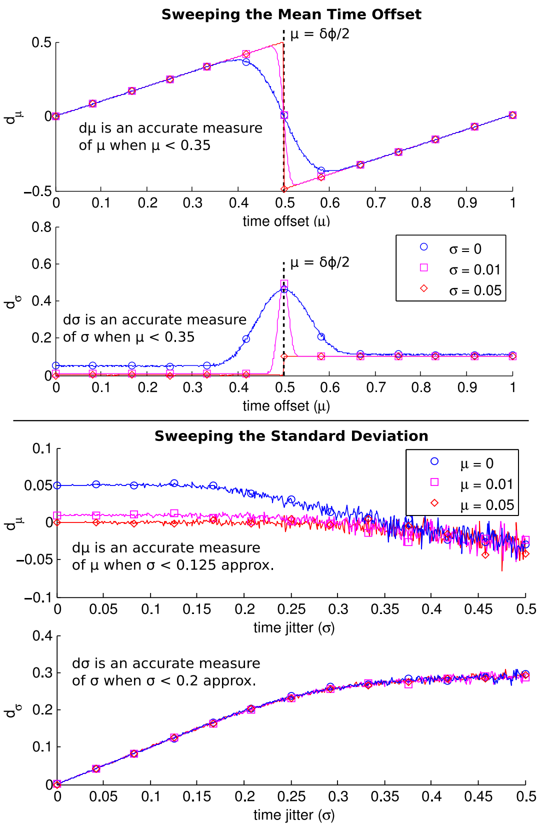

Fig. 7. Behavior of d_μ and d_σ as we sweep $\mu^\star$ and $\sigma^\star$. When condition 12 is true, $d_\mu \approx \mu^\star$ and $d_\sigma \approx \sigma^\star$.

```
function D($A, B$)
    $distances \leftarrow [\,], i \leftarrow 1, j \leftarrow 2$
    $B \leftarrow (-\infty) \cup B \cup (\infty)$
    while $i \le |A|$
        if $d(a_i, b_j) < d(a_i, b_{j-1})$ and $d(a_i, b_j) < d(a_i, b_{j+1})$ then
            append $d(a_i, b_j)$ in $distances$
            $i \leftarrow i + 1$
        else j $\leftarrow j + 1$
```

VI. Multiple-Person Identification Algorithm

In the previous sections we derived and characterized two distance metrics, d_μ and d_σ, which provide a measure of the average shift and jitter between timestamps in two sequences. In order to identify people in the scene, it is necessary to combine those two distances into a single metric. Thus, given two motion signatures $S_k^A = \{H_k^A, M_k^A\}$, $S_\ell^C = \{H_\ell^C, M_\ell^C\}$, we compute the linear combination of the shift and jitter distances, d_μ and d_σ:

$$\begin{aligned} d(S_k^A, S_\ell^C) \ &= \ d_\mu(H_k^A, H_\ell^C) \ + \ \alpha\, d_\sigma(H_k^A, H_\ell^C) \\ &\quad + \ d_\mu(M_k^A, M_\ell^C) \ + \ \alpha\, d_\sigma(M_k^A, M_\ell^C) \end{aligned} \tag{17}$$

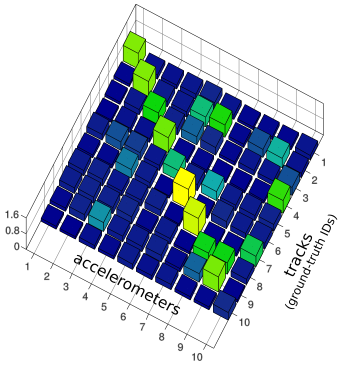

Fig. 8. Similarity matrix comparing each accelerometer signal to each track in a random permutation of 10 experimental traces. For clarity, the value of each bin in the figure is the normalized inverse of the distance metric (hence, taller bars denote higher similarity). Note that item $(10, 10)$ in the diagonal does not represent the best match in its row or column, as would be expected. This situation is rectified by selecting the assignments that optimize the global similarity over all assignments, as described in Section VI.

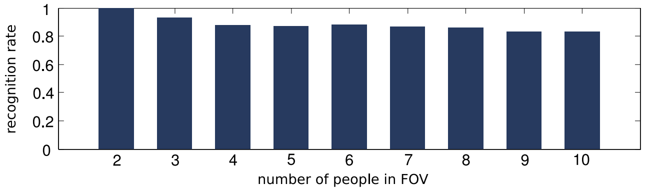

Fig. 9. Average correctness of the ID assignments that result from applying our distance measure to experimentally acquired motion paths. Using different permutations of data from one-person experiments we simulate scenarios of up to 10 people in the FOV, and display the average of 10 such simulations for each datapoint in the figure. The identification rates are above 85% in all cases.

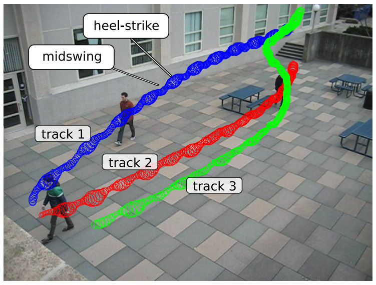

Fig. 10. Instance of multiple-person experiment showing three extracted tracks. The radii of the circles represent the blob aspect ratio at each instant. The PEM-ID system correctly identified Track 1 as corresponding to the single accelerometer-carrying person in the scene. Videos can be found at http://enaweb.vimeo.com/drupal/PEM-ID-videos.

where the α is an experimentally-determined parameter that weights the importance of the time deviations d_σ compared to d_μ. In all our experiments, we use $\alpha = 2$.

Given m tracks and n accelerometers, a distance matrix Ω is constructed where each cell $\Omega_{k\ell}$ is the distance between the motion signature of the k^{th} accelerometer and that of the ℓ^{th} track — i.e. $\Omega_{k\ell} = d(S_k^A, S_\ell^C)$. Figure 8 shows an example of such a matrix where $m = n = 10$, obtained from a permutation of 10 experimentally-acquired traces, as described in the evaluation section (Section VII). The matrix in the figure displays the similarity between the motion signatures (the inverse of their distance, normalized by the maximum) to provide a clearer picture.

If people were identified by simply picking the best value for each row or column of Ω, there could occur multiple matches for the same accelerometer or track. This is undesirable, since each person carries at most one accelerometer, and no accelerometers are shared. This can also lead to situations such as the one in Figure 8, where the best match for accelerometer 10 is not found to be track 10, as would be expected.

For these reasons, we instead identify people based on the *best global assignment* of accelerometers to tracks, as described in equation (1). Our goal is, then, to find the match matrix Λ that minimizes the sum of matched distances, as described in equation (1). We solve (1) through the Munkres assignment algorithm [17], which yields the optimal solution in polynomial time $O(\max(m, n)^3)$. In order to allow accelerometers and cameras to go unmatched (in case there are people who are not carrying a wearable node, for example), we set to ∞ all distances in Ω that are above a threshold value (we used $d > 2.5$ in our experiments).

VII. Evaluation

In order to clearly separate the performance of the matching and identification algorithm from artifacts originating from the multiple-person tracker employed in our prototype, we divide our evaluation into two sets of experiments. All experiments were recorded with a 640×480 camera at $30 fps$. Data from a SparkFun 6DoF IMU carrying a MMA7260Q accelerometer, sampling at $100 Hz$, was collected serially to a small computer attached to the person's body.

In the first set, a single person walks across the FOV of the camera for 12 experimental runs, changing the source and destination points each time. Given the physical constraints of the experiment, the camera captured approximately 10 steps before the person left its FOV. When outside the FOV, the person's steps were captured only by the wearable accelerometer. This was repeated for two different people, resulting in 24 data sets. The person's location was obtained from each video frame by averaging the x and y coordinates of the foreground pixels after a simple background-subtraction step. This allowed the simulation of multiple-person scenarios by considering different permutations of the experimental data. The resulting simulations are often quite challenging datasets, since the walking frequencies obtained from the same person tend to be close to their preferred stepping frequency. In real-world scenarios, however, this would be counter-balanced by the person's stepping phase. We simulate this by randomizing the moment of the first step with a uniformly-distributed time-offset of up to 1 second that is added to each simulated trace. Under these conditions, the distance metric and global optimization procedure result in average recognition rates of over 85% for scenarios with up to 10 people in the scene at a time, as shown in Figure 9. With 2 people in the scene, we obtained an average recognition rate of 100%.

To demonstrate the application of the system to actual multiple-person situations, where there may be tracking errors and occlusions, we recorded a second set of experiments.

In these experiments, three people crossed the FOV of the camera, for 8 experimental runs, each time entering/leaving from different directions. The PEM-ID system was used to system identify the only person in the scene that was wearing an accelerometer node. Figure 10 shows an example frame from one of these multiple-person experiments overlaid with each person's detected track. The blob aspect ratio is depicted in the radii of the circles in the image. Each person in the scene was segmented using a straight-forward background subtraction and connected-component analysis step, and then finally tracked with a mean-shift tracker. The average recognition rate for these experiments was 87.5%. Since only one of the three people in the scene was carrying an accelerometer, the global optimization procedure from equation (1) does not have any impact. If all three people were carrying accelerometers, it is expected that the global optimization step would increase the recognition rate above 90%, as found in the permutation experiments (Figure 9). Hence, the more people in the FOV wear accelerometer nodes, the better the expected recognition rate. Example videos of the PEM-ID system running can be found at http://enaweb.vimeo.com/drupal/PEM-ID-videos.

VIII. Discussion

The limiting factor for the number of people that can be concurrently identified in PEM-ID is the sampling rate of the sensors, which affect the precision of the time measurements. With cameras at $30fps$, and assuming τ_{ab} is at most $1frame$ ($\pm 33ms$), the system should be able to identify at most 10 people walking at the same stepping frequency. However, since people's preferred stepping frequencies are typically different, given enough time it should be possible to disambiguate between them.

Our identification system relies on the low probability of two different people presenting synchronized stepping patterns in real-world scenarios. If the people are in the FOV for a short time, however, this probability increases, and the system becomes less viable. The longer the time when the person is under camera coverage, the better the performance of the algorithm — thus, by using multiple cameras to increase coverage it should be possible to obtain better recognition rates. The probability of occurrence of synchronized stepping patterns also increases with the number of people in the scene. To mitigate this problem, it is possible that the rough proximity of each person to anchor nodes can be approximately measured through RSSI or other means, and used in conjunction with the stepping timestamps. Similarly, additional features such as the direction of each person's motion (captured from wearable magnetometers) can be leverage for further disambiguation.

IX. Conclusion

We have described a system that identifies people using the timestamps of gait landmarks from cameras and accelerometers. The system has diverse applications such as identifying security personnel and separating data from multiple-person deployments into single-person traces. The system uses a new distance metric to compare sequences of timestamps, finding their relative time shift and jitter. Our experiments show that the system is able to correctly identify people over 85% of the time, even in 10-person scenarios. Additional precision in more crowded situations may be possible by incorporating additional constraints such as walking direction and rough proximity to anchor nodes. This as well as an expansion to a multiple-camera system are left for future work.

Acknowledgments

This work was partially funded by the National Science Foundation under projects ECCS 0622133, IIS 0715180 and CNS 0721632. Any opinions, findings and conclusions or recommendation expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

References

[1] Wayne T. Willis, Kathleen J. Ganley, and Richard M. Herman, "Fuel oxidation during human walking," *Metabolism - Clinical and Experimental*, June 2005.

[2] David C. Brogan and Nicholas L. Johnson, "Realistic human walking paths," in *CASA '03: Proceedings of the 16th International Conference on Computer Animation and Social Agents (CASA 2003)*, Washington, DC, USA, 2003, p. 94, IEEE Computer Society.

[3] John E. A. Bertram and Andy Ruina, "Multiple walking speed-frequency relations are predicted by constrained optimization," *Journal of Theoretical Biology*, vol. 209, no. 4, pp. 445 – 453, 2001.

[4] Thiago Teixeira, Deokwoo Jung, Gershon Dublon, and Andreas Savvides, "Identifying people in camera networks using wearable accelerometers," in *Pervasive Technologies Related to Assistive Environments (PETRA)*, 2009.

[5] L. Lee and W.E.L. Grimson, "Gait analysis for recognition and classification," May 2002, pp. 148–155.

[6] Davrondzhon Gafurov, Einar Snekkenes, and Patrick Bours, "Spoof attacks on gait authentication system," in *IEEE Transactions on Information Forensics and Security*, 2007.

[7] Branislav Kusy, Akos Ledeczi, and Xenofon Koutsoukos, "Tracking mobile nodes using RF doppler shifts," in *SenSys '07: Proceedings of the 5th international conference on Embedded networked sensor systems*, New York, NY, USA, 2007, pp. 29–42, ACM.

[8] A. Savvides, C.C. Han, and M. B. Srivastava, "Dynamic fine grained localization in ad-hoc sensor networks," in *Proceedings of the Fifth International Conference on Mobile Computing and Networking, Mobicom 2001, Rome, Italy*, July 2001, pp. pp. 166–179.

[9] Dirk Schulz, Dieter Fox, and Jeffrey Hightower, "People tracking with anonymous and id-sensors using rao-blackwellised particle filters," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2003.

[10] M. Kourogi and T. Kurata, "Personal positioning based on walking locomotion analysis with self-contained sensors and a wearable camera," *Mixed and Augmented Reality, 2003. Proceedings. The Second IEEE and ACM International Symposium on*, pp. 103–112, Oct. 2003.

[11] K. Gold and B. Scassellati, "Deictic pronoun learning and mirror self-identification," in *6th International Conference on Epigenetic Robotics (EpiRob)*, 2006.

[12] Dan Gusfield, *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press, 1 edition, 1997.

[13] Robert H. Shumway and David S. Stoffer, *Time Series Analysis and Its Applications: With R Examples*, Springer, 2 edition, 2006.

[14] Chang-Shing Perng, Haixun Wang, Sylvia R. Zhang, and D. Stott Parker, "Landmarks: A new model for similarity-based pattern querying in time series databases," in *In ICDE*, 2000, pp. 33–42.

[15] J.P. Paul, "History and fundamentals of gait analysis," *Bio-Medical Materials and Engineering*, vol. 8, pp. 123–135, 1998.

[16] Donald E. Knuth, *The Art of Computer Programming*, vol. 2: Seminumerical Algorithms, Addison-Wesley Professional, 1998.

[17] Harold W. Kuhn, "The hungarian method for the assignment problem," in *Naval Research Logistic Quarterly*, 1955, vol. 52.