Multiple-Sensor Indoor Surveillance System

Valery A. Petrushin[1)], Omer Shakil[2)], Damian Roqueiro[3)], Gang Wei[1)], Anatole V. Gershman[1)]
[1)] *Accenture Technology Labs, Chicago, IL, USA*
[2)] *University of Texas at Austin, Austin, TX, USA*
[3)] *University of Illinois at Chicago, Chicago, IL, USA*
valery.a.petrushin@accenture.com

Abstract

This paper describes a surveillance system that uses a network of sensors of different kind for localizing and tracking people in an office environment. The sensor network consists of video cameras, infrared tag readers, a fingerprint reader and a PTZ camera. The system implements a Bayesian framework that uses noisy, but redundant data from multiple sensor streams and incorporates it with the contextual and domain knowledge. The paper describes approaches to camera specification, dynamic background modeling, object modeling and probabilistic inference. The preliminary experimental results are presented and discussed.

1. Introduction

The proliferation of a wide variety of sensors (video cameras, microphones, infra-red badges, RFID tags, etc.) in public places such as airports, train stations, streets, parking lots, hospitals, governmental buildings, shopping malls, and homes has created the opportunities for development of security and business applications. Surveillance for threat detection, monitoring sensitive areas to detect prohibited or unusual events, tracking customers in airports and in retail stores, monitoring movements of assets, and monitoring elderly and sick people at home are examples of some applications that require the ability to automatically detect, recognize and track people and other objects by analyzing multiple streams of often noisy and poorly synchronized sensory data. A scalable system built for this class of tasks should also be able to integrate this sensory data with contextual information and domain knowledge provided by both the humans as well as the physical environment to maintain a coherent picture of the world over time. While video surveillance has been in use for several decades, systems that can automatically detect and track people (or objects) using multiple streams of heterogeneous and noisy sensory data is still a great challenge and an active research area. Since the performance of these systems is not at the level at which they can work autonomously, there are human experts who are still part of the loop. Many approaches have been proposed for object tracking in recent years. They differ in various aspects such as number of cameras used (single [1], two [2] or more [3-5] cameras), type of cameras and their speed and resolution, type of environment (indoors or outdoors), area covered (a room or a hall, a hallway, several connected rooms, a parking lot, a highway, etc.), and location of cameras (with or without overlapping fields of view), using different approaches to background modeling, object modeling (2D or 3D representations, color and/or shape models), and different inference techniques. However, the performance of most systems is still far from what is required for real-world applications.

2. Multiple Sensor Indoor Surveillance Project

This research is a part of Multiple Sensor Indoor Surveillance (MSIS) project. The backbone of the MSIS environment consists of 32 AXIS-2100 webcams, a pan-tilt-zoom (PTZ) camera, a fingerprint reader and an infrared badge ID system (91 readers that are installed on the ceiling) that are sensing an office floor (Figure 1). The webcams and infrared badge system cover two entrances, seven laboratories and demonstration rooms, two meeting rooms, four major hallways, four open space cube areas, two discussion areas and an elevator area. Some areas are covered by multiple cameras, the maximum overlap being with up to four cameras. The total area covered is about 18,000 sq. ft. (1,670 sq. m). The fingerprint reader is installed at the entrance and is used for matching an employee

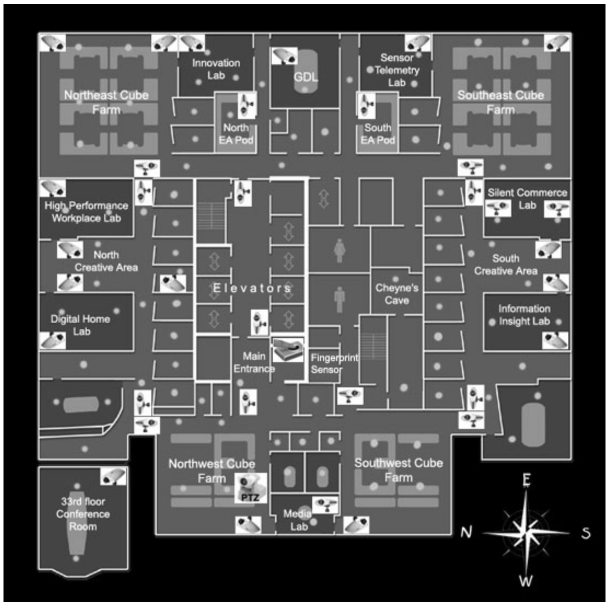

Figure 1. Locations of sensors on the floor. Here IR badge readers are represented as dots; cameras are represented by small images that show their orientation.

with his/her visual representation. The PTZ camera is watching the main entrance and northwestern cube area, and is used for face recognition.

The architecture of the system consists of three layers. The bottom layer deals with real-time image acquisition and feature extraction. It consists of several networked computers, with each computer running an agent that receives signals from 3-4 webcams, detecting "events", storing images for that event in the image repository in JPEG format, extracting features and saving them in the database. The event is defined as any movement in the camera's field of view. The average signal sampling frequency is about 3 frames per second. Three more agents acquire and save in the corresponding databases information about events detected by the infrared badge ID system, and the results of finger-print and face recognition. The event databases serve as a common resource for applications of higher levels.

The middle layer consists of a set of application agents which use the features extracted at the bottom layer. The results of these agents go to the databases. Depending on the objective of the application it may use one, several or all cameras and some other sensors.

The top layer consists of a set of meta-applications, which use the results of the middle layer applications, integrate them, derive behavioral patters of the objects and maintain the consistency of results. The applications of this layer are also responsible for maintaining the databases of the system and creating reports on its performance.

The MSIS project has the following objectives.

- Create a realistic multi-sensor indoor surveillance environment.
- Create an around-the-clock working surveillance system that accumulates data in a database for three consequent days and has a GUI for search and browsing.
- Use this surveillance system as a base for developing more advanced event analysis algorithms.

The following agents or applications have been considered.

- Creating a real-time image acquisition and feature extraction agent.
- Creating an event classification and clustering system.
- Counting how many people are on the floor.
- Creating a people localization system that is based on evidence from multiple sensors and domain knowledge.
- Creating a system that recognizes the behavior patterns of people.
- Creating a system that maintains the consistency of dynamic information about the events that was collected or derived by the other agents.

The above mentioned applications are currently at different stages of completeness. This paper describes our approach to background modeling and presents the people localization system that is based on evidence from multiple sensors and domain knowledge.

3. Feature Extraction

3.1. Camera Specification

We are using for the surveillance task multiple static cameras with low frame sampling rate (3-5 Hz) which is typical for Web cameras. The advantages of indoor environments comparing to the outdoor ones are the following: there are no sharp shadows, illumination changes rather slow, and the speed of the objects is low, because the objects of interest are mostly people. Besides, we can use our knowledge for specifying important areas, such as, working places in cubicles, and unimportant areas, such as reflecting surfaces, in a camera's view. The disadvantages are that many places in an indoor environment are unobservable; people can easily change the direction of movement and the people can be often occluded by furniture or the other people.

Each camera has a specification that includes the following data.

- *Operating zone* is an area that is used for feature extraction. For some cameras only part of their view area is worth to use. Having smaller operating zone expedites processing.
- *Background modeling type* sets the type of background modeling for the camera. The following background models are currently supported: single frame and median filtering.
- *Indicators* are some small areas and associated with them recognizers that allow detecting some local events such as light in an office is on/off, a door is open/closed, etc. The indicators play a double role – they can be used to improve the background modeling, and they are additional pieces of evidence about the state of the environment.
- *Important areas* are areas that the surveillance system pays special attentions, such as doorways, working places in cubicles, armchairs in a hall, etc.
- *Unimportant areas* are areas that must be ignored because they are sources of noise. Such areas are reflective surfaces, screens of TV and computer monitors, etc.
- *Camera calibration data* is location of markers that allow estimating distances to objects in cameras' views. This data is used for estimating objects' location, their geometrical features and speed.

Figure 2 gives an example of camera specification. Here there are two indicators that detect such events as lights are on/off (in a meeting room on the left) and the door to the meeting room is open/closed. Areas of indicators are represented as black and white solid rectangles. The light indicator uses average intensity and a threshold as a recognizer, and the door indicator checks for a horizontal edge as a recognition that the door is open. Three black dotted rectangles on the left represent important areas – the working places in cubicles. Three white dashed rectangles on the right mark unimportant areas, which are surfaces that get shadow when a person is passing by. White markers on the floor are used for camera calibration. Five white dash-dotted rectangles split the walkway into zones. A tool with a GUI has been designed to facilitate camera specification and semi-automatically generate recognizers' parameters.

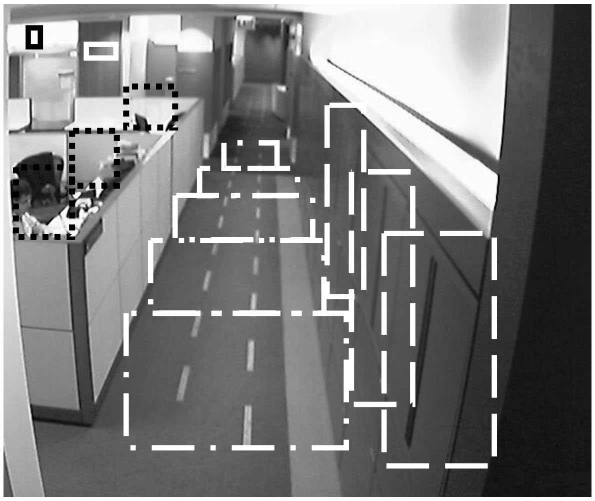

Figure 2. An example of camera specification.

3.2. Background Modeling

The objective of background modeling techniques is to estimate value of pixels of the background frame, i.e. the frame without any moving objects. Then this frame is subtracted pixel-by-pixel from a current frame to detect the pixels that belong to moving objects (foreground pixels). Many approaches for background modeling have been developed [6, 7].

The simplest approach is to use a single frame which is acquired and periodically updated when no motion or changes within a sequence of frames is detected. The advantage of the single frame method is that it does not require any resources for maintaining the model. The disadvantage is that it does not work for scenes with intensive motion.

Another approach is to accumulate and maintain a pool of N frames, where N is an odd number. The value of each pixel of the background model is estimated as median of the corresponding pixels of frames from the pool. This model is called the median filter background model. It works well when each pixel of the scene is covered by moving objects less than 50% of time. The advantage of the median filter model is that it can be used for scenes with high motion, for example, for a camera that is watching cubicles. The disadvantage is that it requires additional memory for storing the pool of frames and computations for maintaining the model [6]. More advanced techniques are required for modeling backgrounds that can have several values for a pixel. One of the approaches is to use Gaussian mixture models (GMM) [7].

Real-time processing puts additional restrictions on background modeling. The real-time system cannot wait to accumulate training data, train the models, and then catch up by processing all postponed frames. It does not have enough processing power to implement such scenario. That is why we adopted an approach that uses two background modeling techniques and switches between them when it is needed. The system loops through the basic cycle that consists of the following steps: image acquisition, motion detection, if motion is not detected then the system is idle till the next cycle otherwise it processes the image, which includes extracting foreground pixels, extracting and storing features and maintaining the background model. A sequence of cycles that have motion forms a dynamic event. The dynamic events are separated by events with no motion or static events.

If the camera uses the single frame model then the system starts by acquiring a single frame background during a static event and updates it not less than in T seconds by picking up a frame from a static event that lasts not less than D seconds. The parameters T and D can be specified for each camera (the default values are $T = 120$ and $D = 60$). There are three kinds of events that require attention for robust background modeling. The fist one is a flicker, which is a short abrupt change in illumination due to camera noise. The second kind of events is a local luminosity change, for example, lights get on or off in an office, which is a part of the scene. And the third kind of events is global luminosity change when more then 60% of pixels are changed, for example, when lights get on/off in the room, which is observed by the camera. The system reacts differently for each kind of events. When a flicker occurs, the system skips the frame. When a local luminosity change occurs, the system recognizes this case using an indicator and patches the *affected area* with values extracted from the new frame. If the system uses an adaptive background model it first generates a single frame model and then patches the affected area. In case of global change the system acquires a new single frame. If the camera uses an adaptive background model then the system acquires a single frame model and starts accumulating data for creating a new adaptive model. It picks up frames from static events using parameters T and D. When the desired number of frames is reached the systems generates and switches to the adaptive model and begins to maintain it.

As an adaptive model we used the median filter model with a pool of size N=51. Maintaining the model requires discarding the oldest frame, adding a new one and sorting the values for each pixel. The system uses a version of a dynamic deletion-insertion algorithm to avoid sorting and improve the speed of model maintenance.

For dynamic events the background model is subtracted from the current frame for detecting foreground pixels. Then some morphological operations are applied to remove noise and shadows. After this, the foreground pixels are separated into blobs using the calibration information for each camera. Finally a set of candidate blobs is selected for feature extraction.

3.3. Visual Feature Extraction

A person's most distinguishable visual identity is his or her face. However, in many practical applications the size and image quality of the face do not allow traditional face recognition algorithms to work reliably, and sometimes the human face is not visible at all. Therefore, our people localization system uses face recognition as an auxiliary means that is applied for only some areas of some cameras. The other salient characteristic of a person are sizes of the body, color of the hair and color of the clothes that is on the person. At any given day, a person usually wears the same clothes, and thus the color of person's clothes is consistent and good discriminator (unless everybody wears a uniform). We use color histograms in different color spaces as major features for distinguishing people based on their clothes. The blob is processing in the following manner. The top 15% of the blob, which represent the head, and bottom 20%, which represent the feet and also often include shadow, are discarded and the rest of the region is used for feature extraction. We used the color histograms in RGB, normalized RGB, and HSV color spaces with the number of bins 8, 16 and 32.

After some experiments we chose the 8 bin color histogram in the normalized RGB space for the red, green, and luminosity components, which gave a good balance between computation efficiency and accuracy.

3.4. People Modeling

We used several approaches for modeling a person based on his or her appearance. The simplest one is to use all pixels of the blob for training a color histogram. Another approach is to fit a Gaussian or a Gaussian Mixture Model to the training data. A more elaborate person modeling includes two models – one for top and another for the bottom part of the body.

Let us assume that we built a model for a human H. To estimate how well the data D extracted from a new blob R fits the model, we can consider the model as a probability density function and estimate the likelihood of the dataset using equation (1), which assumes that pixels' values are independent. The dataset D can include all pixels of the blob or a randomly selected subset of particular length (usually 50-100 pixels are enough for reliable classification). In case when two (top and bottom) models are used for modeling, the equation (1) should be extended to include products of likelihoods for each model over corresponding data points. In practice, a log-likelihood function is used instead of likelihood one.

$$L(D \mid H) = \prod_{i=1}^{N} p(x_i \mid H) \tag{1}$$

where $x_i \in D$ are points of the dataset D, $p(\cdot \mid H)$ is the probability distribution function for the model H, N is the number of points in the dataset D.

The type of probability distribution function depends on the type of the model used. For example, in case of color histogram it can be approximated as a product of corresponding values for pixel's components.

4. People Localization

This section presents a Bayesian framework for people localization which allows the integration of evidence of multiple sensor sources. Our task is to localize and track N objects in a space of known geometry with stationary sensors of different kinds. The number of objects may change dynamically over time when an object arrives or leaves. The sensing zones for some sensors can overlap. We assume that there are two types of objects: known objects (employees) and unknown objects (guests or customers). The space is divided into "locations". Time is sampled into ticks. The tick duration is selected depending on the sampling frequencies of the sensors. It should be large enough to serve as a synchronization unit and small enough so that objects can either stay in the same location or move only to an adjacent one within a tick.

4.1. Sensor Streams

Each object is represented by a set of features extracted from sensor streams. We are currently using four sources of evidence.

4.1.1. Video Cameras

This is very rich data source, but requires a lot of sophisticated processing to extract useful information. Our system is mostly based on this source. We are using two approaches for people localization – people appearance modeling and face recognition. People appearance modeling is based on color features. An object can have several color models – one or more for each location or even for the time of the day. Object models can be defined (through training) prior to the surveillance task or accumulated incrementally during the task. Appearance modeling works for all cameras, whereas face recognition is efficient only for some cameras where size and orientation of faces are appropriate. We use a dedicated PTZ camera that watches the main entrance to the floor for face recognition. The face recognition system uses the OpenCV algorithm [8] and tries to recognize people from a restricted list.

4.1.2. Infra-Red Badge ID System

The second source of evidence is the infra-red (IR) badge system. The system collects data from 91 readers and merges them into a database that indicates where a particular badge was sensed the last time. This source of information is not very reliable because of the following reasons. (1) the badge has to be in the line of sight of a reader on the ceiling (if a person puts his/her badge into a pocket, it cannot be detected); (2) the orientation of the badge affects the detection; (3) a person can leave his/her badge in the office or give it to another person; (4) detection records are written to the database with a delay creating a discrepancy among different sources of evidence for fast moving objects.

Before processing IR badge sensor signals the system must determine and maintain the list of active sensors, which are sensors that both transmit signals and move in space.

4.1.3. Finger Print Reader

The third source of evidence is the fingerprint reader. This is a very reliable source, but located only at the main entrance, has a restricted number of registered users, and a person only uses it one or two times per day for check-in. We mostly use it for acquisition or updating of person appearance models as a person checks-in when entering the office.

4.1.4. Human Intervention

The fourth source of evidence is human intervention. People who participate in a surveillance task can interactively influence the system. They can mark an object in a camera view and associate it with a particular person, which causes the system to set the probability of the person being at this location to 1 and recalculate the previous decisions by tracking the person back in time. This is a very reliable, but costly information source. We use it mostly for initializing and updating person appearance models.

4.2. Identification and Tracking of Objects

The current state of the world is specified by a probability distribution of objects being at particular locations at each time tick. Let us assume that $P(H_i|L_j)$, $i=1,N$, $j = 1,K$ are probabilities to find the object H_i at location L_j. The initial (prior) distribution can be learned from data or assumed to be uniform.

In case of dealing with multiple sensors and multiple locations the major challenge is synchronization of multiple sensors in time and space.

On one hand, some sensors such as video cameras can have large fields of view that can be divided into non-overlapping locations. On the other hand, the sensing zones of different sensors can intersect. These intersections can be considered as natural locations. Sometimes the borders of locations are fuzzy.

The concept of location allows making more precise localization, and having person models for each location to improve person identification. On the other hand, we have to create person models for each person and for each location which often is not possible because the person cannot visit all locations during the day. We assume that a person may have models only for some locations and the system uses the model for the closest location.

Each person also has a transition matrix $T(H_k) = \{t_{ij}(H_k)\}$ $k=1,N$, $i,j = 1,K$ that specifies the probability of person transition from i-th to j-th location.

The major concern for a multi-sensor environment is accuracy of data synchronization. Different sensors may have different sampling rates and can acquire signals in non-regular intervals. In general, surveillance cameras are not synchronized. However, computers' clocks can be synchronized and time stamps can be assigned to frames. This means that we cannot synchronize frames, but we can select frames that belong to time interval of some duration (time tick). The time tick should be big enough to contain at least one frame from each camera, but be small enough to allow people moving only inside the current location or to the one of adjacent locations. In our experiments time tick is equal to one second.

The process of identification and tracking of objects consists of the following steps:

Step 1. Data Collection and Feature Extraction. Collect data from all sensors related to the same time tick. Select data that contains information about a new "event" and extract features.

Step 2. Object Unification from Multiple Sensors. Each sensor detects signals of one or more objects in its sensory field. The signals that come from the same object are merged based on their location and sensory attributes. This gives us a unified model of how different sensors "see" the same entity. For video cameras, the blobs are first mapped into locations based on their coordinates and calibration data from the cameras. Then the blobs from different cameras that belong to the same location are assigned to the same entity based on their color features. For IR badge data, which consists of binary indicators of a badge being detected at a particular location, the system first spreads the probability to the adjacent IR locations taking into account the space geometry, and then maps IR locations into camera based locations and associates evidence with entities. The result is a set of entities $\boldsymbol{O} = \{O_r\}$ and a matrix $\boldsymbol{W} = \{w_{kr}\}$ $k=1,K$, $r=1,M_0$, where M_0 is the number of entities. Each w_{kr} is the membership value of r-th entity to belong to the k-th location.

Step 3. Motion Estimation. The locations are selected in a way that an object can either stay in the same location or move to an adjacent location during any single time tick. The specific transition probabilities among locations for a known object or generalized transition probabilities for the other objects are estimated from historical data or provided as prior knowledge by the people involved in the task. These probabilities are taken into account for re-estimating prior probabilities using equation (2).

$$\widetilde{P}(H_i \mid L_j) = \frac{\left[\sum_{k=1}^{L} P(H_i \mid L_k) \cdot t_{kj}(H_i)\right] \cdot P(H_i \mid L_j)}{\sum_{l=1}^{L}\left[\sum_{k=1}^{L} P(H_i \mid L_k) \cdot t_{kj}(H_i)\right] \cdot P(H_i \mid L_l)} \qquad (2)$$

This is a kind of motion prediction in case when we don't know anything about the person except that he/she was previously in a particular location. Adding more information to a person state, such as direction of movement, velocity, acceleration, etc., allows applying more advanced tracking techniques, such as Kalman filtering, particle filtering or Bayesian filtering.

Step 4. Posterior Probability Estimation Using the features that belong to the same entity and the person models, the conditional probability that the entity represents a person at a given location is estimated for all entities, objects and locations. The result is a sequence of of probabilities $S_r=\{P(R_j,L_k,C_q|H_{i,})\}$ associated with the entity O_r. $r=1,M_0$.. Here R_j $j=1,M_r$ are the feature data extracted from representations of entity O_r, and C_q, $q = 1,Q$ are sensors. For video cameras, the probabilities that a blob represents an object (person) for given cameras and locations are calculated using blob's features and persons' models (see equation (1)). For IR badge data the probabilities distributed to adjacent locations are used as the conditional probabilities. The output of face recognition system is also used as conditional probabilities. The fingerprint and human intervention evidence sets up the prior probabilities directly.

For each entity the estimates of likelihood that the entity represents a particular person at a given location are calculated. If all estimates are less than a threshold, then the entity is marked as "unknown", and a new ID and a new model are generated. Otherwise, the conditional probabilities of signals that are views of the same entity from different sensors are used for estimating posterior probabilities of a person being represented by the entity at the location using Bayes rule (3) and the person's ID that maximizes the conditional probability is assigned to the entity. Then the model of the just assigned person is excluded from the model list for processing the other entities.

$$P(H_i \mid O_r, L_k) = \frac{\widetilde{P}(H_i \mid L_k) \cdot w_{kr} \cdot \prod\limits_{P(R_j, L_k, C_q) \in S_r} P(R_j, L_k, C_q \mid H_i)}{P(O_r)} \tag{3}$$

where

$$P(O_r) = \sum_{i=1}^{N} \widetilde{P}(H_i \mid L_k) \cdot w_{kr} \cdot \prod_{P(R_j, L_k, C_q) \in S_r} P(R_j, L_k, C_q \mid H_i) \cdot$$

Then the probabilities for the entity are normalized over locations using (4).

$$P(H_i \mid L_k) = \frac{P(H_i \mid O_r, L_k)}{\sum_{k=1}^{L} P(H_i \mid O_r, L_k)} \tag{4}$$

Step 5. Re-Estimation.

The steps 1-4 are repeated for each time tick.

Step 6. Post-processing.

This step includes some smoothing procedures for whole events and truth maintenance procedures, which use problem domain knowledge to maintain probabilities when no data available. In case when an object is temporarily invisible, the truth maintenance procedures mark it as "idle" and keep its probability high to be in "hidden" locations that are near the location where the object has been identified last time. The system uses two cameras that watch the elevator area and detects people who are entering or leaving the floor. If a person leaves the floor, his/her model is marked as "inactive". If a person enters the floor, a new object and its appearance model is created and is marked as "new". The system tracks a new object and creates models for it for other locations when it is possible.

5. Experimental Results

For evaluation we used 15 cameras and 44 IR badge readers that are located in the northern half of the floor. In the first experiment we evaluated the system's performance in the closed set case. It means that the system had models for all 15 people, who participated in the experiment. The second experiment was designed for evaluating the system's performance for the open set problem. Besides 15 "known" people it included 10 "unknown" people, i.e. people whose models were not available at the beginning and created during the process. Each experiment lasted for four hours. In both experiments two evaluations have been done. The first evaluation estimated the accuracy of people localization for each camera separately, and then calculated the average for each person. The second evaluation merged the results from all cameras and IR badge readers. Only seven people of 15 "known" (marked by stars in the Table 1) and none of "unknown" had active IR badges. The results were compared to the ground truth data created manually. Table 1 presents results for both experiments. We can see that in case of closed set problem the average recognition accuracy is about 11% higher than for the open set problem for both single camera and integrated evaluations. For the closed set problem the accuracy of localization for individual person is in range from 44% to 99%.The low accuracy for some people can be explained by the following: (1) poor blob extraction for people who are sitting still for a long time; (2) poor blob extraction when a person is (partially) occluded; (3) poor blob separation in the hallways; (4) several people have similar models. Merging evidence from several cameras and IR sensors improves the performance for both cases for about 6%. For the closed set problem the largest improvement (23.7%) was mostly due to IR badge data. In this case, a person ID1006 stayed in the same location for a long time with his badge active. But sometimes merging IR badge data causes a decrease of localization accuracy. It is happen for transient events in the hallways because of poor alignment of visual and IR data (see results for ID1015 and ID1023). The increase from merging visual evidence from several cameras can reach up to 9%.

For the open set problem the localization accuracy for individual person lies in the range from 25% to 94% Low accuracy for some people can be mostly attributed (besides the above mentioned reasons) to confusion with people who have similar models. Merging additional evidence can improve performance up to 15%. Merging only visual evidence from several cameras improves performance by 7-8%.

Person ID	Closed Set Accuracy			Open Set Accuracy		
	Single Camera	Cameras and IR badge	Difference	Single Camera	Cameras and IR badge	Difference
1000	82.14%	87.56%	5.42%	71.39%	77.31%	5.92%
1002	99.27%	99.51%	0.24%	94.54%	95.81%	1.27%
*1003	86.88%	91.07%	4.19%	81.91%	86.10%	4.19%
1005	74.32%	81.69%	7.37%	62.73%	70.59%	7.86%
*1006	45.98%	69.68%	23.70%	43.78%	58.58%	14.80%
*1015	44.08%	41.12%	-2.96%	26.32%	26.58%	0.26%
1020	67.03%	76.71%	9.68%	60.75%	69.24%	8.49%
*1023	64.43%	60.77%	-3.66%	59.82%	58.49%	-1.33%
1024	41.26%	50.72%	9.46%	25.08%	32.78%	7.70%
*1025	69.26%	78.29%	9.03%	57.88%	66.81%	8.93%
1026	71.06%	73.66%	2.60%	41.34%	46.19%	4.85%
*1027	62.04%	73.41%	11.37%	58.13%	67.08%	8.95%
*1029	51.21%	57.88%	6.67%	51.50%	57.21%	5.71%
1064	77.81%	83.42%	5.61%	44.69%	51.69%	7.00%
1072	66.07%	74.49%	8.42%	52.53%	59.67%	7.14%
Average	**66.86%**	**73.33%**	**6.47%**	**55.49%**	**61.61%**	**6.12%**

Table 1. Accuracy of people localization for open and closed set problems

6. Summary

In this paper we described the MSIS project's components for real-time background modeling using both single frame and adaptive models. We also described the people localization systems, which uses a Bayesian framework that enables us to robustly reason from data collected from a network of various kinds of sensors.

As to future work, we see that the system could be improved on many levels. On the low level it needs more robust background modeling and blob extraction and blob separation techniques, search for better features and reliable dynamic modeling of people and other objects' appearance. On the middle level it needs using more advanced tracking approaches such as non-linear filtering and sensorial data fusion approaches. On the high level the system needs more efficient decision merging approach, which can use domain specific knowledge and can produce a consistent "big picture" of events in the area under surveillance. We also plan to put efforts into developing more attractive visualization techniques and a useable user interface.

7. References

[1] Siebel, N.T. and Maybank, S. Fusion of Multiple Tracking Algorithms for Robust People Tracking. Proc. 7th European Conference on Computer Vision (ECCV 2002), Copenhagen, Denmark, May 2002; IV: 373-387.

[2] Krumm, J., Harris, S., Meyers, B., Brumitt, B. , Hale, M. Shafer, S. Multi-camera Multi-person Tracking for EasyLiving. Proc. 3rd IEEE International Workshop on Visual Surveillance, July 1, 2000, Dublin, Ireland.

[3] Mittal ,A. and Davis, L.S. M2Tracker: A Multi-view Approach to Segmenting and Tracking People in a Cluttered Scene. International Journal of Computer Vision, 2003; 51 (3): 189-203.

[4] Cai, Q. and Aggarwal, J.K. Tracking Human Motion in Structured Environments using a Distributed-camera System. IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 2, No. 11, November 1999: 1241-1247.

[5] Javed, O. , Rasheed, Z., Atalas, O. and Shah, M. KnightM: A real Time Surveillance System for Multiple Overlapping and Non-overlapping Cameras. The fourth IEEE International Conference on Multimedia and Expo (ICME 2003), Baltimore, MD, July 6-9, 2003.

[6] Cheung S-CS, Kamath Ch. Robust techniques for background subtraction in urban traffic video. Proc. of SPIE, Visual Communications and Image Processing 2004, S. Panchanathan, B. Vasudev (Eds), January 2004; 5308: 881-892

[7] Stauffer C, Grimson WEL, Adaptive Background Mixture Models for Real-Time Tracking, Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1999; 246-252.

[8] Nefian, A.V., Hayes, M.H. III. "Maximum likelihood training of the embedded HMM for face detection and recognition", IEEE International Conference on Image Processing, September 2000; 1: 33-36.