Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 5 (2011) 378–385

Procedia Computer Science

The 2nd International Conference on Ambient Systems, Networks and Technologies (ANT)

Spatio-Temporal Reasoning in Biometrics Based Smart Environments

Vivek Menon[a,*], Bharat Jayaraman[b], Venu Govindaraju[b]

[a]*Amrita Vishwa Vidyapeetham, Coimbatore 641105, India*
[b]*University at Buffalo, Buffalo, NY 14260, USA*

Abstract

We discuss smart environments that identify and track their occupants using unobtrusive recognition modalities such as face, gait, and voice. In order to alleviate the inherent limitations of recognition, we propose spatio-temporal reasoning techniques based upon an analysis of the occupant tracks. The key technical idea underlying our approach is to determine the identity of a person based upon information from a track of events rather than a single event. We abstract a smart environment by a probabilistic state transition system in which each state records a set of individuals who are present in various zones of the smart environment. An event abstracts a recognition step and the transition function defines the mapping between states upon the occurrence of an event. We define the concepts of 'precision' and 'recall' to quantify the performance of the smart environment. We provide experimental results to show performance improvements from spatio-temporal reasoning. Our conclusion is that the state transition system is an effective abstraction of a smart environment and the application of spatial-temporal reasoning enhances its overall performance.

Selection and/or peer-review under responsibility of Prof. Elhadi Shakshuki and Prof. Muhammad Younas.

Keywords: Smart environments, Biometrics, Recognition, Spatio-Temporal Reasoning, Precision, Recall, Abstract Framework, Events, States, Transitions

1. Introduction

The goal of our research is to develop indoor smart environments that can recognize and track their occupants as unobtrusively as possible and answer queries about the whereabouts of their occupants. The sensors of interest in our work are video cameras, microphones, etc. Such environments are useful in settings ranging from homes for the elderly or disabled and office workplaces, and can be extended to larger arenas such as department stores, shopping complexes, train stations, airports, etc.

Identification of occupants has traditionally relied on tag-based approaches involving RFID badges where the occupant has to continuously retain them or biometrics-based approaches based on fingerprint and

*Corresponding Author
Email addresses: vivek_menon@cb.amrita.edu (Vivek Menon), bharat@buffalo.edu (Bharat Jayaraman), govind@buffalo.edu (Venu Govindaraju)

1877–0509
Selection and/or peer-review under responsibility of Prof. Elhadi Shakshuki and Prof. Muhammad Younas. doi:10.1016/j.procs.2011.07.049

iris scans, which involve a 'pause-and-declare' interaction with respect to the occupant [1]. These modalities are considered less natural than other biometric modalities such as face, voice, height, and gait, which are less obtrusive and therefore better candidates for identification of occupants in our smart environments.

In this paper, we extend our previous research on abstracting the behavior of a multimodal smart environment in terms of a *state transition system*: states, events, and a transition function [2, 3, 4]. The state captures who is present in the different regions, or zones, of the environment. The state changes upon an event, i.e., the movement of an occupant from one zone to another. An event abstracts a biometric recognition step - whether it is face recognition, voice recognition, etc. - and is represented as a set of pairs $\langle o, p(o)\rangle$ where $p(o)$ is the probability that occupant o has been recognized at this event. The state information is thus also probabilistic in nature. The transition function takes as input a state and an event, and determines the next state by assigning revised probabilities to the occupants based upon the probabilities in the event.

We also show in this paper how spatio-temporal reasoning can help alleviate some of the limitations of the underlying recognition methodology. Identification based upon a single event is subject to the vagaries of biometric recognition. For example, in face recognition, the angle of the camera, the amount of illumination and face expression could cause a misidentification. Spatio-temporal reasoning is more robust in that the identity of a person is based upon information from a track of events rather than a single event. The basic idea is that the consecutive track elements of a valid track will mostly obey the zone adjacencies in the physical environment, whereas spurious tracks will mostly violate the zone adjacencies. Thus, an occupant o is not confirmed for any event unless there is a coherent track for o with respect to zone adjacencies.

We incorporate this spatio-temporal reasoning into the transition function of our state transition system. In our earlier paper [2, 3], we proposed a simple transition function of the form $\Delta : S \times E \rightarrow S$. Here, the next state is determined just from the current state and current event. When track-based reasoning is employed, the transition function takes the form $\Delta : \mathcal{P}(\mathcal{S}) \times E \rightarrow S$. That is, the next state is determined only after examining the tracks that are implicit in the set of all previous states. We also present a more refined transition function of the form $\Delta : \mathcal{P}(\mathcal{S}) \times E \rightarrow \mathcal{P}(\mathcal{S})$. Here, in addition to track analysis, the transition function also determines a revised set of previous states. Since track-based reasoning on shorter tracks is less effective than on longer tracks, the errors in initial states can be corrected only retrospectively when more events have taken place.

We also extend previous research on quantitative metrics for identification and tracking in a smart environment based upon two metrics: precision and recall. Precision captures the 'false positives' while recall captures the 'false negatives'. These are complementary concepts and together capture the overall performance of a smart environment. These are standard performance measures in the information retrieval literature [5], but we have adapted their definitions to suit our context.

We present results from a prototype implementation of our concepts based upon biometric data that was captured from continuous video frames. Our results confirm that the state transition model serves as an elegant abstraction of a smart environment and that spatio-temporal reasoning enhances its overall precision-recall. The rest of this paper is organized as follows. Related work is surveyed in section 2, and the details of the state transition model as well as precision and recall are discussed in section 3. Spatio-temporal reasoning and results from our experimental system are presented in section 5. Conclusions and further work are described in section 6.

2. Related Work

There exists a number of biometric based approaches to identification in smart environments [6, 7, 8]. Our research on multimodal identification and tracking in smart environments is similar to the previous approaches [9, 10]. However, our focus is on unobtrusive identification and tracking in larger environments like office workplaces, hospitals or other campuses which could be partitioned into zones or blocks. In our work, the tracking is discrete, generating location cum identity updates of an occupant only at zone or block level. This obviates the need for deploying cameras or other sensors with overlapping views, as in continuous tracking models.

A major difference between our approach and several of the approaches surveyed earlier is our use of a state transition model in which multimodal recognition output is uniformly abstracted as events. In this

paper we build on the novel idea of integrating recognition and reasoning for enhancing the overall accuracy of recognition in smart environments. This paper extends our previous work [3] and discusses the details of a track-based reasoning approach for alleviating the shortcomings of a pure recognition based approach.

HMMs and their variants, such as Factorial HMMs and Coupled HMMs, may be regarded as examples of Dynamic Bayesian Networks (DBNs) [11]. Here, transition probabilities are typically learned from empirical data of the movements of people through the space, gathered over a period of time. We do not adopt this approach as we cannot assume a predictable pattern of movement of people through various zones of the smart environment. In our state transition system approach, biometric capture devices provide direct information on the occurrence of events in specific zones (i.e. movement of people) and given an event occurring in a zone, the next state can be unambiguously determined, albeit the probabilistic nature of the state information. Furthermore, a state with n occupants and m zones requires only $m * n$ storage, since for each of the m zones we record the probabilities of each of the n occupants being present in that zone.

Spatio-temporal reasoning has been investigated from a logic and constraint perspective, with applications in geographical information systems (GIS), computer vision, planning, etc [12]. Spatio-temporal reasoning over occupant tracks is similar to a higher-order Markov process, since the next state depends upon multiple previous states. When the transition function also updates the information in previous states, the resulting inference is closer to that of a Markov Random Field (MRF) analysis. In the MRF approach, the operation of a smart environment may be modeled by an undirected graph whose nodes correspond to space-time (or zone-event) points and edges capture space-time adjacency. Spatio-temporal reasoning with MRF is based upon a neighborhood analysis around the zone of occurrence of an event. While it is more general in principle, it is also computationally more complex than track-based reasoning, which is more specialized and hence can more efficiently incorporate a global view of the system.

3. Abstract Framework

We shall consider a smart environment as being made up of a number of zones, each of which is a region – a room or a set of rooms. An n-person smart environment is abstracted as a state transition system (S, E, Δ) where S is the set of states labeled $s_0, s_1, \ldots s_x$, E is the set of events labeled $e_1, e_2, \ldots e_x$, and $\Delta : S \times E \rightarrow S$ is a function that models the state transition on the occurrence of an event. The state transitions may be depicted as follows: $s_0 \stackrel{e_1}{\rightarrow} s_1 \stackrel{e_2}{\rightarrow} s_2 \ldots \stackrel{e_x}{\rightarrow} s_x$

- A *state* records for each zone the probability of presence of each occupant in that zone. For each occupant, the sum of probabilities across all zones equals one.
- An *event* abstracts a biometric recognition step and is represented as a set of person-probability pairs, $\langle o_i, p(o_i)\rangle$, where $p(o_i)$ is the probability that occupant o_i was recognized at this event. We also have $\sum_{i=1}^{n} p(o_i) = 1$.
- The *transition function* abstracts the reasoning necessary to effect state transitions. In the zone of occurrence, we define $p_s(o_i) = p(o_i) + x_i * p'_s(o_i)$, where $x_i = 1 - p(o_i)$ and $p'_s(o_i)$ is probability of the occupant in the previous state. For all other zones, we define $p_s(o_i) = x_i * p'_s(o_i)$. This ensures that the sum of probabilities for an occupant across all zones in the resultant state equals one. A more detailed account of the transition function may be found in [4].

For simplicity, we assume that events happen sequentially in time, i.e., simultaneous events across different zones are ordered arbitrarily in time. That is, the entry of an occupant o_i into zone z_i and occupant o_j to zone z_j at the same time t can be modeled as o_i before o_j or o_j before o_i. Thus events are assumed to be independent, but the transition function captures the dependency on the previous state, as in a Markov process.

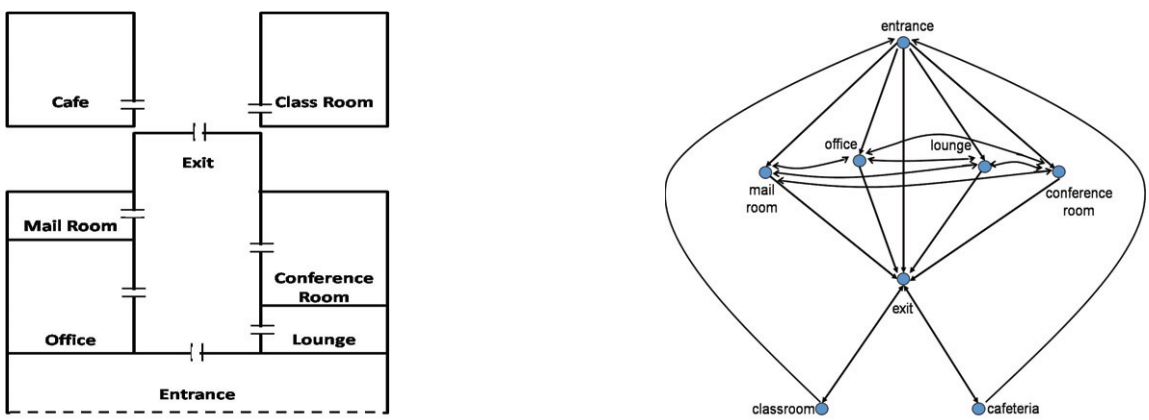
We define the concepts of precision and recall for a smart environment in terms of the ground truth, which, for a given input event sequence, is a sequence of states of the environment wherein the presence or absence of any occupant in any zone is known with certainty (0 or 1). Precision captures the extent of

'false positives' while recall captures the extent of 'false negatives'. These definitions are stated in terms of a *recognition threshold* θ; only those persons with a probability $\geq \theta$ are assumed to be present. When a person's probability in two or more zones is $\geq \theta$, the zone with the highest probability is taken as the zone of his presence. We refer to the set of persons occurring in a ground truth G as $occ(G)$.

1. $\pi = tp/(tp + fp)$, where tp is the set of 'true positives' and fp is the set of 'false positives'. The set $tp = \{o_i : p_s(o_i) \geq \theta \wedge o_i \in occ(G)\}$, while the set $(tp + fp) = \{o_i : p_s(o_i) \geq \theta\}$.
2. $\rho = tp/(tp + fn)$, where tp is defined as above, and fn is the set of 'false negatives'. The set $(tp + fn) = \{o_i : o_i \in occ(G)\}$.

4. Recognition sans Reasoning

An automated and unobtrusive approach to biometric recognition introduces errors in the overall recognition process. There are two broad factors leading to the errors in recognition - extrinsic and intrinsic. Extrinsic sources of the error includes errors in sensors (cameras), availability of lighting, distance of subject from sensor, occlusions, pose variations, number of subjects in the frame etc. The extrinsic factors compound the inexact nature of automated biometric recognition and produce scenarios where the ground truth does not emerge as the top estimate. The state information is probabilistic in nature as it is based on the occurrence of an event that produces probability estimates of the occupants registered in the database based on the distance scores generated by the biometric recognition algorithm. The probabilistic notion of identity introduces errors that accumulate over time in the state transition system model.



(a) (b)

Fig. 1: Layout of a Multimodal Smart Environment

Let us consider a smart environment with adjacencies as shown in figure 1a, where some of the registered occupants have entered the environment and moved between various zones of the environment triggering a sequence of biometric events and associated state transitions. We describe basic strategies for estimation of the event sequence and occupant tracks from the sequence of states of a smart environment.

4.1. Estimation of Event Sequence

Given a sequence of observed events and the corresponding states, our smart environment model will estimate the identity of the occupant associated with each state transition. For any two consecutive states, the maximum difference in occupant probabilities in the zone of occurrence of the event is taken as the criterion for determining the occupant who moved. We do not consider the person with the highest probability in an event as the one who moved, for two reasons: (i) event information could be erroneous; and (ii) comparing consecutive states gives due importance to both event and historical information.

The projection of the estimated occupant sequence with respect to any occupant of interest yields the occupant's track. However, the errors in estimation of events lead to the introduction of false positives and false negatives in the estimated event sequence of the smart environment. This in turn leads to generation of spurious track entries which wrongly associate an occupant with an event occurring in a zone at a point of time. These misidentified events involve either occupants already in the environment or registered occupants who never entered the smart environments. A misidentified event creates a false negative in the track of the actual occupant and introduces a false positive in the track of the estimated occupant.

To limit the erroneous estimated tracks, only those plausible occupant tracks where the average probability is at least the recognition threshold are chosen. The key idea behind this filtering is from the observation that erroneous occupant tracks either exhibit wide variations in their constituent zone level probabilities or possess low occupant probabilities throughout. Since the notion of the recognition threshold is central to the consideration of valid occupants of a state, the use of this threshold is extended for evaluation of plausible tracks. Thus, an occupant o is not confirmed for any event unless there is a coherent average track for o with respect to recognition threshold θ. It is worthwhile to note that the erroneous occupant tracks also exhibit inconsistent spatial and temporal properties in terms of their constituent track points (zones and time stamps of occurrence). However, a recognition only approach does not factor these spatio-temporal constraints during track estimation and hence spurious occupant tracks exist within the set of estimated tracks.

5. Spatio-Temporal Reasoning

In this section, we discuss how spatio-temporal reasoning can help alleviate some of the limitations of the underlying recognition methodology by minimizing the impact of recognition errors on the system. The error detection and correction strategies revolve around the concept of valid occupant tracks. We first revise the set of misidentified events from an analysis of estimated tracks by classifying them as valid or spurious, and then proceed to construct an improved set of tracks that minimize the number of misidentified events. The elements of tracks classified as spurious are appended to the existing set of misidentified events.

As noted in the introduction, our approach in spatio-temporal reasoning is to identify a person using a track of events and corresponding states rather than just a single event. This in turn requires us to determine which tracks (of the occupants) are spurious and which are valid. In order to determine spurious tracks, we observe that consecutive track elements of a valid track will mostly obey the zone adjacencies in the physical environment, whereas spurious tracks will mostly violate the zone adjacencies. Thus, an occupant o is not confirmed for any event unless there is a coherent track for o with respect to zone adjacencies. Factoring multiple states to determine a coherent track is similar in spirit to an n-order Markov chain discussed earlier. This track-based reasoning can be captured by a revised transition function of the form $\Delta : \mathcal{P}(\mathcal{S}) \times E \rightarrow S$, that is upon the occurrence of an event, a set of previous states are used to compute the occupant tracks and thus determine the next state.

It is possible that a valid track may have a few events that are misidentified as well as few transitions that do not obey the zone adjacencies, i.e., there may be one or more missing events. When a track is determined to be spurious, it means that all events in this track are considered to be misidentified and therefore they are candidates for re-identification: Some of these events can be reassigned to the valid tracks at those places where there is a missing event (or events). For the remaining events, their probabilities are re-determined using knowledge of the occupants in adjacent zones given the time and location of the event as well as the valid tracks. Once we determine these occupants, we can map their distance scores to probabilities as described earlier. Since the set of neighboring occupants will in general be a much smaller than the set of all registered occupants, the resulting probabilities will be better. The revised probabilities serve as a basis for determining a new event sequence which in turn a revised sequence of states and an improved set of tracks. The zone adjacencies of a layout specifies the connectivity between the different zones of the environment under consideration and is illustrated as a directed graph as in 1b.

5.1. Experimental Testbed

We illustrate in this subsection the modeling of a 8-zone university building with 45 registered occupants as a smart environment. We map each of the frequented areas as belonging to a separate zone and name the

zone according to the room/area it covers – entrance, office, mailroom, lounge, conference room, classroom, cafeteria and exit as per the layout illustrated in figure 1b. A training database was prepared by enrolling multiple face images of each occupant. The entry of an occupant at a zone is captured by a video camera which triggers a face recognition event. The face recognition module was customized from an OpenCV [13] implementation of the eigenface algorithm [14]. The distance score generated by the recognition algorithm with respect to each registered occupant is recast as a probability value [15] which denotes the posterior probability of the detected face matching the pre-registered occupants. This set of person-probability pairs generated essentially constitutes an event as defined in section 3. Our formulation of sensor quality σ abstracts intrinsic and extrinsic factors that can affect the recognition output. Since we have 10 different event templates (face images) for every person in the database, when the sensor quality is reduced (using the slider bar at the top left of the GUI in figure 3), our system will choose a lower quality image such that the event probability for the person recognized is correspondingly lower. A varying number of false positives across these event templates also factors the variability due to noise and errors in unconstrained biometric recognition.

We discuss the results for a sample runs with 5, 10, 15, 20 and 25 occupants inside the smart environment. For each occupant, a script randomly generates a a trajectory, which is represented by a sequence of zones visited by the occupant. The movement of an occupant between any two consecutive zones is assigned an event randomly drawn from the occupant's pool of 10 event templates. The state changes of the smart environment are driven by the events associated with the trajectories of its occupants. Each event corresponds to unique combination of time of occurrence, zone of occurrence, and probabilities generated for an occupant's trajectory points.

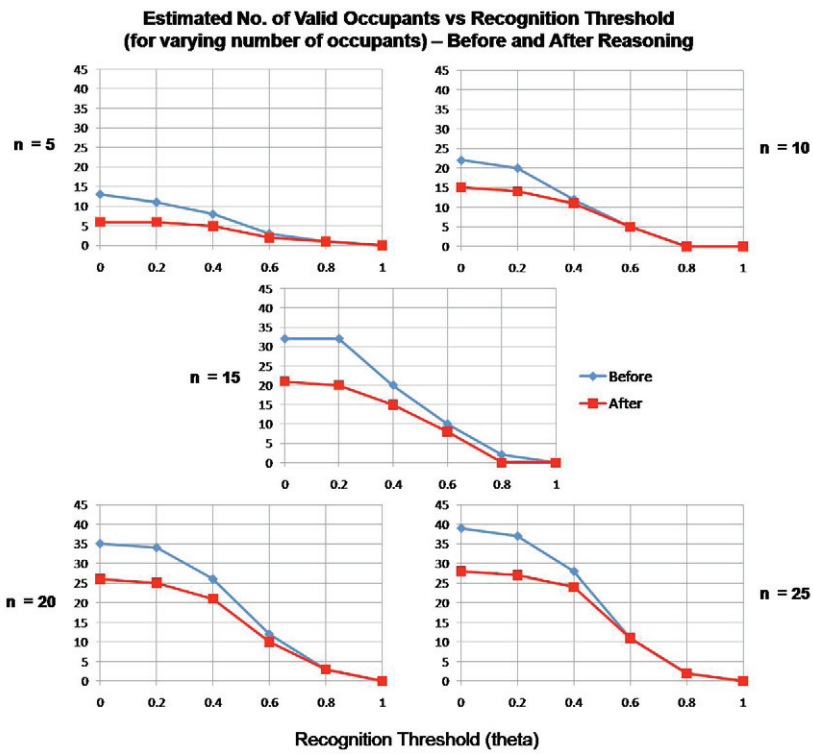


Fig. 2: Estimated Number of Occupants - Before and After Reasoning

Figure 2 shows the benefits of integrating recognition and track-based reasoning (indicated by the red curve) so as to reduce the extent of spuriously identified occupants. The benefits of reasoning are more pronounced at lower value of θ where the number of false positives are higher. Track-based reasoning on

shorter tracks is less effective in mitigating the errors due to recognition and hence the initial states of a smart environment are likely to be more error prone. In due course of time, as longer tracks are formed, the reasoning process is able to determine the subsequent states with less error and additionally can also correct the errors in the initial states. This transition function would now have the form $\Delta : \mathcal{P}(\mathcal{S}) \times E \rightarrow \mathcal{P}(\mathcal{S})$. Here the function takes a set of states as input, computes the tracks from these states, and determines as output the next state along with a revised set of previous states.

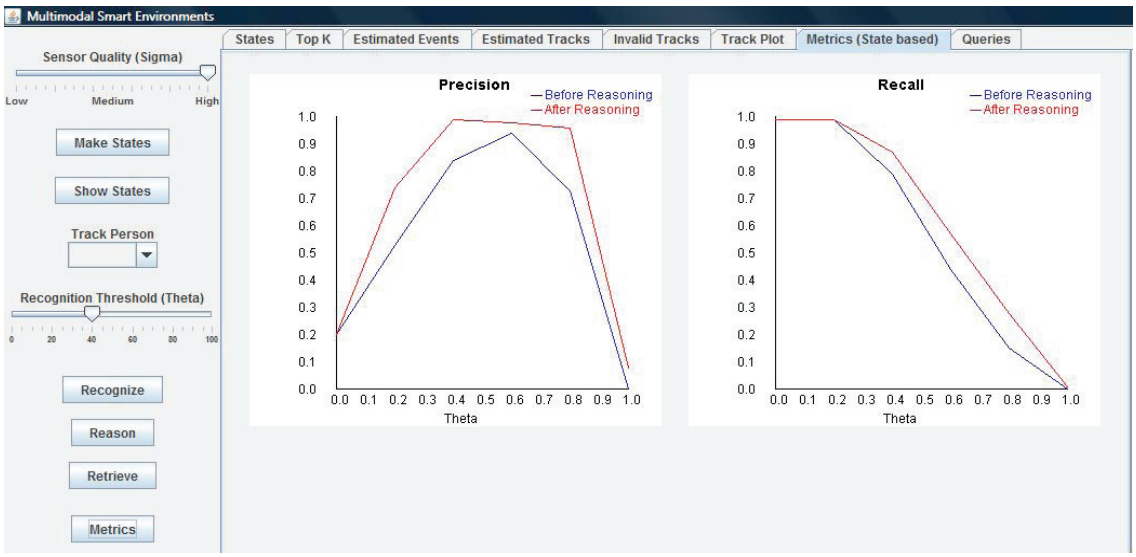


Fig. 3: Precision and Recall Plots (State based)

Figure 3 is a screen shot of precision and recall metrics produced by our prototype. Average precision and recall for varying values of recognition threshold θ are shown. As theta increases, the average precision increases up to $\theta = 0.6$, and then declines. At low values of θ, a high proportion of false positives makes the average precision low. The proportion of false positives in the set of recognized occupants reduces with increasing θ, until a point of inflexion from where the true positives also fail to get recognized, resulting in a drop in average precision.

Average recall on the other hand decreases with increasing θ. At low values of θ, there are hardly any false negatives, thereby leading to a recall value of nearly 1. As the θ increases, the proportion of false negatives increases which in turn reduces the average recall. Average precision and average recall drop to 0 at $\theta = 1.0$, as the true positives are not recognized.

Parameter	Before	After
Total Number of Events	156	156
Correctly Identified Events	115	126
Misidentified Events	41	30
Error Percentage	26.28	19.23
Valid Tracks	12	10
Spurious Tracks	31	5

Table 1: Improvements from Spatio-temporal reasoning

Figure 3 also shows the clear improvement in precision and recall as a result of spatio-temporal reasoning. Precision improves as events are correctly identified and recall also improves as false negatives are eliminated through correct identification of events. Our experiments show the average number of mismatched

events (between the ground truth and estimated occupants) decreases by about 10% through spatio-temporal reasoning (see table 1).

While it is possible to define precision and recall metrics with respect to any query of interest, we have formulated them in a query-independent manner. Our initial experiments show that query-dependent metrics tend to fare better than query-independent metrics, as typical queries are not concerned with every single event that occurred.

6. Conclusion

We have presented a novel model for unobtrusive identification and tracking in smart environments with a provision for integrating recognition and reasoning in a uniform manner. The two main contributions of this paper are:

1. A state transition framework in which events abstract different biometric recognition steps and transitions abstract different reasoning steps.
2. A demonstration of the improvement in the performance metrics by integrating recognition and spatio-temporal reasoning.

Our experiments show that recognition alone is insufficient to achieve the highest degree of precision and recall of a smart environment. We show that improved precision and recall is possible by augmenting recognition with spatio-temporal reasoning. While our experiments have focused on face recognition, our model can also be extended to incorporate other biometric modalities of an individual which can be fused using multimodal fusion techniques. Though we have formulated precision and recall metrics in a generic and query-independent manner to estimate the overall performance of a smart environment, these metrics can also be applied in the conventional manner with respect to a query of interest.

References

[1] A. Pentland, T. Choudhury, Face Recognition for Smart Environments, IEEE Computer 33 (2) (2000) 50–55.
[2] V. Menon, B. Jayaraman, V. Govindaraju, Biometrics Driven Smart Environments: Abstract Framework and Evaluation, in: Proc. of the 5th International Conference on Ubiquitous Intelligence and Computing (UIC-08), Springer-Verlag, 2008, pp. 75–89.
[3] V. Menon, B. Jayaraman, V. Govindaraju, Integrating Recognition and Reasoning in Smart Environments, in: Proc. of the 4th IET International Conference on Intelligent Environments (IE'08), 2008, pp. 1–8.
[4] V. Menon, B. Jayaraman, V. Govindaraju, Multimodal identification and tracking in smart environments, Personal and Ubiquitous Computing 14 (8) (2010) 685 – 694.
[5] C. J. van Rijsbergen, Information Retrieval, London: Butterworths, 1979.
[6] K. Bernardin, R. Stiefelhagen, A. Waibel, Probabilistic integration of sparse audio-visual cues for identity tracking, in: MM '08: Proceeding of the 16th ACM international conference on Multimedia, ACM, 2008, pp. 151–158.
[7] Y. Gao, S. C. Hui, A. C. M. Fong, A MultiView Facial Analysis Technique for Identity Authentication, IEEE Pervasive Computing 2 (1) (2003) 38–45.
[8] K. Hong, J. Min, W. Lee, J. Kim, Real Time Face Detection and Recognition System Using Haar-Like Feature/HMM in Ubiquitous Network Environments, in: Computational Science and its Applications (ICCSA 2005), Vol. 3480 of Lecture Notes in Computer Science, Springer, 2005, pp. 1154–1161.
[9] K. Bernardin, R. Stiefelhagen, Audio-visual Multi-person Tracking and Identification for Smart Environments, in: Proc. of the 15th International Conference on Multimedia (MULTIMEDIA '07), ACM, 2007, pp. 661–670.
[10] J. Luque, et al., Audio, Video and Multimodal Person Identification in a Smart Room, in: R. Stiefelhagen, J. Garofolo (Eds.), Multimodal Technologies for Perception of Humans, Vol. 4122 of Lecture Notes in Computer Science, Springer, 2007, pp. 258–269.
[11] Z. Ghahramani, Adaptive Processing of Sequences and Data Structures, Lecture Notes in Artificial Intelligence, Springer-Verlag, 1998, Ch. Learning Dynamic Bayesian Networks, pp. 168–197.
[12] A. Gerevini, B. Nebel, Qualitative spatio-temporal reasoning with rcc-8 and allen's interval calculus: Computational complexity, in: Proc. European Conf. on AI (ECAI 2002), IOS Press, 2002, pp. 312–316.
[13] OpenCV, `http://www.intel.com/technology/computing/opencv/index.htm`.
[14] R. Hewitt, Seeing With OpenCV: Implementing Eigenface, SERVO Magazine (2007) 44–50.
[15] H. Cao, V. Govindaraju, Vector Model Based Indexing and Retrieval of Handwritten Medical Forms, in: International Conference on Document Analysis and Recognition (ICDAR07), IEEE Computer Society, 2007, pp. 88–92.