Tasking Networked CCTV Cameras and Mobile Phones to Identify and Localize Multiple People

Thiago Teixeira, Deokwoo Jung, and Andreas Savvides
Yale University
10 Hillhouse Ave
New Haven, CT
{firstname}.{lastname}@yale.edu

ABSTRACT

We present a method to identify and localize people by leveraging existing CCTV camera infrastructure along with inertial sensors (accelerometer and magnetometer) within each person's mobile phones. Since a person's motion path, as observed by the camera, must match the local motion measurements from their phone, we are able to uniquely identify people with the phones' IDs by detecting the statistical dependence between the phone and camera measurements. For this, we express the problem as consisting of a two-measurement HMM for each person, with one camera measurement and one phone measurement. Then we use a maximum *a posteriori* formulation to find the most likely ID assignments. Through sensor fusion, our method largely bypasses the motion correspondence problem from computer vision and is able to track people across large spatial or temporal gaps in sensing. We evaluate the system through simulations and experiments in a real camera network testbed.

Author Keywords

Localization, Person Identification, Cameras, Inertial sensors

ACM Classification Keywords

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis – *Sensor Fusion, Tracking*; I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms

Algorithms, Design, Experimentation

INTRODUCTION

Smart environments of the future are poised to revolutionize the interactions between people and computer systems by infering intents and behaviors. Cameras are strong candidates for this purpose, as they can not only localize multiple people but also detect poses and interactions of each person with the surrounding environment, objects, and other people.

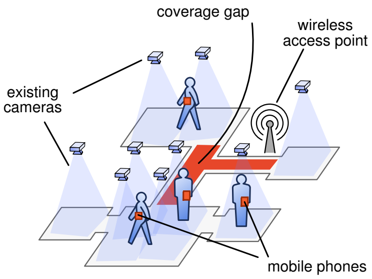

Figure 1. System overview: a network of overhead cameras is used to detect and localize people, and inertial sensors on people's mobile phones are used to identify them.

However, when multiple people are involved, it is of utmost importance to consistently label each one, even in the presence of ambiguities and spatial or temporal gaps in sensing (e.g. when a person leaves for a lunch break and returns one hour later). This is known in the computer vision literature as the correspondence problem. Its main challenge lies in the rank of the solution space, which grows faster than exponentially as a function of time and number of people.

In this paper we propose a method to bypass these issues and, furthermore, uniquely identify each person in the scene. For this, we leverage local measurements from accelerometers and magnetometers — two inertial sensors that are increasingly popular on devices such as the Apple iPhone 3GS, the HTC Hero, the Motorola Droid and the Google Nexus One. The proposed method works as follows: People's locations are detected using existing CCTV cameras, and are made available over the air through WiFi or other infrastructure (Figure 1). Then the phones of users within range read measurements from their onboard inertial sensors to pick the most probable location out of the stream. This location estimate may conflict with the estimates made by other phones. Hence, the phones may, then, send their results back to the camera network where a global optimization procedure is executed to pick the best non-conflicting phone-to-location assignments. This way, our system inherently provides two core services: (1) *identification*, since the people detected in the video streams become labeled with their phones' unique ID; and (2) *localization*, as each phone learns its own location in the process. The focus of our evaluation is on the identification service, given that the position estimates of the localization service are forwarded directly from the output of

UbiComp '10, Sep 26-Sep 29, 2010, Copenhagen, Denmark.

the CCTV camera network, which is taken as a black box.

In sum, the proposed method focuses on finding the association between a phone's ID and the person's location, detected by the camera network. This is done by using the phone's local motion measurements as the "glue" between the two. This approach exhibits three main advantages: (1) it reuses existing infrastructure; (2) it can identify/localize people even in the presence of ambiguities and sensing gaps; and (3) it automatically provides each person the anonymous location of all surrounding people.

The main contributions of this paper are:

1. The derivation of a probabilistic data-association framework to assign IDs to location measurements by introducing a set of helper measurements for which association is known. We model this as a hidden Markov model (HMM) where the main unknown is not the state estimate — which can be found using one's preferred tracking technique — but instead to discover the underlying association between measurements.

2. The use of this framework to localize and identify people in existing camera networks by leveraging the inertial sensors present on their mobile phones. In this paper we perform a proof-of-concept validation of the localization service using our existing testbed of iMote2 camera sensor nodes.

RELATED WORK

Locations and IDs can be extracted in a number of different ways. Wearable devices simplify the process of target detection and motion correspondence since targets make an effort to identify themselves. The most prominent and widely available localization technology today is undoubtedly GPS. However, GPS is limited by coarse spatial resolution and by signal attenuation when indoors, which frequently leads to complete failure to localize. For indoor situations, multiple methods exist in the literature, employing RF and ultrasound signaling properties such as propagation delays [22], Doppler-shifts [14] or signal absorption [36]. The main setbacks of these technologies are interference from multipath signals [14], variability to antenna orientations [18], and the need for a large infrastructure of anchor nodes [36]. Increasingly popular these days are the approaches based on RF fingerprinting [37], but these require a laborious training process and can only attain lower resolutions.

An alternate approach is to detect and track the targets using external sensors, such as a camera network. The main disadvantage of cameras is that tracking multiple targets across multiple cameras is a highly complex open problem, typically encompassing camera calibration/registration, and visual and/or motion models in order to track targets across different views. And, due to ambiguities when people cross paths, the number of different track combinations grows exponentially in time and becomes unmanageable without heuristic track-pruning methods. Furthermore, unique identification with cameras (such as through gait or facial recognition) requires extensive training for each person to be identified, and tend to fail given large ID databases.

Our intention in this work is to build upon the wide availability of camera networks and mobile phones to provide the best of both worlds, wearable- and external-sensing. This combines the advantages of cameras (the ability to localize people with a small number of infrastructure nodes) with those of wearable nodes (simple person identification and tracking over large sensing gaps by using the node's ID). Although the fusion of cameras and inertial sensors has previously been considered for motion tracking [29][23], augmented reality [12], and SLAM (simultaneous localization and mapping) [20], the work presented in this paper bears more similarity to the multiple-target tracking literature. The reason is that the main problems that we tackle are data association and motion correspondence, rather than exact position estimation. In sharp contrast, in motion tracking the data association problem is generally bypassed through the use of color markers. Similarly, in augmented reality and SLAM the inertial sensors are placed at known locations on the camera, and therefore need not be localized. Perhaps the most relevant fusion approach is the work by Schulz et al. [24], which combines a dense network of infrared/ultrasound ID sensors to identify tagged people as they are detected by a laser ranger. Their formulation, like ours, is based on identifying anonymous location measurements using ID-carrying measurements. However, our solutions deviate on four principal levels: (1) We consider the issue of *identification* separately from that of *localization*, which greatly reduces the state space of the problem. (2) This, when combined with our bipartite graph matching solution, allows our method to execute in real time, differently from their particle-filter approach. Of course, the tracks labeled by our method can, later, be processed with any state-of-the-art tracking technique to produce higher-precision location estimates. (3) In addition, the approach in [24] cannot recover from losing the correct ID hypotheses. (4) And finally, while the implementation of their system would require the widespread installation of unorthodox sensors, we emphasize the pressing need to reuse existing infrastructure, by employing ubiquitous CCTV cameras and mobile phones. Finally, the key difference between this work and our previous accelerometer-based person-identification research is that here we make no assumption regarding people's motion paths (as is the case in [31], where the acceleration is implicitly assumed to vary frequently), and can identify people even if they are partially occluded (as opposed to [30]) so long as the camera network can still detect them. In addition, the work presented here uses an entirely new probabilistic formulation that is general to ID-assignment problems and which can be seamlessly extended to other sensing modalities by simply modifying the state vector and emission probabilities.

PROBLEM FORMULATION

As discussed in the previous section, cameras can cheaply and unobtrusively detect and localize people within their field-of-view. However, since the *identity* of each person detected by a camera is not known, it becomes challenging to track people when there are path discontinuities (due to crossing, entering, leaving, clutter, occlusions, etc.). Indeed, the

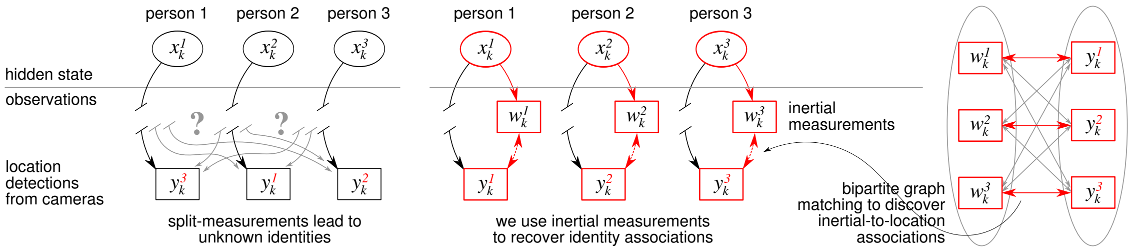

Figure 2. Main idea: to use inertial measurements from wearable nodes (with known ID) as a "glue" between the location-detections from a camera network and people's IDs.

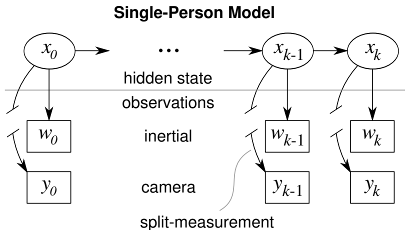

Figure 3. Independence diagram of relating the variables that pertain to the same person. The broken arrows are used to indicate that the camera observations do not carry the person's ID (i.e. they are *split measurements*).

anonymity of a camera's detections also means people cannot be uniquely identified nor, equivalently, localized. To this end, we propose the use of motion measurements to uniquely identify each person, according to the following formulation sketch:

- Given anonymous location measurements from a camera network, and inertial measurements (acceleration and direction) from mobile phones of known ID,
- Find the $(location, inertial)$ association pairs which maximize the likelihood that the measurement pair originated from the same person.

We model this as a missing data problem where each person is observed twice: once from the viewpoint of the camera network, and once from the inertial sensors on the mobile phones. What makes this problem distinct from traditional missing data problems is that, in addition to unknown true states, we also lack knowledge regarding the data association of location observations. This is shown in Figure 2. We define the term *split measurements* to denote the observations which do not have a known association (the broken arrows in the figure).

Below we describe the problem within a more formal framework. In this discussion we assume the extrinsic calibration parameters of the cameras have been computed a priori using a method such as [9] or [1], and thus camera placement is known.

Suppose a number of people are present within the sensor network's coverage area at timestep k. Let person i's true state at k be represented by the variable x_k^i. In our implementation, x_k^i is composed of four components (s_x, s_y, $|\ddot{\mathbf{s}}|$, $\ddot{\mathbf{s}}_{yaw}$) consisting of x, y position coordinates, acceleration magnitude and acceleration yaw. Nevertheless, the derivation that follows applies to any other state definition. We consider the evolution of x_k^i in discrete time to be indexed by $k \in \mathbb{N}^*$. Since it is widely accepted that human motion can be approximated by a Markov process, we assume x_{k-1}^i is sufficient to predict x_k^i.

Let $\beta_k \in \mathbb{N}$ be the number of people detected with the camera network at time k. Then we can denote a detection instance by y_k^j (where j is in the set $\{1, 2, \cdots, \beta_k\}$, or, more compactly, $1\!:\!\beta_k$). The set of all people detected at timestep k is, then, written as $y_k^{1:\beta}$. Note that, when the context is unambiguous we may drop the subscripts and superscripts to reduce clutter. Assuming additive noise, we can write:

$$y_k^j = x_k^i + \nu_k^i \qquad \text{for some } i \tag{1}$$

where ν_k^i is the measurement noise. Since the location detections are split measurements, the mapping between indices i and j is unknown. In fact, this is what we aim to infer.

If a person is known to be carrying phone i, then the person's motion is recorded with inertial measurements w_k^i:

$$w_k^i = x_k^i + \epsilon_k^i \tag{2}$$

where ϵ_k^i is the measurement noise, which is independent of ν_k. Note that the same index i is used to specify both the person's true state and the person's phone. For now we make no assumptions regarding the probability distributions of the ν's and ϵ's, but in our implementation these will be modeled as Gaussian, which simplifies our likelihood calculation.

The relationship between a person's x, y and w is shown in Figure 3. As portrayed in the figure, the x's and w's form a hidden Markov model[1] (HMM) with observations $\{w_k^i\}_k$ (from mobile phones) that are conditioned on states $\{x_k^i\}_k$. In multiple person scenarios, several such HMMs will coexist, each indexed by a distinct i. What is atypical in our problem is the existence of a second set of HMM-like structures whose observations $\{y_k^j\}$, despite being conditioned on *some* state $\{x_k^i\}_k$, do not carry the implicit information of *which* state they are conditioned upon (that is,

[1] In this paper we use the generalized definition of hidden Markov models used in [6], which allows for continuous state spaces.

which i goes with which j). We denote these structures split-measurement HMMs. When multiple people are present, their split-measurements will be shuffled. Then, person-localization will depend on *unshuffling* the split measurements to assign IDs to each anonymous detection from the camera. This is equivalent to discovering the association matrix M for the bipartite graph shown on the right side of Figure 2.

At this point we can finally state the problem as follows:

Identification Problem
Input: *location detections y_k^j (from a camera network) and inertial measurements w_k^i (from mobile phones)*
Output: *the $\gamma \times \beta$ match matrix M_k that assigns each w_k^i to at most one y_k^j with maximum global probability over all timesteps $1\!:\!K$.*

where $\gamma \in \mathbb{N}$ is the number of people equipped with a mobile phone. The matrix M is such that $M_k^{ij} \in \{0, 1\}$ and $M_k^{ij} = 1$ if and only if detection j is identified as person i. This implies $\sum_{\forall i} M_k^{ij} \in \{0, 1\}$ and $\sum_{\forall j} M_k^{ij} \in \{0, 1\}$.

Note that throughout this paper we will use the terms "localization" and "identification" as duals of one another in the following sense: when we assign an ID i to a detection y_k^j we say y_k^j has been identified, and that person i has been localized. Also note that any solution to the identification problem necessarily solves the motion correspondence problem in the process, since each person becomes consistently labeled with their real-world ID.

PROBLEM ANALYSIS

Our derivation is divided into two parts. First we demonstrate the foundation of our method by considering only the instantaneous information from a single timestep. Then, we use the Markov assumption to derive a more precise ID inference by considering all past timesteps.

Optimal Instantaneous ID Assignments

From equations (1) and (2) it is clear that there exists a relation of statistical dependence between the y and w that belong to the same person. This can be easily quantified by subtracting the two equations:

$$y_k^j = w_k^i + (\nu_k^j - \epsilon_k^i) \tag{3}$$

Our goal is to infer which combinations of i, j follow the above equation, that is, which (y, w)-pairs display a measurable statistical dependence.

From (1) and (2) it follows that if the probability distributions of ν and ϵ are known, then so are the emission probabilities $p(y_k^j | x_k^i)$ and $p(w_k^i | x_k^i)$. Then, the likelihood that y_k^j and w_k^i were emitted from the same x_k^i (no matter the actual value of x_k^i) can be found by marginalizing x_k^i:

$$\mathrm{L}(y_k^j, w_k^i) = \int p(x_k^i, y_k^j, w_k^i)\, dx_k^i \tag{4}$$

$$= \int p(y_k^j | w_k^i, x_k^i)\, p(w_k^i | x_k^i)\, p(x_k^i)\, dx_k^i \tag{5}$$

$$= \int p(y_k^j | x_k^i)\, p(w_k^i | x_k^i)\, p(x_k^i)\, dx_k^i \tag{6}$$

where the last equality arises from the conditional independence of y_k^j and w_k^i given x_k^i. In addition, if the prior of x_k^i is uniformly distributed, then the term $p(x_k^i)$ can be cancelled out without without adverse effects. By calculating the likelihood in (6) over all combinations of inertial nodes and detections, we obtain a likelihood matrix Ω_k:

$$\Omega_k = \begin{bmatrix} \mathrm{L}(y_k^1, w_k^1) & \mathrm{L}(y_k^2, w_k^1) & \cdots & \mathrm{L}(y_k^{\beta_k}, w_k^1) \\ \mathrm{L}(y_k^1, w_k^2) & \mathrm{L}(y_k^2, w_k^2) & \cdots & \mathrm{L}(y_k^{\beta_k}, w_k^2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{L}(y_k^1, w_k^\gamma) & \mathrm{L}(y_k^2, w_k^\gamma) & \cdots & \mathrm{L}(y_k^{\beta_k}, w_k^\gamma) \end{bmatrix} \tag{7}$$

The likelihoods in Ω_k constitute the edge weights in the bipartite graph from Figure 2. The most likely global ID assignments are, then, the ones that maximize the joint likelihood, as found using the following optimization:

$$\arg\max_M \prod_{i \in 1:\gamma} \prod_{j \in 1:\beta_k} \Omega_k^{ij} M^{ij} \tag{8}$$

In order to increase robustness against false positives, and to allow for people that are not carrying phones, we set Ω_k^{ij} to 0 if it is below some threshold Ω_{min}. The optimization in (8) can be efficiently solved (in polynomial time) using the Hungarian assignment algorithm [13], as is common in the computer vision literature. Prior to that, it is necessary to convert the multiplications into summations by using log-probabilities.

Maximum a Posteriori Estimate

Our hidden Markov model formulation (Figure 3) hints that a more precise estimate can be found by following the evolutions of w, y and x through all timesteps $k \in 1\!:\!K$. For this, let us consider $x_{1:K}$, $w_{1:K}$ and $y_{1:K}^{j_{1:K}}$, where the latter is a track obtained by associating multiple consecutive location-detections:

$$y_{1:K}^{j_{1:K}} = (y_1^{j_1}, y_2^{j_2}, \cdots, y_K^{j_K}) \tag{9}$$

with $j_k \in 1\!:\!\beta_k$.

In the single-timestep case from the previous section, we marginalized x to compute the emission probability for each y, w given a common x. In the multiple-timestep case, this would translate to marginalizing $x_{1:K}^i$ and computing all possible combinations of $w_{1:K}^i$ and $y_{1:K}^{j_{1:K}}$. This, however, is not feasible, as the rank of the space of all possible tracks is exponentially large. Assuming, for the sake of example, that the number of people detected by the camera network is known to be constant and equal to $\beta_k = \beta$, then the number of possible tracks during $k \in 1 : K$ is β^K. If, additionally, people are allowed to enter or leave at any timestep, then the exponent gains a factorial function, and the number becomes $\beta^{K!}$ [27]. Hence, to marginalize $x_{1:K}^i$ one would need to solve $\beta^{K!}$ K-variable integrals! Clearly, this situation quickly becomes unmanageable.

Rather than marginalize the multiple-timestep hidden state, our solution is to recursively compute the maximum *a posteriori* (MAP) estimate $\hat{x}^i_K$ under the assumption what y and w did originate from the same person. We, then, use $p(\hat{x})$ to quantify the likelihood of our assumption, and generate a likelihood matrix much like (7). For this, let θ^h_K compactly denote a track hypothesis $\theta^h_K = \{y^{j_1}_1, y^{j_2}_2, \cdots, y^{j_K}_K\}$. Then $\Theta_K = \{\theta^{h_1}_K, \theta^{h_2}_K, ..., \theta^{h_{\zeta_K}}_K\}$ is the set of all track hypotheses up to frame K. Then we can calculate the following joint probability:

$$
\begin{aligned}
&p(x^i_{1:K}, \theta^h_K, w^i_{1:K}) = \\
&= p(x^i_{1:K}, \theta^h_K)\, p(w^i_{1:K}|x^i_{1:K}, \theta^h_K) && (10)\\
&= p(x^i_{1:K}, \theta^h_K)\, p(w^i_{1:K}|x^i_{1:K}) && (11)\\
&= p(x^i_{1:K})\, p(\theta^h_K|x^i_{1:K})\, p(w^i_{1:K}|x^i_{1:K}) && (12)\\
&= p(x^i_1) \prod_{k=2:K} p(x^i_k|x^i_{k-1}) \prod_{k=1:K} p(y^{j_k}_k|x^i_k) p(w^i_k|x^i_k) && (13)\\
&= p(x^i_K|x^i_{K-1})\, p(y^{j_K}_K|x^i_K)\, p(w^i_K|x^i_K) \times \\
&\times p(x^i_1) \prod_{k=2:K-1} p(x^i_k|x^i_{k-1}) \prod_{k=1:K-1} p(y^{j_k}_k|x^i_k) p(w^i_k|x^i_k) && (14)\\
&= p(x^i_K|x^i_{K-1})\, p(y^{j_K}_K|x^i_K)\, p(w^i_K|x^i_K) \times \\
&\times p(x^i_{1:K-1}, \theta^h_{K-1}, w^i_{1:K-1}) && (15)
\end{aligned}
$$

where (11) arises from the conditional independence of w, y given x, and (13) from the Markov assumption.

Then we may use (15) to derive the MAP estimate $\hat{x}^i_K$ of the latest hidden state (x_K):

$$
\begin{aligned}
\hat{x}^i_K &= \arg\max_{x^i_K} p(x^i_K|x^i_{1:K-1}, \theta^h_K, w^i_{1:K}) && (16)\\
&= \arg\max_{x^i_K} p(x^i_{1:K}, \theta^h_K, w^i_{1:K}) / p(x^i_{1:K-1}, \theta^h_K, w^i_{1:K}) && (17)\\
&= \arg\max_{x^i_K} p(x^i_{1:K}, \theta^h_K, w^i_{1:K}) && (18)\\
&= \arg\max_{x^i_K} p(x^i_K|x^i_{K-1})\, p(y^{j_K}_K|x^i_K)\, p(w^i_K|x^i_K) \times \\
&\times p(x^i_{1:K-1}, \theta^h_{K-1}, w^i_{1:K-1}) && (19)
\end{aligned}
$$

where the denominator in (17) is cancelled out as it does not change the result of the maximization. The likelihood that all y and w originated from a given sequence of $\hat{x}$ is simply:

$$
\begin{aligned}
\mathrm{L}_{MAP}(\theta^h_K, w^i_{1:K}) &= p(\hat{x}^i_{1:K}, \theta^h_K, w^i_{1:K}) = \\
&= p(\hat{x}^i_K|\hat{x}^i_{K-1})\, p(y^{j_K}_K|\hat{x}^i_K)\, p(w^i_K|\hat{x}^i_K) \times \\
&\times p(\hat{x}^i_{1:K-1}, \theta^h_{K-1}, w^i_{1:K-1}) && (20)
\end{aligned}
$$

As was done in (7) for the single-timestep case, we assign the edge weights of the bipartite graph in Figure 2 by setting $\Omega_K = [\mathrm{L}_{MAP}(\theta^h_K, w^i_{1:K})]^{\forall\, i,h}$. It is important to note that the matrix Ω_K considers only the tracks that are still "alive" at time K, rather than all tracks from $1:K$. The solution to the multiple-timestep identification problem can, then, be found as the match matrix M that maximizes the global probability:

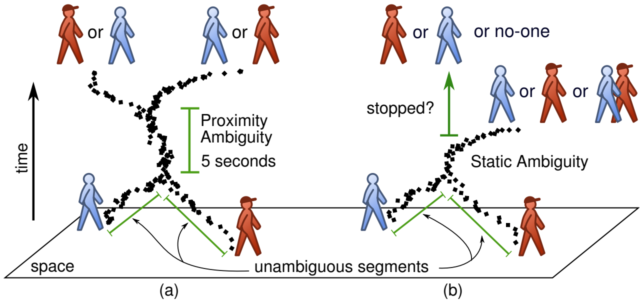

Figure 4. (a) Proximity ambiguity: Two people approach one another, talk for a few seconds, then resume moving. It is not possible to disambiguate the tracks based on motion alone, even if the system is able to detect stopped people. (b) Static ambiguity, which exists when using motion sensitive cameras: since a negative detection does not imply negative presence, it is not possible to tell whether or not one of the people was left behind, completely motionless.

$$
\arg\max_{M} \prod_{i\in 1:\gamma} \prod_{h\in 1:\zeta_k} \Omega^{ih}_K M^{ih} \qquad (21)
$$

Returning to the physical context of our solution, each mobile phone can locally generate its own row in Ω from the broadcast position measurements. At that point, the locally-best solution for each phone is simply the maximum of that row. However, without inter-communication, multiple phones may end up selecting the same coordinates as their location, leading to conflicts. This is resolved by transmiting each locally calculated row back to a central location to piece together the likelihood matrix Ω, with which the optimization in Equation (21) may be performed.

Notice that the central part of this derivation, equation (15), is a recursive relation. This leads to efficient computation of the MAP estimate and its likelihood at each new timestep by simply multiplying the previous value with the latest transition and emission probabilities.

Discussion

An apparent limitation of Equation (21) is that the innermost multiplication is iterated over the set of all track hypotheses ($h \in 1:\zeta_k$). This is problematic because, in the worst case scenario, ζ_k is a very fast-growing integer on the order of $\beta^{k!}$, as discussed in the previous section. Yet, since the primary cause for this combinatorial explosion lies on *proximity ambiguities* (Figure 4a), it is possible to use our identification method to largely bypasses this problem, as we describe next.

A proximity ambiguity is the event that two or more people occupy the same approximate location, so that the person-detecting and tracking layers of the camera network may confuse them. For example, in Figure 4a two people meet for $5s$ and then separate, leaving the tracker to decide how to connect the pre-ambiguity track segments with the post-ambiguity measurements (motion correspondence problem). In the worst case, that 5-second ambiguity can lead to as many as $\zeta_k = 2^{5s \times 30Hz}$ hypotheses (where $30Hz$ is the camera's sampling rate). However, if we postpone making decisions about track correspondence until the very end of an ambiguity, then the ζ_k for Figure 4a will be reduced to merely 4 track hypotheses. The use of such a lazy tracking method is generally risky, as it increases the chances of losing the correct track. Nonetheless, for the purpose of Equation 21 this is not the case, since the solution to the identification problem inherently solves the correspondence problem. Thus, it is justified to employ a lazy tracker in our optimization, greatly reducing the size of ζ_k.

The main disadvantage of using a lazy tracker is that people are not tracked *during* ambiguities, but only *after* they end. Of course, this may be a problematic choice in overcrowded scenarios. For such situations, a more traditional tracker such as [7] is required. There exists a vast literature devoted to multiple-target tracking techniques (based on MHT [2], JPDA [2], and Particle Filtering [34]), any of which may be used to generate a reduced set Θ of track hypotheses for our person-identification solution. This way, rather than committing to a specific tracking method, our solution automatically reaps the benefits of any advances in the field of multiple-target tracking. Further, our solution may be used alongside traditional tools that aid in motion correspondence, such as color histograms [28] and other types of feature matching [17].

IMPLEMENTATION DETAILS

Camera Network

We implemented the method described above in a testbed where multiple camera-nodes are placed on the ceiling, facing down (to minimize occlusions). Since this testbed was originally conceived for use in our assisted-living research [19], its architecture emulates privacy-preserving motion-sensitive cameras [15][16] that can only perceive moving objects, and cannot be used to extract photographs. This brings additional challenges to multiple-target multiple-camera tracking, which would not come up with a more traditional camera setup. For one, traditional person-detection algorithms based on background-subtraction or image segmentation do not work in this setup. Instead, we employ our motion histogram approach [32] which is able to locate multiple moving people in a scene, albeit with a much lower resolution. More importantly, the primary disadvantage of these privacy-preserving cameras is that they cannot detect people who are relatively motionless. Therefore, a "no motion" detection from the camera tells very little about whether or not someone is present. We define this situation as a *static ambiguity*. For instance, consider Figure 4b. Two people meet at the same location, then one of them stops moving while the other continues walking. Immediately after the meeting time, three hypotheses exist: either both people have moved together, or "Person 1" walked away while "Person 2" stayed still at the meeting place, or "Person 2" was the one who walked away. Here, again, we use the concept of a lazy tracker and abstain from making a decision about static ambiguities at the tracker level. Instead, it is left to our person-identification method to automatically reestablish correspondence. The intuition here is that the wearable inertial sensors will be able to disambiguate between a walking person and a stopped one. This privacy-preserving implementation effectively stress-tests our person-identification solution by displaying a much higher level of track segmentation than would traditional CCTV cameras.

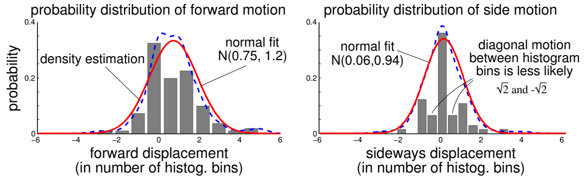

Figure 5. Dataset used for learning the emission and transmission probabilities related to a person walking in our deployment.

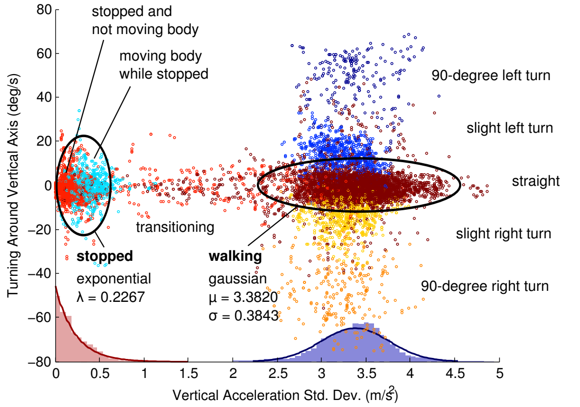

Figure 6. Dataset used for training the classification of 'walking' versus 'stopped'.

Nonetheless, with the motion histogram algorithm, we are able to directly obtain the x, y position of each moving person. Then, the acceleration magnitude and yaw are simply computed from the derivative of the detected person's position. Finally, the two key pieces of information that must be known *a priori* for the HMM, i.e. the emission probability $p(y_k^i|x_k^i)$ and the transmission probability $p(x_k^i|x_{k-1}^i)$, were heuristically estimated from a set of simple experiments. In these, a person walked aimlessly within our camera network 10 times, for a duration of 1 minute each. The extracted data for the transition probabilities of the histogram bins, for instance, is shown in Figure 5.

Wearable Inertial Sensors

Inertial sensors are often used to infer a person's motion path through a process called dead-reckoning. Dead-reckoning demands tightly-calibrated high-precision sensors, and still suffers from drift and abundant noise. It is common to alleviate these issues using sporadic location measurements (e.g. from GPS) or by placing the inertial sensors on the person's foot [11]. The latter approach allows a drift-correction layer

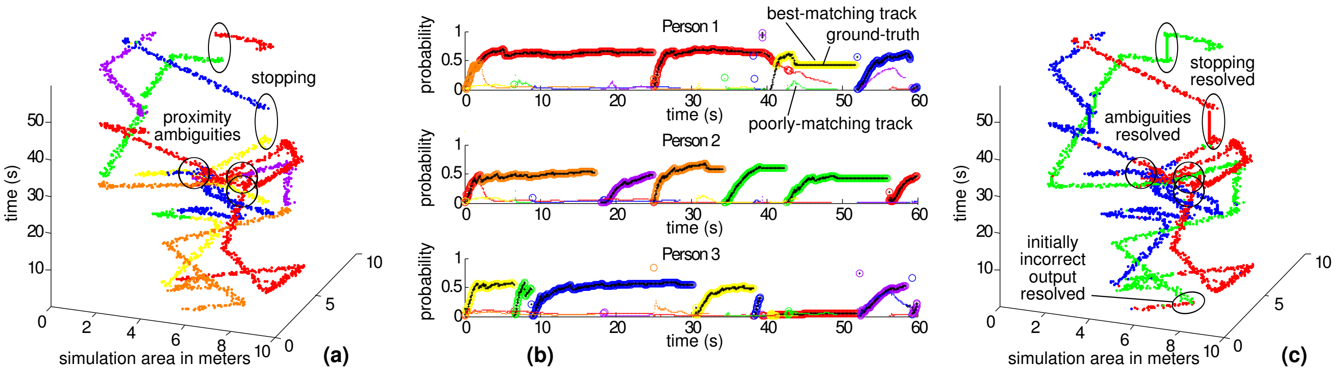

Figure 7. (a) Simulation showing three people moving in a $10m \times 10m$ area. Track segments that are unambiguous are shown with different colors. (b) Calculated probabilities for each track segment from (a), where the tracks selected after global optimization are shown with thick lines, and the ground-truth assignments are shown in black (mostly within the thick lines). (c) Our identification method joins together the segments from (a) into three long tracks with 3 people's IDs.

to execute when the foot is standing still, between steps, with good results over short distances. However, such a method does not work when the sensors are elsewhere on the person's body.

Instead, we eschew the use of dead-reckoning altogether and opt for a less complex, entirely probabilistic approach, relying on the known emission probability $p(w_k^i|x_k^i)$. We employ the magnetometer on the phones to extract the direction of motion, and the accelerometer to extract a binary signal indicating whether the person is walking or stopped. From our experiments we have found that people generally walk in long segments of nearly-constant speed, which makes the use of non-binary motion sensing largely superfluous. This agrees with the literature on human locomotion [35][5], where it is found that walking paths are mostly linear and at a piecewise constant preferred walking speed [4]. We classify motion as 'walking' or 'stopped' by measuring the amount of vertical bobbing observed by the wearable accelerometer. For this, we acquire a training dataset (shown in Figure 6) and fit the 'walking' and 'stopped' classes to a Gaussian and exponential distribution. We have found that the same distribution parameters can be used for different people without any noticeable loss in precision. We use the learned parameters, then, to infer the binary motion state at run time from a simple likelihood ratio test. Similarly, the direction observations can be assumed to follow the true hidden angle plus Gaussian noise, although additional provisions must be taken since angles make up a circular space (modulus 2π *radians*). This causes noisy measurements near 0 or 2π *radians* to wrap around, landing far from the true value. Our answer is to, instead, use a circular variation of the Normal distribution, known as the Von Mises distribution [10].

It is important to point out that magnetic readings can be affected by the presence of metal or additional magnetic fields in a room. Although in our experience this unwanted effect has not been noticeable, it can be corrected by constructing *a priori* a map of the environment's magnetic fields. A similar process is already done in many mobile phones to account for deviations in the Earth's magnetic field using the World Magnetic Model produced by the U.S. National Geospatial-Intelligence Agency.

EVALUATION

Simulation

We used a simulator to characterize the performance of our method given different target densities (the number of people in the network divided by the coverage area). The simulator uses a random-waypoint model to generate scenarios where multiple people move at different speeds and possibly stop for some random time duration. It emulates cameras, motion-sensors and wearable inertial sensors with different noise characteristics and sampling rates, including a sine-squared model to emulate gait perturbations on the accelerometer signals.

The traces from a three-person simulation is shown in Figure 7(a). In the figure, the detections from the simulated camera are colored according to which piecewise-unambiguous track segments they belong to. The problem we aim to solve is that of (1) stitching these tracks together, and (2) identifying who they belong to. For this, we compute the probabilities in the likelihood matrix Ω at each timestep. These probabilities are shown in Figure 7(b). Coloring is used to indicate the same track segments as in Figure 7(a). Ground-truth is shown in thin black lines, and our current belief after the global optimization step is shown with a thick stroke. In the figure, most locally-best matches (the tracks with highest probability for each person) happen to coincide with the global optimum, but this is often not true in other scenarios. Finally, a plot of the best global matches in space is shown in Figure 7(c). The figure shows that people's IDs have been recovered, i.e. track segments belonging to the same person are correctly joined into long same-color paths without spatio-temporal gaps.

To quantify the accuracy of the system, we simulated 100 scenarios consisting of 1 to 10 people in a $10m \times 10m$ area. The simulated cameras were sampled at 20Hz and the inertial sensors at 100Hz. We considered the three following sensor setups:

1. **Ideal sensors** — As a sanity-check, we consider the sim-

ulation of ideal noiseless sensors to verify the correctness of our approach.

2. **Non-ideal sensors and a regular camera** — We simulate non-ideal cameras and inertial sensors to assess the identification accuracy when using a regular camera under realistic conditions. For this, zero-mean Gaussian noise is added to all sensor readings, with $\sigma = 0.15m$ (cameras), $0.03m/s^2$ (accelerometers) and $0.02 \times H$ (magnetometers), where H is the magnitude of the Earth's magnetic field.

3. **Non-ideal sensors and privacy-preserving motion cameras** — Finally, to provide a baseline against which our experimental evaluation can be compared, we simulate the camera network that we developed for our assisted living deployments. For this, the inertial measurements were simulated like in the previous case, while the location measurements were additionally quantized to $15cm$ increments. This is in agreement with the reported tracking accuracy for our motion histogram method of detecting people in a privacy-preserving motion image [32]. This setup has the coarsest resolution of all three simulated scenarios, and should present the toughest conditions for person-identification.

We quantify the accuracy of our method using the *multiple-object tracking accuracy* (MOTA) metric proposed by Bernardin et al. [3]. This is defined as:

$$MOTA = \frac{\sum_{\forall k} \text{correct identifications in } k}{\sum_{\forall k} \text{all identifications in } k} \qquad (22)$$

Thus MOTA is a measure of how accurate our identification attempts are. The difference between MOTA and the classic *precision/recall* metrics is that MOTA considers the output of the system at all moments in time, rather than solely the final inference. This is designed to catch even momentary variations in accuracy, giving a good picture of the real-time performance of the system. Bernardin also proposes a second metric, the *multiple-object tracking precision* (MOTP), for measuring the precision of the location estimates. However, since the person-identification approach described in this paper simply forwards the location estimates of the camera network (which is considered as a black box), we focus our evaluation on the precision of the ID assignments instead. The localization precision is entirely dependent on the black-box tracker and, therefore, follows the state-of-the-art in computer vision.

Figure 8 shows the simulated accuracy of our method averaged over 10 runs for each datapoint. The accuracy found for the ideal simulation is approximately 100% for all cases, which serves to corroborate the correctness of our approach. When using noisy data, our method achieves accuracy of over 95% with the regular camera, and over 80% with the privacy-preserving motion cameras. The performance loss in the latter case can be explained by its low resolution, which adds considerable quantization noise to the y_k^j's. The data in Figure 8 can be better interpreted with the knowledge that larger target-densities lead to shorter unambiguous track lengths, which are harder to identify. A reduced inter-ambiguitiy time presents a challenge to our method in that they reduce the probability that people will act to differentiate themselves within that time. It makes sense, therefore, to analyse the accuracy of the system as a function of the mean interambiguity time, as shown in Figure 9. Our data shows that the proposed identification procedure has an accuracy of over 90% for tracks lasting as little as $2.9s$, even when using the motion histogram. In most assisted-living situations, however, we expect interambiguity times to be much longer, leading to increased accuracy as shown in the right side of the plot.

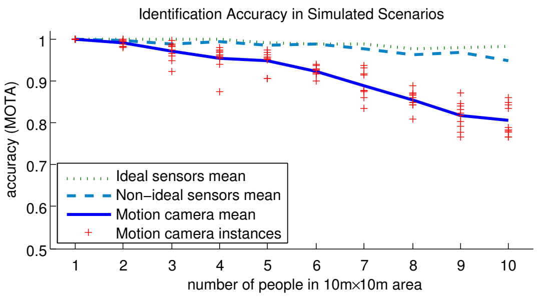

Figure 8. Identification accuracy (MOTA) as a function of target density. The figure shows 100 simulations where the number of people in the scene varies from 1 to 10. The mean is shown with a thick blue line.

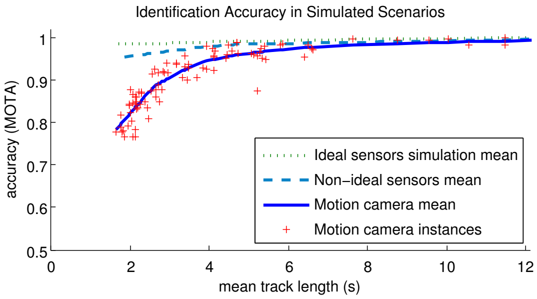

Figure 9. Plotting the same data as Figure 8, but as a function of inter-ambiguity time. We obtain a accuracy (MOTA) of over 0.9 for piecewise-unambiguous tracks lasting as short as $3.5s$. Of course, the longer the track segments, the higher the chance our method will correctly identify the person.

Experiments

In addition to the simulations described in the previous section, we also performed experiments on a testbed deployment of 4 Intel iMote2 sensor nodes equipped with custom camera boards. The iMote2's PXA271 processor was set to operate at $208MHz$, allowing it to detect people in using the motion histogram approach at frame rate of approximately $14Hz$. The cameras were placed on the ceiling, facing down, at a height of $2.1m$. We used a 162° wide-angle lens to be able to capture the *full height* of a person in an area of approximately $3m \times 4m$ for each camera (partial images of people could be seen from areas much larger). The location of each detected person was transmitted over $802.15.4$ and recorded at a nearby laptop. Although we have recently started implementing the inertial-sensing and local optimization layers of our method on a Google Android phone (the

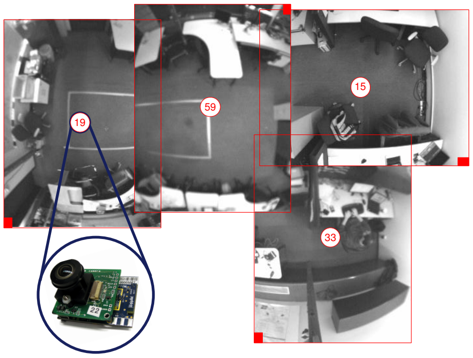

Figure 10. Experimental testbed with 4 iMote2 sensor nodes instrumented with our custom camera boards.

Nexus One), in this proof-of-concept implementation, we use a SparkFun 6DoF inertial measurement unit (IMU) instead. The IMU, attached to the person's belt, transmitted the measured 3D acceleration and magnetic force through a Bluetooth link at a sampling frequency of $100Hz$. The nodes' internal clocks were loosely synchronized by simply transmitting a beacon with the global time at the beginning of the experiments, leading to synchronization disparities as high as $50ms$. In addition, whereas multiple camera systems in the literature often use high-speed links to synchronize the cameras' sampling rates, our camera nodes sampled each frame in a completely asynchronous manner. This is in agreement with the conditions of existing camera infrastructure.

We performed two sets of experiments in our person-identification testbed. On the first set, we acquired 15 experimental traces where 1 person freely walked for 1 minute within the 4-node testbed carrying the inertial sensor. We, then, superimposed N of these 1-person traces onto one another to generate datasets where the ground truth was perfectly known for all N people involved. The number N of people was varied from 1 to 4, with 5 repetitions each. The results, shown in Figure 11 as a function of interambiguity time, are in agreement with the trend found in our simulations (dashed line). As can be seen from the plot, the interambiguity time in our experiments were found to be quite short, lower than $4.5s$. This was caused by two factors: (1) the large density of people for such a small deployment caused frequent proximity ambiguities, and (2) the privacy-preserving motion cameras often failed to detect people slowed down before making a turn, leading to high track fragmentation. Nonetheless, accuracy exceeded 90% for interambiguity rates as high as one every 3.5 seconds.

For our second set of experiments, we evaluated the capability of the system to disambiguate between two people after an ambiguity. For this, we acquired 36 traces where the two people walked from one end of the deployment to the other, in trajectories that lasted approximately 4 to 5 seconds, spanning 3 different camera nodes on average. Only one of the persons was carrying an inertial sensor node. These traces are quite challenging given their short duration, and serve as a stress test on the ability of the system to quickly recover from ambiguities. There were 9 experimental runs consisting of one of four scenarios: (1) two people walking in opposite directions, crossing paths in the middle of the trace; (2) two people walking in opposite directions, meeting in the middle of the trace, but not crossing paths; (3) two people walking in the same direction, crossing paths; (4) two people walking in the same direction, meeting but not crossing. The accuracy of our identification method is shown in Figure 12. The average accuracy (median of the set of 'all data') was found to be 0.8437. This agrees with our simulation for tracks lasting $2.25s$ — or approximately half the duration of our traces, given that the piecewise-unambiguous tracks were interrupted at the middle. As expected, the accuracy for opposite-direction traces is on average higher than for same-direction ones, owing to a larger contribution from the magnetometer measurements. Finally, of all our simulations and experiments, the worst case running time for the proposed identification method was approximately 6× faster than real-time.

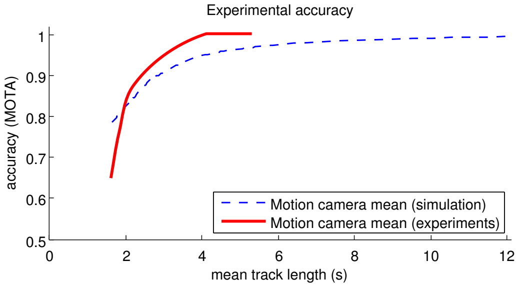

Figure 11. Experimental results from overlapping up to four 1-person experiments at a time. The experimentally-found accuracy closely follows the trend from our simulation results.

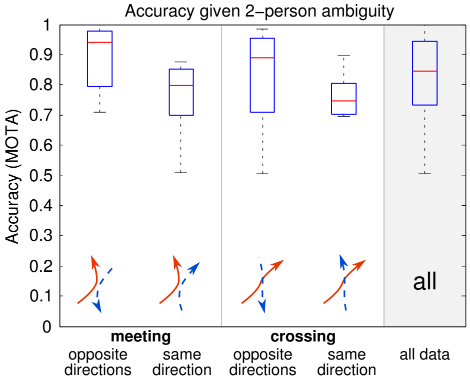

Figure 12. Experimental results for 36 ambiguity-resolution tests. The median accuracy for all experiments was found to be 84.37%, as predicted by our simulations.

CONCLUSION

We have presented a method to identify/localize people using camera networks and mobile phones. The system has applications for tracking people in smart environments, locating personnel in a building, generating alerts regarding the presence of unauthorized individuals (i.e. people detected by the cameras, but who do not carry the mobile phone client), and many others. The main advantage of our proposed method is that it requires little or no modification of the infrastructure currently found in shopping malls, airports, and other public spaces.

The problem was formulated as one hidden Markov model for each person in the scene, with two observations per model: the first observation comes from the camera network, carrying location information but not the person's IDs (i.e. split measurements); and the second observation comes from the person's inertial node, which carries a unique ID but cannot localize. The solution of the identification problem was described as a matching between inertial nodes and detections from cameras, thus simultaneously obtaining IDs and locations. We showed that this could be achieved by recovering the dependence relation between the inertial measurements and the person's motion as observed by the cameras. Results demonstrates that our method can achieve high accuracy (in excess of 90%) for scenarios with interambiguity time as little as 3.5 seconds. Our method is limited mostly by the accuracy of the vision subsystem's person-detection and tracking modules, and can be improved by simply utilizing a more capable tracker.

REFERENCES

1. N. Anjum and A. Cavallaro. Multi-camera calibration and global trajectory fusion. In Y. Ma and G. Qian, editors, *Intelligent Video Surveillance: Systems and Technologies*. CRC Press, 2009.
2. Y. Bar-Shalom and W. D. Blair, editors. *Multitarget-Multisensor Tracking*. Artech House Publishers, 1990.
3. K. Bernardin, A. Elbs, and R. Stiefelhagen. Multiple object tracking performance metrics and evaluation in a smart room environment. In *IEEE VS*, 2006.
4. J. E. A. Bertram and A. Ruina. Multiple walking speed-frequency relations are predicted by constrained optimization. *Journal of Theoretical Biology*, 209(4):445 – 453, 2001.
5. D. C. Brogan and N. L. Johnson. Realistic human walking paths. In *IEEE CASA*, 2003
6. O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer Verlag, 2005.
7. J. Connell, A. Senior, A. Hampapur, Y. Tian, L. Brown, and S. Pankanti. Detection and tracking in the IBM PeopleVision system. In *IEEE ICME*, 2004.
8. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR*, 2005.
9. D. Devarajan and R. Radke. Distributed metric calibration for large-scale camera networks. In *BASENETS*, 2004.
10. N. Fisher. *Statistical analysis of circular data*. Cambridge Univ Pr, 1996.
11. E. Foxlin. Pedestrian tracking with shoe-mounted inertial sensors. *IEEE Computer Graphics and Applications*, 25(6):38–46, 2005.
12. M. Kourogi and T. Kurata. Personal positioning based on walking locomotion analysis with self-contained sensors and a wearable camera. *IEEE/ACM ISMAR*, 2003.
13. H. W. Kuhn. The hungarian method for the assignment problem. In *Naval Research Logistic Quarterly*, volume 52, 1955.
14. B. Kusy, A. Ledeczi, and X. Koutsoukos. Tracking mobile nodes using rf doppler shifts. In *ACM SenSys*, 2007. ACM.
15. P. Lichtsteiner, J. Kramer, and T. Delbruck. Improved on/off temporally differentiating address-event imager. In *IEEE ICECS*, 2004.
16. P. Lichtsteiner, C. Posch, and T. Delbruck. A 128× 128 120dB 15us latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid State Circuits*, 43(2):566–576, 2008.
17. D. G. Lowe. Distinctive image features from scale-invariant keypoints. volume 60, pages 91–110, 2004.
18. 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 *Lecture Notes in Computer Science*, volume 3868, page 326, 2006.
19. D. Lymberopoulos, T. Teixeira, and A. Savvides. Macroscopic human behavior interpretation using distributed sensor networks. *Proceedings of IEEE*, October 2008.
20. W. Mayol, A. Davison, B. Tordoff, and D. Murray. Applying active vision and slam to wearables. In *IEEE ISRR*, 2003.
21. T. Murakita, T. Ikeda, and H. Ishiguro. Human tracking using floor sensors based on the Markov chain Monte Carlo method. In *IEEE ICPR*, 2004.
22. A. Savvides, C. Han, and M. B. Srivastava. Dynamic fine grained localization in ad-hoc sensor networks. In *ACM Mobicom 2001*, 2001.
23. E. Schoonderwaldt, N. Rasamimanana, and F. Bevilacqua. Combining accelerometer and video camera: Reconstruction of bow velocity profiles. In *NIME*, 2006.
24. D. Schulz, D. Fox, and J. Hightower. People tracking with anonymous and id-sensors using rao-blackwellised particle filters. In *IJCAI*, 2003.
25. M. Shankar, J. Burchett, Q. Hao, B. Guenther, D. Brady, et al. Human-tracking systems using pyroelectric infrared detectors. *Optical Engineering*, 45:106401, 2006.
26. N. Shrivastava, R. Madhow, and S. Suri. Target tracking with binary proximity sensors: fundamental limits, minimal descriptions, and algorithms. In *ACM SenSys*, 2006.
27. L. D. Stone, C. A. Barlow, and T. L. Corwin. *Bayesian Multiple Target Tracking*. Artech House Publishers, 1999.
28. M. Taj, E. Maggio, and A. Cavallaro. Multi-feature graph-based object tracking. In *CLEAR*, pages 190–199, 2006.
29. Y. Tao, H. Hu, and H. Zhou. Integration of vision and inertial sensors for 3D arm motion tracking in home-based rehabilitation. *The International Journal of Robotics Research*, 26(6):607, 2007.
30. T. Teixeira, D. Jung, G. Dublon, and A. Savvides. Identifying people by gait-matching using cameras and wearable accelerometers. In *ACM/IEEE ICDSC*, 2009.
31. T. Teixeira, D. Jung, G. Dublon, and A. Savvides. Identifying people in camera networks using wearable accelerometers. In *ACM PETRA*, 2009.
32. T. Teixeira and A. Savvides. Lightweight people counting and localizing in indoor spaces using camera sensor nodes. In *ACM/IEEE (ICDSC)*, 2007.
33. 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.
34. J. Vermaak, S. Godsill, and P. Perez. Monte Carlo filtering for multi-target tracking and data association. *IEEE Transactions on Aerospace and Electronic systems*, 41(1):309–332, 2005.
35. W. T. Willis, K. J. Ganley, and R. M. Herman. Fuel oxidation during human walking. *Metabolism - Clinical and Experimental*, June 2005.
36. J. Wilson and N. Patwari. Radio tomographic imaging with wireless networks. *IEEE Trans. Mobile Computing*, 2009.
37. Z. Xiang, S. Song, J. Chen, H. Wang, J. Huang, and X. Gao. A wireless lan-based indoor positioning technology. *IBM J. Res. Dev.*, 48(5/6):617–626, 2004.