Identifying People in Camera Networks using Wearable Accelerometers

Thiago Teixeira, Deokwoo Jung, Gershon Dublon and Andreas Savvides
Yale University
10 Hillhouse Ave
New Haven, CT
thiago.teixeira@yale.edu

ABSTRACT

We propose a system to identify people in a sensor network. The system fuses motion information measured from wearable accelerometer nodes with motion traces of each person detected by a camera node. This allows people to be uniquely identified with the IDs the accelerometer-node that they wear, while their positions are measured using the cameras. The system can run in real time, with high precision and recall results. A prototype implementation using iMote2s with camera boards and wearable TI EZ430 nodes with accelerometer sensorboards is also described.

Categories and Subject Descriptors

C.3 [**Special-Purpose and Application-Based Systems**]: Real-time and embedded systems; I.4 [**Image Processing and Computer Vision**]: Scene Analysis—*Sensor fusion, Tracking*

General Terms

Measurement, Design, Experimentation

Keywords

Unique identification, Consistent labelling, Association problem

1. INTRODUCTION

A large obstacle to the deployment of assisted-living systems in multiple-person or family homes is the problem of differentiating between people and uniquely identifying them in order to properly attend their individual needs. For this reason, much of the current assisted-living technology focuses on single-person scenarios — and often break as soon as visitors are invited into the home. Additionally, multiple-person homes present complex privacy requirements for assistive technologies, in the sense that only those who voluntarily choose to utilize the system people should have any private information captured and stored. Meanwhile, cameras are becoming increasingly popular sensors in assistive environments, given their long sensing range and ability to measure distinct information modalities (such as location, pose, motion path and ambient lighting). However, the problem of associating detected people across multiple image frames as well as robustly identifying them solely with visual features is still a topic of much research in computer vision.

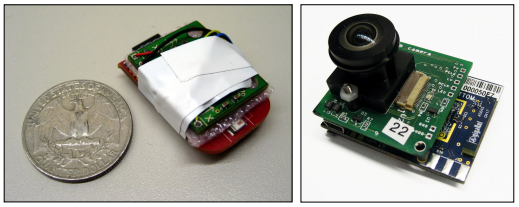

Figure 1: Sensor nodes used in the prototype system. Left: TI EZ430-RF2480 ZigBee node with accelerometer sensorboard. Right: Intel Mote2 with custom camera board.

In this paper we present a real time system that identifies people in camera networks with high accuracy through the use of wearable accelerometer nodes with known IDs (Figure 1). We bypass the computer vision correspondence problem by matching local motion signatures from wearable accelerometers with those observed from infrastructure cameras. This way, we are able to obtain the location of each person using the camera detections and to estimate their identities by matching their motion characteristics. The advantage of this approach is that it provides reliable operation at reduced cost for assistive applications. Instead of relying on expensive video analytics to identify people, we make use of very limited information from the cameras and construct a unique modality pair by coupling cameras and accelerometers through wireless links. An overview of this process is shown in Figure 2, and more detail is provided in Sections 3 and 4. As described in the evaluation section (Section 5), the system runs in real time with high precision and recall metrics. We start the paper with a discussion regarding related work (Section 2) following by description of the problem in Section 3.

2. RELATED WORK

Accelerometers and cameras are often combined to track the camera motion, generally for use in robot navigation [1],

PETRA'09 June 9-13, 2009, Corfu, Greece

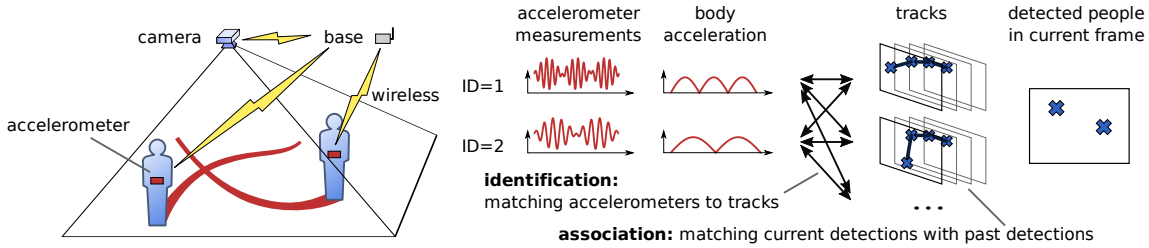

Figure 2: To associate IDs with each detected people, we find the best match between measured body acceleration and each person's motion in the image plane. To obtain this motion information, *tracks* must be formed from combinations of detected people in the image sequence. Since the number of tracks grows exponentially, the main issue that must be solved is how to keep the track count down.

and virtual or augmented reality [2]. In such cases, the accelerometer is placed on the same rigid object as the camera, which moves in relation to its environment. This contrasts with the setup described in this paper, where the camera is stationary and accelerometers are placed on the moving people in the camera's field-of-view (FOV) in order to identify them.

In the literature, people in sensor networks have been localized and identified using wearable sensors such as ID badges [3][4], the combination of ultrasound with radios [5], radios signal properties [6], and inertial sensing units [7]. Ultrasound-based approaches require bulky nodes that consume relatively large quantities of energy. Although ID badges present low spatial resolution when used by themselves, motion models [3] or additional sensing modalities can be used to improve spatial accuracy [4]. Radio signal strength localization is subject to many random factors in uncontrolled environments, such as antenna orientation [8]. Radio Doppler-shift has been used to localize targets [6], but require a large number of infrastructure nodes. In [7], accelerometers and magnetometers are used along with ID sensors. This approach requires the knowledge of the building map to constrain the location of the particles in their particle filters. One similarity among many of the multi-person solutions is that they require an exact association between detected people from one frame to people detected in the next. This problem is known to be NP-hard [9]. One of the seminal works in this area is the multiple-hypothesis tracking algorithm [10]. When only the position of each detected person is used to perform this association, this is called the motion correspondence problem and is the subject of much research [11]. Other times, additional image features (such as size, color, shape or motion gradient) [12][13] or motion models are used to offer additional clues regarding frame-to-frame associations, but usually with mixed results in uncontrolled environments. In contrast, the algorithm presented here is lightweight, does not make assumptions about motion models, and does not require an exact solution of the association problem as input.

3. PROBLEM DESCRIPTION

The problem we solve in this paper is the matching of the locations of people detected with a camera network to their accelerometer signals, in order to obtain location-ID pairs. The core of the identification problem can be described as finding the matching between accelerometers and detected locations that maximizes a similarity measure. Therefore, if Z_k is the set of all accelerometer measurements at time k, and X_k is the set of all detected locations, then at each time k we must find the match matrix M_k according to the expression below:

$$\arg\max_{M_k} \sum_{i=1}^{|Z_k|} \sum_{j=1}^{|X_k|} f(z_k^i, x_k^j) M_k^{ij} \tag{1}$$

where z_k^i is the i^{th} accelerometer measurement at time k, and x_k^j is the j^{th} detected position at that time. Note that the index i of the accelerometer measurements is the ID of the nodes that transmitted them, while the j's are random internal IDs of each detected person without actual physical relevance. The match matrix M_k is a matrix of size $|Z_k| \times |X_k|$ describing the associations between accelerometers and detected people in the image frame. Since the same person cannot be wearing two accelerometers, and the same accelerometer cannot be in more than one place at a time, M_k must follow a few constraints:

$$M_k^{ij} = \begin{cases} 1 & \Rightarrow \ z_k^i, x_k^j \text{ are associated} \\ 0 & \Rightarrow \ \text{no association} \end{cases} \tag{2}$$

$$M_k^{ij} = 1 \Rightarrow \begin{cases} M_k^{i\ell} = 0 \ \forall \ \ell \in [1, |Z_k|], \ell \neq j \\ M_k^{\ell j} = 0 \ \forall \ \ell \in [1, |X_k|], \ell \neq i \end{cases} \tag{3}$$

Despite the brief definition, the problem that is targeted in this paper cannot be solved by directly associating accelerometers to detected people as described in Equation 1. Instead, the two types of measurements (accelerations and positions) must be brought to a common representation in order to be compared, which, as will be described later, leads to an exponentially complex problem. As shown in Figure 2, our solution is divided into two parts:

Identification — In order to match the motion data from the wearable accelerometers with detections from the camera nodes, we transform each into a signal that is proportional to the person's floor-plane acceleration. We, then, measure their similarity by computing their correlation coefficient. This is described in Section 3.1. To obtain obtain these acceleration measurements from the detected locations, however, we must first obtain a time-series of locations for each person in the scene (tracks). This is called the multi-dimensional association problem, and it is known to be NP-hard [9].

Association — Rather than solving the association problem, we make use of the fact that correct tracks must belong to real people in the scene, and therefore must correlate with *some* accelerometer (exactly one, in fact). We use this to

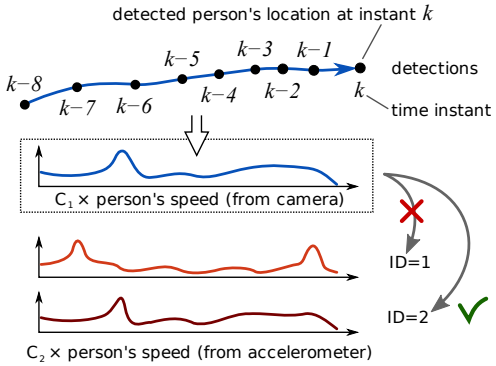

Figure 3: Base case described in Section 3.1, where a single person is in the camera's field of view, while two accelerometers are within communication range. Signals proportional to the person's speed are compared, and the best matching accelerometer is found to have $ID = 2$.

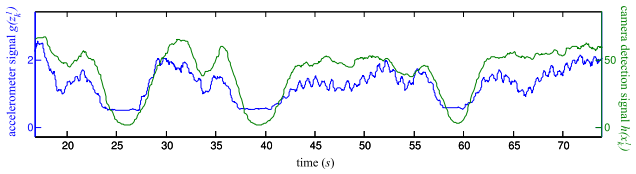

Figure 4: Superimposition of aligned signals from accelerometer and camera, showing the approximate proportionality between them.

approximate the multi-dimensional association problem as a one-dimensional association with polynomial complexity. This is described in Section 3.2.

In the following subsections we describe two base-case scenarios from which our solutions to the identification and association stages are derived.

3.1 Base-Case for Identification Problem: 1 Person in FOV, 2 Accelerometers

Consider the scenario where there is a single person in the camera FOV while two accelerometer nodes can be heard through the wireless channel (Figure 3). If it is known a priori that there is only one person in the FOV, then it is simple to create a time-history of the person's locations, as there are no frame-to-frame association ambiguities: the person detected in image frame I_k at time k is always associated to the person detected in frame I_{k+1}. This reduces the association problem to a trivial step and allows us to focus solely on the identification. This section describes the process by which we compare and match signals in order to *identify* people in the FOV.

Position measurements from the camera contain instantaneous information with an absolute frame of reference in space, and with no association among previous measurements (no frame of reference in time). Meanwhile, acceleration measurements obtained from the body-mounted accelerometer nodes have no spatial frame of reference, but have a clearly defined temporal frame of reference. To find the similarity between these two signals, we must first convert them into a common representation, or intermediary format. This process is known as data alignment [14]. We align the two signals to the same temporal frame of reference by associating all position measurements that belong to the same person into a time series. From this time series, the person's acceleration in the image plane can be easily extracted by double differentiation, as shown in Figure 3. We also align the accelerometer measurements into a signal proportional to the overall body acceleration by calculating the magnitude of the 3D acceleration vector and finding the envelope of the signal to remove noise caused by the stepping motion and by accelerometer-bouncing artifacts [15][16]. Figure 4 shows the similarity between two matching signals that were processed in this manner. These two signals are proportional to the person's floor-plane acceleration, and, therefore, also proportional to one another.

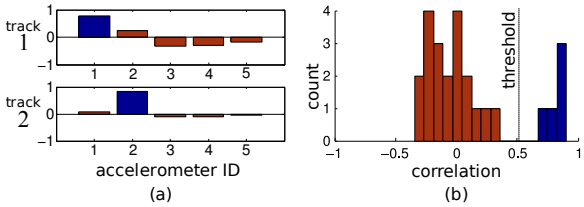

Figure 5: (a) Correlation between 5 tracks and 2 accelerometer signals, showing the large difference between correct and incorrect matches. (b) Histogram with sampling distribution of correlation coefficient. A clear threshold separates correct matches from incorrect ones.

If α and β are the functions which align accelerometer and camera signals into the same common representation, then the similarity $g(\cdot,\cdot)$ between the two signals can be calculated by detecting whether the two signals are proportional using Pearson's correlation coefficient r:

$$g(z_k^i, \theta_k^\ell) = r(\alpha(z_k^i), \beta(\theta_k^\ell)) \tag{4}$$

where θ_k^ℓ is a *track* containing a time series of consecutive person detections $\theta_k^\ell = (x_{k-n}^{\ell_0}, ..., x_k^{\ell_n})$ with $n \in \mathbb{N}$ and $0 < n < k$. Figure 5(a) shows the experimental value of the correlation coefficient between 5 tracks and 2 accelerometer signals. The correct matches can be easily seen by their strong correlations.

Note, however, that $g : Z_k \times \Theta_k \mapsto \mathbb{R}$ from Equation 4 has a different domain than the similarity function f from Equation 1. To assign IDs to detected people using g, the maximization problem in Equation 1 must be modified to use tracks rather than person detections:

$$\arg\max_{\Omega_k} \sum_{i=1}^{|Z_k|} \sum_{\ell=1}^{|\Theta_k|} g(z_k^i, \theta_k^\ell)\Omega_k^{i\ell} \tag{5}$$

where $\Theta_k = \{\theta_k^\ell\}$ is the set of all tracks at instant k, and Ω_k is a match matrix associating accelerometer signals to tracks. The matrix Ω_k follows similar rules as M_k (Equation 3) but it additionally does not allow the same detected person to be assigned to multiple tracks *at any time instant*. So Ω_k must follow the additional rule that for any two elements equal to 1, the corresponding tracks must have an empty intersection:

$$\Omega_k^{i\ell_1} = \Omega_k^{i\ell_2} = 1, \ell_1 \neq \ell_2 \Longrightarrow \theta_k^{\ell_1} \cap \theta_k^{\ell_2} = \emptyset \tag{6}$$

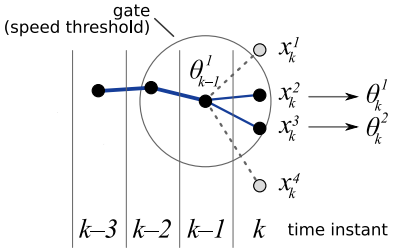

Figure 6: Limiting number of track hypotheses through gating. When a single detection is within the gate, identification can be done directly (Section 3.1). Otherwise, association must take place (Section 3.2).

We call this the "strong no-intersection" property, which will be relaxed in Section 4 in order to approximate the solution for real time operation.

3.2 Base-Case for Association Problem: 2 People in FOV with Accelerometers

In the previous section we outlined a signal-comparison method to identify people given a trivial base scenario. In the same vein, in this section we will employ a base-case scenario to describe a method by which accelerometer measurements can be used to influence and simplify association decisions. To understand the underlying problem, consider the situation where it is known that exactly two people are in the FOV, and they are within a large distance of one another. Then it is easy to infer their position histories (association) by creating tracks connecting each detection at frame k to the only detection at $k+1$ that is within a physically plausible speed threshold — a process known as *gating*.

However, if the two people approach one another (Figure 6) tracking ambiguities arise, giving rise to multiple competing track hypotheses. If all possible track hypotheses are considered by the tracking algorithm, then due to combinatorial explosion the complexity of the problem quickly becomes unmanageable. This is shown in Figure 7, where two people meet for 6 time instants (at 15 frames per second this corresponds to $0.4s$) generating more than 64 hypotheses. If it is known that there are exactly N people in the FOV, then the number of hypotheses after K ambiguous frames is $N!^K$. If the people are allowed to enter/leave, and a realistic detector is assumed (with the possibility of false positive detections), then the number is even larger.

This association problem can be described as selecting the set of tracks that, at each time instant k, globally minimizes some distance metric h:

$$\arg\min_{\Phi_k} \sum_{\ell=1}^{|\Theta_{k-1}|} \sum_{j=1}^{|X_k|} h(\theta_{k-1}^{\ell}, x_k^j)\Phi_k^{\ell j} \tag{7}$$

where Φ_k is a match matrix that follows the same rules as M in Equation 3 (which causes the constructed tracks to naturally follow the strong no-intersection rule from Equation 6). From Φ_k, the set of tracks $\Theta_k^\star$ which solves Equation 7 for time instant k is directly obtained. The simplest similarity metric for track-to-location association is the Euclidean distance between the track's latest location $x_{k-1}^{\ell_n}$ and the detection x_k^j:

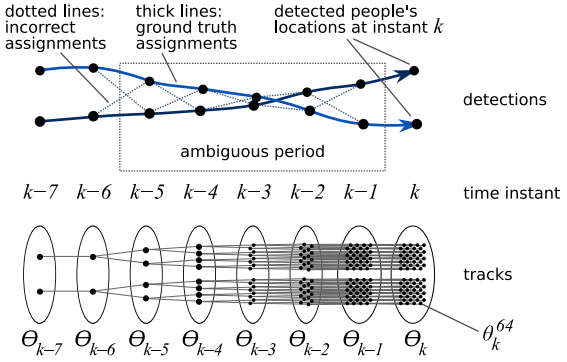

Figure 7: Base-case for Section 3.2. After this 6-sample-long ambiguity, the number of track hypotheses grows as an exponential of a factorial.

$$h(\theta_{k-1}^{\ell}, x_k^j) = \text{dist}(x_{k-1}^{\ell_n}, x_k^j) \tag{8}$$

where n is the track length. This is often called nearest-neighbor association.

As described before, it is not tractable to exactly solve Equation 7 due to the expansive number of tracks. Luckily, to identify the people in the scene it is not necessary to solve this complex association problem, since the no-intersection property is handled later in the process by the maximization in Equation 5. So we bypass this problem by generating several conflicting *track hypotheses* (Θ_k), rather than finding the best non-conflicting solution ($\Theta_k^\star$). The set Θ_k of track hypotheses is defined to contain all tracks that pass a goodness criterion:

$$\Theta_k = \left\{ \theta_k^{\ell} \in \Theta_{k-1} \times X_k \;:\; \begin{array}{l} h(\theta_{k-1}^{\ell}, x_k^j) < \tau_\theta, \\ \exists\, z_k^i \in Z_k \mid g(z_k^i, \theta_k^{\ell}) > \tau_r \end{array} \right\} \tag{9}$$

where τ_θ and τ_r are thresholds that filter out bad hypotheses. Thus, only tracks within the gate are considered ($h(\cdot,\cdot) < \tau_\theta$). Here, the similarity measure g is used in a manner analogous to the use of additional image attributes (size, color, shape) and motion models that are usually employed in multiple-target trackers. In this case, we keep only the tracks that can be explained to some degree by at least one of the accelerometer signals. This is described in greater detail in Section 4.2. Of course, image and motion attributes from the literature can be used *in addition to* the accelerometer signal, for increased robustness if necessary.

4. PERSON-IDENTIFICATION ALGORITHM

As our person-identification formulation is composed of two interconnected parts (association and identification), we design our algorithm as a cycle consisting of two blocks: a tracker and a comparator.

The **tracker** generates a set Θ_k of tracks from sequences of person detections, filtering them according to the parameters that rely solely on track properties (i.e. the τ_θ filter from Equation 9). The comparator is in charge of performing the maximization in Equation 5 taking Θ_k as input, and pruning tracks that do not pass the τ_r filter.

The **comparator** then passes the set Θ_k' of filtered tracks

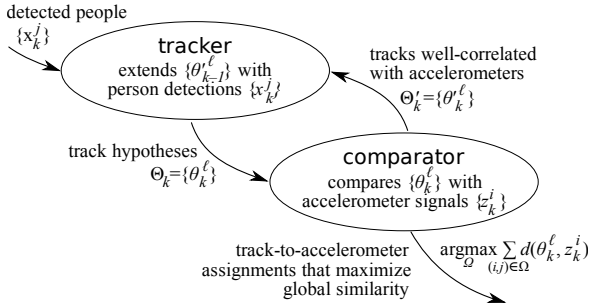

Figure 8: Overview of the entire algorithm showing the interactions of its two main logical blocks: a *tracker* generates a small set of track hypotheses from the large pool of possible associations; and a *comparator* solves the association problem described in Equation 5, assigning IDs to each detected people.

back into the tracker as input. The output of the algorithm is found by using the Hungarian method [17] to solve the one-dimensional association problem from Equation 5 with complexity $O(\max(|X_k|, |\Theta_k|)^3)$.

This cycle is the core of our method, and is summarized in Figure 8.

Although in the problem description we discussed the similarity between each track and each accelerometer, $g(z_k^i, \theta_k^\ell)$, this is not exactly how things take place in the identification algorithm. Instead, each track is marked as belonging to a single accelerometer, which is the only one it will ever be compared to. The reason for this is that the correlation coefficient requires the two input signals to be of the same length. When each track is created, we must bootstrap the sufficient statistics to compute its correlation with each specific accelerometer. This also allows us to keep the complexity low, since tracks that have historically not correlated well with a given accelerometer can be pruned and never compared to that accelerometer again.

Other than this, the algorithm takes two main approaches to allow real time operation: (1) it simplifies the accelerometer-to-track assignment problem in Equation 5 by weakening the track no-intersection property; (2) it restrains the number of track hypotheses to a minimum through several means.

4.1 Relaxing the No-Intersection Property

Since our algorithm aims to provide the best immediate results without the intention to reconstruct past traces, we relax the strong no-intersection constraint of Equation 6 to require only that *the newest* position measurements in each matched track do not intersect. That is, the following weak no-intersection constraint is used instead:

$$\Omega_k^{i\ell_1} = \Omega_k^{i\ell_2} = 1, \ell_1 \neq \ell_2 \Longrightarrow x_k^{j_1} \in \theta_k^{\ell_1} \neq x_k^{j_2} \in \theta_k^{\ell_2} \quad (10)$$

Although this relaxes the strong no-intersection property of Equation 6, the similarity measure g used in the identification (Equation 5) guarantees that tracks correlate well with their matched accelerometer. So, as long as the motion of the people in the scene is not too similar and synchronized with one another, most tracks selected by Equation 10 will still be strongly non-intersecting. In the case that their motion is correlated, then it is not possible to identify them based on motion characteristics alone, whether strong no-intersection is enforced or not. Hence, this simplification has little negative effect on the quality of the tracks, while greatly limiting the problem's complexity.

4.2 Adjustments to Control the Number of Hypotheses

Combinatorial Contention — When there are ambiguous situations, such as in Figure 7, the number of tracks grows exponentially. In order to contain this growth, we only resolve ambiguities *after* the people move apart. For this, the algorithm keeps track of the number of people inside each track's gate (a circle of radius R). If the number is greater than one, then the track is marked as being ambiguous. Otherwise, it is marked as unambiguous. Each ambiguous track θ_{k-1}^ℓ gets extended into time k as θ_k^ℓ by assigning it the closest detection x_k, rather than forking into one track for each within-gate detection. When a track transitions from ambiguous to non-ambiguous, however, it is forked for each detection inside a gate with radius $2R$. If N_{2R} is the number of people in the $2R$ gate, then, instead of ending up with $N!^K$ tracks as before, each track splits into just N_{2R} alternatives, most of which are pruned within a few seconds by a track-pruning process.

Pruning Tracks and Allowing "Leaving" — If a track correlates badly with all accelerometer signals, then it cannot belong to an accelerometer-wearing person, and should be pruned. Figure 5(b) shows a histogram of the correlation values of correct and incorrect accelerometer-to-track assignments. It is clear from the plot that the two can be easily distinguished, and that a threshold value $\tau_r \approx 0.55$ can be used for this purpose. However, as shown in Figure 9(a), the correlation r between an accelerometer and a track takes a few seconds to converge. Oftentimes the correct accelerometer-track association has a poor correlation ($< \tau_r$) for the first few seconds, which can cause correct tracks to be prematurely pruned.

For this reason, we compute the estimated correlation error as a function of track age by using confidence intervals. But since Pearson's correlation coefficient does not have a Gaussian sampling distribution, we must first convert it with Fisher's z' transformation, for which confidence intervals can be calculated:

$$z'(r) = \tfrac{1}{2} \ln[(1+r)/(1-r)] \quad (11)$$

The standard error of z' is known to be $SE = 1/\sqrt{n-3}$, where n is the number of samples used in the computation of the correlation. With this, we compute the 90% confidence interval of f as ranging from z'_{low} to z'_{high}:

$$z'_{low}(r) = r - \tfrac{1.645}{\sqrt{n-3}} \qquad z'_{high}(r) = r + \tfrac{1.645}{\sqrt{n-3}} \quad (12)$$

where the number 1.645 comes from the 90% confidence interval of a Normally distributed random variable (i.e. 90% of the density is within 1.645 standard deviations from the mean). Equation 9 is, then, modified to apply the τ_r threshold on z'_{high} instead. That is, $\alpha(\cdot) > \tau_r$ becomes $z'_{high}(\cdot) > z'(\tau_r)$. This way, the only tracks that get pruned are those where there is a 95% confidence that the track does not correlate above τ_r (95% because the threshold acts on a single-sided confidence interval). For comparison, Figure 9(b) shows the z' and confidence intervals for the signals from Figure 9(a).

Since correlations of longer signals have a smaller standard error, they are inherently more trustworthy. We, therefore,

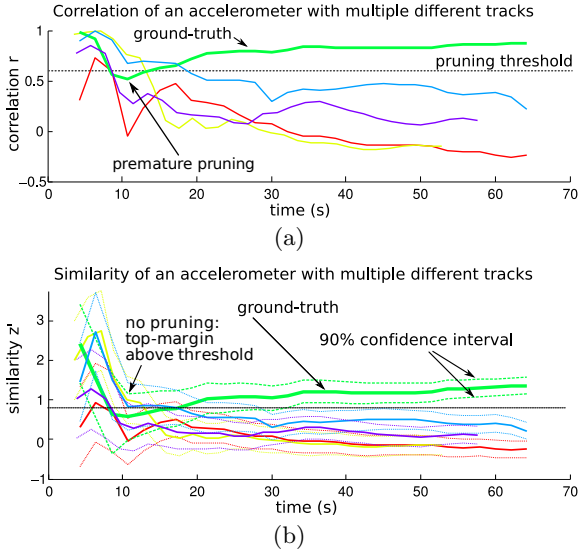

Figure 9: Top: Since the correlation takes time to converge, the use of a fixed threshold for track-pruning can result in the correct track being prematurely deleted. Bottom: We guard against premature pruning this by applying the threshold on the top margin of the 90% confidence interval, resulting in 95% confidence pruning.

prioritize longer tracks by using z'_{low} instead of r in Equation 4. So if two tracks have the same r (and, hence, the same z') the older track will be given a higher weight in g since the z'_{low} will be higher for the older track. With this change, Equation 4 becomes:

$$g(z_k^i, \theta_k^\ell) = z'[\, r(\alpha(z_k^i), \beta(\theta_k^\ell))\,] - 1.645/\sqrt{n-3} \qquad (13)$$

Note that, as the standard error cannot be computed for tracks smaller than 4 samples, we only allow a track to be pruned if its size is greater than 4. Given that new tracks are created at the end of each ambiguous period, this causes the number of tracks to depend on the number of ambiguities.

Faster Error Recovery and Allowing "Entering" — When a new person enters the camera FOV, a new track must be created for comparison with each accelerometer. Similarly, when a new accelerometer is detected, it must be included for comparison with each existing track. For this reason, the algorithm always keeps at least one track for each accelerometer-location combination. If one does not exist, it is created. This can happen either because a new person or accelerometer has been detected ("entering") or because an existing track has been pruned. The end result is that tracks that may or may not represent a correct ground-truth trace are constantly created (and constantly pruned, if they do not pass the τ_r threshold). This ensures that there is always one alternative for each accelerometer-location assignment, which allows for quick recovery in case a correct track becomes associated with the wrong detection due to tracking errors. This puts a lower bound of $|Z_k| \times |X_k|$ on the number $|\Theta_k|$ of track hypotheses at any time k. When there are no ambiguities, the track-pruning process ensures that the lower-bound is reached. Hence, for most real-world cases, it is expected for the average number of track hypotheses to be close to $|Z_k| \times |X_k|$.

count	2	3	4	5
precision	0.951	0.875	0.790	0.627
recall	0.956	0.931	0.887	0.821
proc. time (s)	2.57	6.29	11.02	16.74
ambig./pers.	38.25	68.57	93.90	116.6
avg. tracks	3.92	8.81	15.63	24.39
max. tracks	8	15	24	30

Table 1: Experimental results for algorithm when 2, 3, 4 or 5 people are in the FOV at the same time.

5. EVALUATION

We first performed a set of experiments where data was gathered with a wide-angle USB camera and an off-the-shelf inertial measurement unit. These were used to verify the correctness of the algorithm independently from implementation-dependent effects, such as the performance of the person detector or of the network layer. A second set of experiments were performed using the iMote2 sensor node with our custom camera board [18], as well as TI EZ430-RF2480 nodes equipped with a SparkFun IMU 5DOF board, containing an Analog Devices ADXL330 accelerometer (Figure 1). The purpose of these is to demonstrate the viability of the system in actual multiple-person deployments. For all of these experiments, the cameras were mounted on a $3m$ high ceiling, facing down. This gives a total area of $3m \times 2m$ where people are entirely contained in the FOV. This is the area within which the people were asked to stay. The accelerometer nodes were placed on the front of each person's belt. The orientation of the accelerometer is unimportant, given that it is the magnitude of the 3D acceleration vector that is used in the similarity metric.

We captured five separate videos and the corresponding accelerometer traces of a single person walking in a room for approximately 1 minute. The person detector used in this experiment computed the person in the scene by comparing each frame to an image of the empty room (background subtraction). Since the traces were captured separately in a static, controlled environment, we were able to obtain high precision image-plane coordinates for each person by calculating the center of mass (centroid) of the foreground pixels. The accelerometers were sampled at $100Hz$, and the camera at $15Hz$. Time was roughly synchronized by hand, by visually matching the features from an acceleration magnitude plot for each accelerometer to a plot of the corresponding centroid's speed.

We ran the algorithm for all different 2-person, 3-person, 4-person and 5-person combinations of the five traces. The centroid traces were overlaid onto the same image plane and the centroids' internal index were randomly shuffled for each frame. We additionally simulated people entering and leaving the field of view at random times while still being in range for the accelerometer sensing. This was done by randomly cropping the beginning and end of the centroid traces, and leaving the accelerometer traces intact. For all of these, the ground truth frame-by-frame associations and absolute person IDs were known, given that the traces were acquired separately. Using the ground truth data, we calculated the following metrics: **Precision** answers the question: when the system identifies a person, how often is the ID assignment correct? The precision is calculated as $TP/(TP+FP)$, where TP is the number of true positives (correct assign-

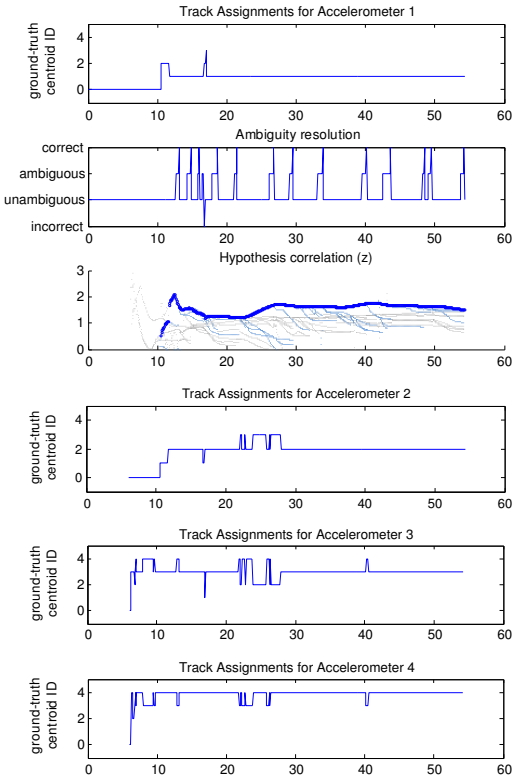

Figure 10: Output of the algorithm for a 4-person trace. The x-axis is time in seconds. The ambiguity resolution and hypothesis correlation plots for accelerometers 2, 3 and 4 are omitted to conserve space.

ments) and FP is the number of false positives (incorrect assignments). **Recall** answers the question: when a given person is in the scene, how often does the system correctly identify him/her? The recall is calculated as $TP/(TP + FN)$, where FN is the number of false negatives, that is, the number of times the person was deemed absent when they were actually present.

The averaged experimental values of these two metrics are shown in Table 1. The algorithm shows strong performance for 2, 3 and 4 people. For 5 people, the precision falls under 0.7, but the recall stays high throughout. The measured processing time is, as expected, proportional to the average number of tracks. Since the number of tracks stays within the predicted value of $|Z_k| \times |X_k|$, the processing time remains nearly constant. Also note that the system always executed many times faster than real-time.

The number of ambiguous frames per person ($ambigs/per$) is also reported. Were the tree-pruning process not present, the expected number of tracks would be in the order of $|X_k|^{ambigs/per}$. This is at least $2^{38.25} = 3.26 \times 10^{11}$ for the 2-person case, and much more for the others. The correctness of the algorithm has a stronger dependence on the $ambig/person$ rate than on the number of people in the FOV per se.

Figure 10 shows the output of the algorithm for an example trace where data for 4 people were overlaid. The track assignment plots show the ground-truth ID of the centroid that was associated by the algorithm to each accelerometer. For an accelerometer with $ID = A$ (where A is some integer), it is desirable for the plot to be a constant line at $y = A$. This is often the case after enough time has passed for the correlations to converge, as seen in the plots. Meanwhile, the ambiguity resolution plot (shown only for person 1, due to space restrictions in this paper) shows how often ambiguities occur (usually consisting of multiple frames at a time), and whether the algorithm is able to resolve them correctly or incorrectly. On average, ambiguities were correctly resolved 80.72% of the times. For the remaining 19.28% when the ambiguity resolution failed, the algorithm eventually found the correct assignment through the correlation metric. That is, the algorithm is able to automatically recover from incorrect hypotheses. Finally, the third type of plot in the figure, the hypotheses correlation plot, shows the z' metric of the selected hypothesis (thick blue line) compared to that of the losing hypotheses for the same accelerometer (light blue). The hypotheses for other accelerometers are shown in light gray. Note how after ambiguous periods small tracks fork from the correct one. They are quickly pruned by the combinatorial contention process described in Section 4.2. You can find videos of these experiments at `http://enaweb.eng.yale.edu/drupal/InertialIdentification`.

To assess the viability of the system as an online sensor network service, we also tested a prototype implementation consisting of an iMote2 camera node mounted on the ceiling, and two people carrying wearable EZ430 sensor nodes with accelerometers. The centroid of each foreground blob was extracted by segmenting them through 8-neighbor connected component analysis. As expected, this often resulted in the typical blob-merging and splitting artifacts that are a product of small occlusions and visual similarity with the background scene. Detections were collected into packets containing pairs of centroids and timestamps, and transmitted wirelessly to a base node. The whole process took place at a rate of around $15Hz$ in the sensor node, fluctuating based on the number of people in the FOV. The wearable nodes used in the experiment were programmed to sample the accelerometer at a rate of $50Hz$, calculating the signal envelope locally, and transmitting it to the base through its ZigBee radio. The collected data was then parsed in a nearby computer, resulting in precision and recall measurements comparable to those in Table 1. This prototype system demonstrates that it is possible to identify people using acceleration and camera measurements under non-ideal real-world sensing conditions (including false positives, false negatives and other types of misdetections) as well as under the constraints of limited local processing and networking bandwidth.

6. CONCLUSION

We have presented a system that uses infrastructure camera nodes and wearable accelerometers to identify people in a sensor network, achieving good precision and recall. Other than memory and processing requirements, there is no limit on the number of tag-wearing people or the number of people in the FOV. We have also described a set of approximations that allow for real-time execution. Although these approximations increase the number incorrect matches immediately following ambiguous periods, experimental results show the algorithm is able to quickly recover.

Possible improvements include utilizing additional image

features for increased robustness against ambiguities. By coupling this system with color histograms, for example, better detection rates should be easily achieved. Future work includes expanding the algorithm to make use of multiple cameras as a single seamless sensor, as well as considering deployments where there are large gaps in camera coverage. Before the system can be used in a long-term deployment, power consumption and network utilization must be properly analyzed. To this end, it is possible that an adaptive sampling and transmission scheme can be devised, which preprocesses accelerometer samples and only transmits them if it is deemed that they can significantly impact the correlation metric.

Acknowledgments

This work was partially funded by the National Science Foundation under projects CNS 0448082 and CNS 0725706. 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).

7. REFERENCES

[1] KyuCheol Park, Dohyoung Chung, Hakyoung Chung, and Jang Gyu Lee, "Dead reckoning navigation of a mobile robot using an indirect kalman filter," *Multisensor Fusion and Integration for Intelligent Systems, 1996. IEEE/SICE/RSJ International Conference on*, pp. 132–138, Dec 1996.

[2] M.S. Keir, C.E. Hann, J.G. Chase, and X.Q. Chen, "A new approach to accelerometer-based head tracking for augmented reality & other applications," Sept. 2007, pp. 603–608.

[3] N. Shrivastava, R. Mudumbai U. Madhow, and S. Suri, "Target tracking with binary proximity sensors: fundamental limits, minimal descriptions, and algorithms," in *SenSys '06: Proceedings of the 4th international conference on Embedded networked sensor systems*, New York, NY, USA, 2006, pp. 251–264, ACM Press.

[4] 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.

[5] 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.

[6] 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.

[7] Lasse Klingbeil and Tim Wark, "A wireless sensor network for real-time indoor localisation and motion monitoring," in *IPSN '08: Proceedings of the 2008 International Conference on Information Processing in Sensor Networks (ipsn 2008)*, Washington, DC, USA, 2008, pp. 39–50, IEEE Computer Society.

[8] D. Lymberopoulos, Q. Lindsey, and A. Savvides, "An empirical analysis of radio signal strength variability in ieee 802.15.4 networks using monopole antennas," in *under submission*, April 2005.

[9] Lawrence D. Stone, Carl A. Barlow, and Thomas L. Corwin, *Bayesian Multiple Target Tracking*, Artech House Publishers, 1999.

[10] Donald B. Reid, "An algorithm for tracking multiple targets," in *IEEE Transactions on Automatic Control*, December 1979, vol. 24, pp. 843–854.

[11] C. J. Veenman, M. Reinders, and E. Backer, "Resolving motion correspondence for densely moving points," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, March.

[12] Omar Javed, Zeeshan Rasheed, Khurram Shafique, and Mubarak Shah, "Tracking across multiple cameras with disjoint views," in *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, Washington, DC, USA, 2003, p. 952, IEEE Computer Society.

[13] M. Taj, E. Maggio, and A. Cavallaro, "Multi-feature graph-based object tracking," in *CLEAR*, 2006, pp. 190–199.

[14] L. Wald, "Definitions and terms of reference in data fusion," in *International Archives of Photogrammetry and Remote Sensing*, 1999.

[15] A. Forner-Cordero, M. Mateu-Arce, I. Forner-Cordero, E. Alcántara, J. C. Moreno, and J. L. Pons, "Study of the motion artefacts of skin-mounted inertial sensors under different attachment conditions," in *Physiological Measurement*, April 2008, vol. 29, pp. 21–31.

[16] Thong Y.K., Woolfson M.S., Crowe J.A., Hayes-Gill B.R., and Challis R.E., "Dependence of inertial measurements of distance on accelerometer noise," *Measurement Science and Technology*, vol. 13, pp. 1163–1172(10), 2002.

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

[18] T. Teixeira and A. Savvides, "Lightweight people counting and localizing in indoor spaces using camera sensor nodes," in *ACM/IEEE International Conference on Distributed Smart Cameras*, September 2007.