

# Using Bayesian Reasoning from Sensor Network for Indoor Surveillance

Valery A. Petrushin, Gang Wei, Rayid Ghani, Anatole V. Gershman

Accenture Technology Labs, 161 N. Clark St., Chicago, IL 60601, USA  
{valery.a.petrushin,gang.wei,rayid.ghani,anatole.v.gershman}@accenture.com

**Abstract.** In this paper we define a Bayesian framework that uses noisy, but redundant data from a network of sensors that include multiple sensor streams of different types. It merges the data with the contextual and domain knowledge that is provided by both the physical constraints imposed by the local environment and by the people that are involved in the surveillance tasks. The paper also presents the results of applying the Bayesian framework to the people localization problem in indoor environment using a sensor network that consists of video cameras, infrared tag readers and a fingerprint reader.

## 1. Introduction

The proliferation of a wide variety of sensors (video cameras, microphones, infrared badges, RFID tags, etc.) in public places such as airports, train stations, streets, parking lots, hospitals, governmental buildings, and shopping malls has created many opportunities for homeland security and business applications. Surveillance for threat detection, monitoring sensitive areas and detecting unusual events, tracking customers in retail stores, controlling and monitoring the movement of assets, and monitoring elderly and sick people at home are just some of the applications that require the ability to automatically detect, recognize and track people and other objects by analyzing multiple streams of often unreliable and poorly synchronized sensory data. A scalable and robust system built for this task should also be able to integrate this sensory data with contextual information and domain knowledge provided by humans to maintain a coherent logical picture of the world over time. While video surveillance has been in use for decades, systems that can automatically detect and track people (or objects) in multiple locations using multiple streams of heterogeneous and noisy sensory data is still a great challenge and an active research area. Many approaches have been proposed for video surveillance in recent years [1-6]. They differ in various aspects such as number of cameras used, type of cameras (grayscale or color, mono or stereo, CCD or Webcams, etc.) 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). Below we shall focus on some research that deals with indoor people identification and tracking.

A system described in work [2], which is a single camera system that was created for tracking people in subway stations, used the luminance contrast in YUV color space to separate people blobs from the background. The coordinates and geometric features of the blobs are estimated and two-way matching matrices algorithm has been used to track (overlapping) blobs.

In the Microsoft's EasyLiving project [3] two color stereo cameras have been used for real-time identification and tracking up to three people in a rather small room (5 m by 5 m). The system evaluates 3D models of blobs and clusters them to fit a people-shaped blob models. Then the centroids of the blobs are projected into the room ground plan. The quantized RGB color histogram and histogram intersection are used for person's identity maintenance. A histogram is estimated for each person viewed by each camera in each visited cell of 10x10 grid of the floor plan. The person tracker module keeps the history of the person's past locations and uses it to predict current location. If the predicted location contains several candidates then color histograms are used to disambiguate them. If no candidates found the system keeps unsupported person tracks active until new data arrive. For supported track their histories are updated and new predictions are calculated. In spite of low image processing rate (about 3.5 Hz) the system works well with up to three people, who are not moving too fast and not wearing similarly colored outfits.

The system presented in [4] uses several non-overlapping cameras and knowledge about topology of paths between cameras. It probabilistically models the chain of observation intervals for each tracked person using Bayesian formalization of the problem. To estimate the optimal chain of observation the authors transform the maximum a posteriori estimation problem into a linear program optimization.

The approach proposed in [5] uses multiple synchronized grayscale overlapping cameras for tracking people and selecting a camera that gives the best view. The system consist of three modules: single view tracking, multiple view transition tracking and automatic camera switching. The system uses the following features for each person: locations of selected feature points, intensity of the selected feature points and geometric information related to a coarse 2D human body model. The multivariate Gaussian models and Mahalanobis distances are used for people modeling and tracking. The class-conditional distribution for spatial and spatial-temporal matching is used for the multiple view transition tracking for matching predicted location and body model size. The automatic camera switching is necessary if the person is moving out of the current camera's field of view, or the person moves too far away, or the person is occluded by another person. The system selects a camera that will contain the person over the largest time accordingly the current prediction of the person's movement. The experiments with three cameras in various indoor environments showed high robustness of people tracking (96-98%).

The KNIGHTM system [6] is a surveillance system that uses several overlapping and/or non-overlapping uncalibrated color cameras for people tracking. The system uses spatial and color Gaussian probability distributions for each person to identify and track people in one camera view. The person identification is based on voting of foreground pixels. If two or more people receive essential percentage of votes from the same region then the systems assumes that partial occlusion of people happens. In case of complete occlusion a linear velocity predictor is used for disambiguation. In order to track people across multiple cameras the system during the training period

learns the field of view lines of each camera as viewed in the other cameras. This information and knowledge of cameras' location are used for identification of moving people. The experiments with three cameras and three different camera setups gave promising results.

The M2Tracker system [7] uses from 4 to 16 synchronized cameras to track up to six people walking in a restricted area (3.5 m by 3.5 m). The system identifies people using the following models for segmenting images in each camera view: color models at different heights, presence probabilities along the horizontal direction at different heights, and ground plane positions tracked using a Kalman filter. Then the results of one camera segmentation are matched for pairs of cameras to estimate 3D models for each person and estimate the object location on the ground plane using Gaussian kernels to create location likelihood map. The system merges results from several pairs of cameras until the ground plane positions are stable. Then the current positions of people are updated and new predictions are calculated. Due to high computational complexity the system cannot work in real time, but the authors hope that code optimization efforts and advances in computing will make it possible in the future.

In spite of essential progress in the field, the performance of most systems is still far from what is required for real-world applications. To bridge the gap between the needs of practical applications and the performance of current surveillance algorithms, we seek solutions in the following directions:

- Develop a framework for logical integration of noisy sensory data from multiple heterogeneous sensory sources that combines probabilistic and knowledge-based approaches. The probabilistic part is used for object identification and tracking and the knowledge-based part is used for maintaining overall coherence of reasoning.
- Exploit local semantics from the environment of each sensor. For example, if a camera is pointed at a location where people usually tend to stand, the local semantics enable the system to use the "standing people" statistical models, as opposed to a camera pointing at an office space where people are usually sitting.
- Take advantage of data and sensor redundancy to improve accuracy and robustness while avoiding the combinatorial explosion.
- Take advantage of human guidance when it is available.
- Develop robust and scalable systems that work in real environments.

This paper describes our probabilistic framework for identifying and tracking moving objects using multiple streams of sensory data. To validate our framework, we built a sensor network consisting of 30 video cameras as well as about 90 infrared tag readers and a biometric station for fingerprint reading. We present preliminary experimental results of applying our approach to creating a people localization system.

## 2. Experimental environment

This research is a part of the Multiple Sensor Indoor Surveillance (MSIS) project. The backbone of the projects consists of 30 AXIS-2100 webcams, a PTZ camera,

fingerprint reader and infrared (IR) badge system (91 reader mounted on the ceiling) that are sensing an office floor for Accenture Technology Labs. 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 elevator area. Some areas overlap 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 allows matching an employee with his/her visual representation. The backbone architecture also includes several computers, with each computer receiving signals from 3-4 webcams, detecting "events" and recording the images for that event in JPEG format. The event is defined as any movement in the camera's field of view. The signal sampling frequency is about 3 frames per second. The computers also create event records in an SQL database. Another database contains events detected by the infrared badge system. The event databases serve as a common repository for both people who are doing manual search of events and automatic analysis.

### 3. Probabilistic Framework

Our task is to localize and track  $N$  objects in a space of known geometry with stationary sensors of different kinds. The sensing zones for some sensors can overlap. The number of objects can change dynamically when an object arrives or leaves. 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 time 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 either stay in the same location or only can move to an adjacent one. Each object is represented by a set of features extracted from sensor streams. An object can have several 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. 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. Each object has a set of models that are location and sensor specific. Each object has a matrix of transition probabilities  $T(H_k) = \{t_{ij}(H_k)\}$   $k=1,N$ ,  $i,j=1,K$  that is learned from training data.

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  $\mathbf{O} = \{O_r\}$  and a matrix  $\mathbf{W} = \{w_{kr}\}$   $k=1, K, r=1, M_0$ , 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 (1).

$$\tilde{P}(H_i | L_j) = \frac{\left[ \sum_{k=1}^L P(H_i | L_k) \cdot t_{kj}(H_i) \right] \cdot P(H_i | L_j)}{\sum_{l=1}^L \left[ \sum_{k=1}^L P(H_i | L_k) \cdot t_{kj}(H_i) \right] \cdot P(H_i | L_l)} \quad (1)$$

This is a kind of motion prediction in case when we don't know anything about the person movement 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., makes possible applying more advanced tracking techniques, such as Kalman or particle 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 probabilities  $S_r = \{P(R_j, L_k, C_q | H_r)\}$  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. For IR badge data the probabilities distributed to adjacent locations are used as the 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 (2) 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.

$$P(H_i | O_r, L_k) = \frac{\tilde{P}(H_i | L_k) \cdot w_{kr} \cdot \prod_{P(R_j, L_k, C_q) \in S_r} P(R_j, L_k, C_q | H_i)}{P(O_r)} \quad (2)$$

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

Then the probabilities for the entity are normalized over locations using (3).

$$P(H_i | L_k) = \frac{P(H_i | O_r, L_k)}{\sum_{k=1}^L P(H_i | O_r, L_k)} \quad (3)$$

#### Step 5. Re-Estimation.

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

#### Step 6. Post-processing.

This step include 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.

## 4. Preliminary 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. For modeling we used color histograms in RGB and HSV color spaces for top and bottom parts of human body. The training samples were selected manually for each known person for each camera and each location that a person visited. On average the number of training samples was about 2-5% of all samples available for a given person. 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 total for each person. The accuracy of localization is estimated in the following manner. For each event of length  $T$  time ticks the system produces the result which first goes through smoothing and then is compared to the ground

truth. If a person  $A$  was correctly localized  $N_A$  times out of  $G_A$ , then the accuracy is  $N_A/G_A$ , where  $G_A$  is obtained from the ground truth labels for each time tick. If a time tick contains more than one frame then the system analyzes each frame and aggregates results based on the probabilities of localized objects. After this the event level results are aggregated by camera and by person.

The second evaluation gives the results when the data from all cameras and IR badge readers were merged. Only seven people of 15 "known" and none of "unknown" had active IR badges. The results were compared to the ground truth data created manually.

For the closed set and single camera the accuracy of individual person recognition is in the range from 44% to 99% with the average of 66.86%. It goes up to 73.33% when the data from several cameras and IR badge system were used. The low accuracy for some people can be explained by the following reasons: (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; (5) poor synchronization between camera and IR badge data.

For the open set problem with single camera the localization accuracy for individual person lies in the range from 25% to 94% with the average 55.49%. Using data from multiple cameras and IR badge system gives the average accuracy 61.61%. Low accuracy for some people can be mostly attributed (besides the above mentioned reasons) to confusion with people who have similar models.

The preliminary results look moderate, but we believe that there is plenty room for improvements in several aspects, such as signal processing (better background modeling, blob extraction and separation, feature extraction), object modeling (using more advanced appearance models that include color, geometrical, and structural features), object tracking (using better techniques for motion prediction, such as Kalman and particle filtering), and decision making (using more sophisticated rules that describe physical constraints of the environment).

## 5. Conclusions

We describe a Bayesian framework that enables us to robustly reason from data collected from a network of sensors. In most practical situations, sensors are producing streams of redundant, but noisy data. A wide variety of surveillance tasks require the ability to have scalable and robust systems that can make inferences from such noisy data spanning a large network of sensors. The probabilistic framework presented here gives us the ability to reason from this data by also incorporating the local semantics of the sensors as well as any domain knowledge that can be provided by people involved in these tasks. Although the preliminary experiments presented in this paper use video cameras and IR badges as sensors, we believe that this framework is applicable in the larger context of creating robust and scalable systems that can reason and make inferences from different kinds of sensors that are present in the world today.

## 6. References

- [1] LEE, L., ROMANO, R, and STEIN, G. Monitoring Activities from Multiple Video Streams: Establishing a Common Coordinate Frame. *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 22, No. 8, August 2000, pp. 758-767.
- [2] FUENTES, L. M. and VELASTIN, S. A. People tracking in surveillance applications. *Proc. 2nd IEEE International Workshop on PETS*, Kauai, Hawaii, USA, December, 2001.
- [3] 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.
- [4] KETTNAKER, V. and ZABIH, R. Bayesian Multi-camera Surveillance. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 23 - 25, 1999, Fort Collins, Colorado, pp. 2253-2259.
- [5] 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. 21, No. 11, November 1999, pp. 1241-1247.
- [6] 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)*, July 6-9, 2003, Baltimore, MD.
- [7] 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.