

A Bayesian Framework for Robust Reasoning from Sensor Networks

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

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

Abstract

The work described in this paper defines a Bayesian framework to use noisy, but redundant data from multiple sensor streams and incorporate it with the contextual and domain knowledge that is provided by both the physical constraints imposed by the local environment where the sensors are located and by the people that are involved in the surveillance tasks. The paper also presents the preliminary 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.

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). However, 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 if it is available.
- Develop robust and scalable systems that work in real environments.

This paper describes our probabilistic framework for identifying and tracking objects (people) using multiple streams of sensory data. To validate our framework, we built a sensor network consisting of 30 video cameras as well as several 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. 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 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 transition 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:

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. For example, for video cameras, we select all video frames that have some movement and obtain “blobs” to extract geometric and color features.

2. Object Unification from Multiple Sensors

Each sensor detects “reflections” of one or more objects in its sensory field. The reflections that come from the same object are merged based on their location and sensory attributes. This gives us an unified model of how different sensors “see” the same entity. For video cameras, the blobs are first mapped into locations based on their geometric features 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 geometric and color features. The result is a set of entities $O = \{O_r\}$ and a matrix $W = \{w_{kr}\}$ $k=1,K$, $r=1,M_0$, 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.

3. Posterior Probability Estimation

We estimate the conditional probabilities for each object being in each location during each time tick. Using the features that belong to the same entity and the object models, the conditional probability that the entity represents an object at a given location is estimated for all entities, objects and locations. The result is a sequence of probabilities $S_r = \{P(R_j, L_k, C_q | H_i)\}$ 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, which are camera and location specific. The blobs that are views of the same entity from different cameras are used for estimating posterior probabilities of an object being represented by the entity at the location using Bayes rule.

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

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

Then the probabilities for entities are merged to obtain the probability $P(H_i | L_k)$ for the object H_i being observed in location L_k .

$$\tilde{P}(H_i | L_k) = \frac{\sum_{j=1}^M P(H_i | R_j, L_k)}{\sum_{i=1}^N \sum_{j=1}^M P(H_i | R_j, L_k)}$$

4. Update Probabilities using Transition Matrices

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.

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

5. Re-Estimation.

The steps 1-4 repeat for each time tick. The sensors we consider are fixed and stationary. Knowing of where the sensors are located enables us to define areas, which the system should pay extra attention to as well as regions the system should ignore. For example, areas, such as working places in cubicles, armchairs in halls, doors, passages, entrances where new objects appear and disappear, etc. are more important to watch than blinking computer or TV monitors, lamps, reflective surfaces, etc. Using a layout for each camera that marks all important/unimportant areas and assigning semantics to them increases the accuracy of object localization. The inclusion of this domain knowledge is done through a Bayesian probabilistic framework by assigning prior probabilities, and fits naturally in the probabilistic framework that was described above.

3. Experimental environment

To validate our vision and framework, we started the Multiple Sensor Indoor Surveillance (MSIS) project. The backbone of the projects consists of 30 AXIS-2100 webcams, a PTZ and an infrared cameras, fingerprint reader and infrared badge system that are sensing an office floor for Accenture Technology Labs and a meeting room that is located on another floor. 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 and two discussion areas. 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. Another computer collects events’ pictures, converts them into MPEG-1 movie and creates an event record in an SQL database. Another database contains events detected by the infrared badge system. The event databases serves as a common repository for both people who are doing manual search of events and automatic analysis.

The objectives of the MSIS project are to create a realistic multi-sensor indoor surveillance environment that serves a base for developing more advanced event analysis algorithms, such as people recognition and tracking, using collaborating agents and domain knowledge.

The following analyses and prototypes have been developed or are currently under development.

- Search and browsing of the Event Repository database using a Web browser.
- Creating a people localization system that is based on evidence from multiple sensors and domain knowledge.
- Creating a real-time people tracking system that is based on multiple sensors and prediction of person’s behavior.
- Creating an event classification and clustering system.

4. Preliminary Experimental Results

For our pilot experiment we used eight video cameras that are integrated into four clusters. Four people served as volunteers to be localized in the experiment. Every day up to fifty people are working on the floor. The color features of most of them have been used for building the “unknown person” model. To evaluate the accuracy, we recorded four days of video data, with eight hours per day. Data from one of the days was used for prior probability estimations, and the rest for testing. We used precision (P) and recall (R) as measures for evaluating the performance of the system:

$$P = C / T \text{ and } R = C / A,$$

where C is the number of events where people were correctly localized by the system, A is the number of events where people are actually visible (ground truth), and T is the number of events that the system claimed that a person was located in that location. Using only visual features, the system obtained average recall of 68.23% and precision of 59.16%. Using the domain knowledge and semantics of the sensors locations, the performance increased to average recall and precision of 87.21% and 73.55%, respectively.

The experiments that use more sensor streams and integrate video data with infrared badge system are planned.

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 homeland security 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 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] Kettner, 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.