Multi-Camera People Tracking Using Bayesian Networks

Meng Howe Tan and Surendra Ranganath

Department of Electrical and Computer Engineering
National University of Singapore

Abstract

A multi-camera tracking system is considered for video surveillance, where the cameras have non-overlapping views and the system is required to be robust under different lighting conditions. As people enter the scene, they are segmented out as foreground objects. Facial features, texture and color features of clothing, and likely location of people are extracted and matched with like data stored in a database that is dynamically constructed. The top 3 matches are used in a Bayesian Network, together with a face detection confidence measure, and the likelihood of a person’s location for inferencing. To establish identity, results over 16 consecutive frames are used in a majority voting scheme. When the person is identified, the extracted data is stored in the database for future use. To assess the performance of the system, experiments were conducted on a database of 11 people, simulating 4 different camera views. Identification accuracies in excess of 95.45% were obtained in different experiments.

1. Introduction

Most of the work on tracking in video surveillance systems has concentrated on single camera systems, and addressed problems such as dealing with occlusion and clutter, using techniques such as Kalman filter [1] and CONDENSATION [2], with a variety of cues such as colour, motion, shape and stereo information.

There are much fewer works on tracking objects through multiple cameras, which is required in realistic environments to cover large areas. To track (match) objects as they move under different camera views methods as in [3] and [4] require explicit camera calibration and geometric features like the person’s height for cross-camera correspondence. Another approach for solving the problem is to match objects using color cues [5], without the need for overlapping views or calibration. [6] combines color and height of subject for tracking between cameras. However, this method still requires a self-calibration method. Except for [5], these methods require overlapping camera views which entail high cost of deployment if the system is to be used in a large environment. Other approaches use Bayesian tracking methods for traffic surveillance [7]. [8] uses contextual constraints for counting people in multi-camera systems.

In this paper, we consider a system for identifying people as they appear under different cameras with non-overlapping views. The basic approach is to identify people using clothing color and clothing texture information, combined with facial information. This is augmented with their previous recognized location to improve matching. Since color information is used, the matching could be sensitive to different illumination of scenes, and different camera responses to color. We explicitly consider building robustness with respect to this. Also, we use intermediate classifiers for face and clothes matching, and integrate them naturally under a Bayesian Network inference scheme. This framework allows classifiers to be easily added to the system. Also, no retraining of the system is required as people enter and exit the surveillance environment. At present, the system is based on the following assumptions:

1. There are a limited number of entry points and exit points, which is reasonable for indoor environments. However in the implementation of the system, only entry points were used to keep track of people entering the system. Exit points are currently being implemented.
2. The foreground objects are all (upright) humans who move through the environment.
3. The people appear facing the camera. This is reasonable, for example, in building corridors. If however, bi-directional tracking is required, another camera could be installed viewing the opposite direction.

2. Feature Extraction

The scenes in the environment are classified into entrance scenes, where people enter the environment, or normal scenes. The processing of entrance scenes and normal scenes was slightly different. Though feature extraction was carried out in both types of scenes, matching and inferencing were only carried out in the normal scenes.

Figure 1 summarizes the steps in the feature extraction process. Adaptive background modeling is done in normalized rgb space to reduce sensitivity to illumination changes. The pixels are modeled using Gaussian distributions as in [9], and suitable thresholds are used to segment out the foreground objects in each frame. The raw segmentation is filtered using simple morphological operators and subjected to connected component analysis. Components above a threshold size are further processed.
Simple heuristics based on the height of the extracted foreground object are used to estimate the neckline. A rectangular region at a suitable distance below the neckline is normalized to a standard size of 60 x 90 pixels and taken to represent the clothes region. The region above the neckline is subjected to facial processing.

For face detection and feature extraction, we followed the method proposed in [10]. The first step is skin color detection as in [11]. A skin color model was constructed from 200 images with skin color regions under different types of lighting. The detected skin color region gives an approximate face size. The face detector examines the skin region to locate the face more precisely and encloses it in a circle centered on the nose. The method is inherently scale tolerant and uses a multiple orientation eigenface representation derived from log-polar mapping of face appearances for matching. It returns the detected face enclosed in a circle which is centered on the nose. A radial grid is then laid on this region, and average intensities within each cell are used to form the face feature vector. The face detector also returns the distance of the located face from face space to use as a parameter in the person identification scheme.

Color and texture features are extracted from the clothes region. To reduce sensitivity to illumination, we use an illuminant color invariant method proposed in [12] that maps a RGB image to an illuminant invariant image which depends on object reflectance only and is independent of the illuminant color and intensity. The histogram of this invariant image is obtained and used for identification of the clothing. The Blackman-Tukey power spectrum of the cloth region is used to represent the texture of the clothing. The power spectrum is sampled over a radial grid structure, and the average power in each radial grid cell is used to form the texture feature vector.

3. Intermediate Matching

The extracted features are used in intermediate matching schemes and their results are combined in a Bayesian Network (BN) inferencing scheme. The features stored in the database are gathered from people as they move through the environment. In an entrance scene, where a person enters the environment, the extracted face feature vectors, the illuminant invariant clothing color feature vectors and the clothing texture feature vectors are extracted and entered into the database, along with a person ID. The feature vectors that are extracted from a person in a normal scene are also used to augment the database but only after the identity of the person has been established.

In the intermediate level matching, each matching process postulates the identity of the owner of these features in rank order. From each of these three matching processes, the top 3 closest matches are input to the Bayesian Network to identify the person in a normal scene.

For face feature matching, the inner product similarity measure

$$\mathop{\text{arg max}}_{v \in \mathcal{I}} \frac{\mathbf{x}_i^T \mathbf{x}_j}{\| \mathbf{x}_i \| \| \mathbf{x}_j \|}$$

(1)

is used to find the closest match between a face feature vector and like feature vectors in the database.

For the matching of the clothing color feature, a cross-correlation measure

$$R_{xy}(n) = \sum_{k=0}^{N-1} x(k) y(k + n)$$

(2)

was found to give best results. Here $x(k)$ and $y(k)$ are normalized histogram bin values, representing the intensities of the illuminant invariant images. $N$ is the number of bins in the histograms and $n$ is the lag. Using (2), $R_{xy}(n)$ for different lag values is obtained and $\max R_{xy}(n)$ is taken as the similarity measure. Experiments showed that the differences in CCD sensor sensitivities of different cameras did not play a role in the illuminant invariant matching method.

The clothes texture matching uses the inner product similarity measure in (1).

4. Identification Using Bayesian Nets

In our system, Bayesian Networks (BNs) [13] are used to identify people as they move through the various camera views. We designed an open system where the number of people in the environment can change over time. Rather than using one node in a single BN to represent all the people in the environment, each person is represented by his own BN where the Person ID node only contains two states, TRUE and FALSE. All the BNs have the same structure and parameters. This eliminates the need to retrain Bayesian Networks whenever a new person enters or exits the environment.

Figure 2 shows the common BN structure used. All nodes in the network are binary valued. A total of 5 different features are used in the BNs. The face feature vector, illumination invariant colour feature vector and texture feature vector (Section 3) are matched with the features stored in the database to obtain the top 3 matches for each of these features. The observed values of these features are denoted as F1-F3, I1-I3 and T1-T3, respectively. Each of these observed values represents whether the match obtained represents the person, and can only be TRUE or FALSE. These match values are combined with 2 other pieces of information, namely, the eigen-distance of the
face feature vector from face space, and the likelihood of the person being in a particular scene based on his previous location.

The eigen-distance gives a measure of the reliability of the face detection. In order to use this feature, the distribution of this distance for the correctly detected faces and falsely detected faces was studied. From this distribution, the eigen-distance values were divided into two groups, based on a threshold, which separated eigen-distance values into a range where the feature vector was likely to be a face and a range where the face detector might detect a face incorrectly.

The likelihood of a person being in the present location is also used as one of the features. By making use of the knowledge of the camera locations, it is possible to know whether the person is likely to be at present location based on his past location. Hence, the Location node of our BN consists of two states: LIKELY and UNLIKELY.

The conditional probability tables of the different nodes in the BN are found by training with data obtained from people moving in all the scenes in the environment. When a person appears in the entrance scene, his features are extracted and stored in the database. When he moves to other scenes, the features are again extracted and matched to the features in the database. This same process takes place as he moves through the other scenes. As more people enter the scene, the same capturing of data and matching with the features in the databases continues. Once all the matching results were available, the BN is trained by using Expectation Maximization (EM) algorithm in the Bayes Net Toolkit [14].

After training, one BN is instantiated for every new person who enters the environment. The BN with the highest probability for the TRUE state of the person ID node specifies the identity of the person. In the system implemented, this inferencing is done independently over 16 frames (or less depending on data availability) and majority voting is used over these 16 frames to finally specify the identity of the person.

After the ID of a person has been established as above, the person’s features are stored and used for further matching. A tracker is required to track a person within a scene. Our tracker uses very simple heuristic rules to track the person within a scene and does not take into consideration cluster formation and occlusions. We do not attempt to solve these problems here as the main aim is to investigate inter-camera tracking of people rather than tracking within a single scene.

5. Results and Analysis

We captured 11 subjects in 4 different scenes (which we name A, B, C and D) with different lighting. A background image of each scene is shown in Figure 3. Scene A was used as the entrance scene to the environment, while Scenes B, C and D were used as normal scenes. The lighting in these scenes is a typical mixture of what may be found in buildings. The data was captured using a Panasonic 3 CCD digital video camera at a resolution of 576x720 and a frame rate of 25 frames/sec. The video was transferred to a computer via firewire interface for processing. Figure 4 shows a subject in the entrance scene.

A virtual environment was simulated to evaluate the system, as we did not have all the cameras and other necessary resources for actual implementation and testing. The layout of the virtual setup is shown in Figure 5. This virtual layout is different from the real location of the scenes where the data was obtained. The letters A, B, C and D in the layout represent the camera location in the scenes in our test dataset and the arrows show the imagined flow of the people through the environment. Locations E and F are unmonitored by cameras.

The system was tested in two ways. In the first experiment, all the 11 subjects were used for training and testing. In the second experiment, 8 subjects were used for training and the other 3 were used for testing.

Experiment 1

When a person enters the scene, the training data was obtained from the 3rd frame (after the foreground region exceeded the threshold size) and every subsequent 5th frame after that (8th, 13th, 18th,….) for a total of 16 frames. For training the BN, features extracted from the 11 people in various scenes were matched with like features from other scenes to obtain the matching results. The location likelihood information was obtained through the simulated movement of people through the 4 scenes in the environment. Table 1 shows the simulated location of people in the virtual environment at each time instant. The time taken for different people to travel between scenes can be different, for example, by moving to unmonitored locations E and F in Figure 5 before going to the next scene. For testing the 1st frame and every 5th frame after that (6th, 11th, 16th,….) for a total of 16 frames was used.

The raw matching accuracies from the intermediate classifiers are shown in Table 2. When these classifiers are used in the BN scheme, the results are much improved. On a frame-by-frame basis, the system was able to identify the people in the scene at an accuracy of 95.74% (337 out of 352 matches). To obtain the identification of people based on a scene by scene basis, the intra-camera view tracker was used to track a person over 16 frames and majority voting was carried out on the 16 sets of inference results. The system was able to identify people correctly 21 times out of the 22 times that person identification was carried out (95.45%).
Experiment II
This experiment was similar to Experiment I except for the training and testing data. The testing data was obtained from the first 3 subjects from the dataset and the training data was obtained from the other 8 subjects. On a frame-by-frame basis, the system was able to identify the people in the scene at an accuracy of 100% (96 out of 96 matches). The identification rate of the people based on a frame by frame basis was also 100%. A possible reason for the good result is that limited testing data was used here.

6. Conclusion
We have proposed a system for tracking people with multiple cameras using Bayesian Networks. The BNs use intermediate matching results from facial feature, illuminant invariant clothing colour feature and clothing texture feature along with the eigen-distance and the likelihood of the person being at that location. The BN with the highest posterior probability for the TRUE state of the Person node was used to identify the person. Two experiments that were conducted to assess the overall system performance yielded accuracy rates in excess of 95.45% for 11 people and 4 scenes in the database.

An advantage of the proposed system is that is possible to easily integrate matching results from various sensors. Also, inferencing is possible even when some of the nodes in the Bayesian network are not instantiated. For example, if face matching results are unavailable because the face detector failed, the system can still identify the person based on other data. The modular design of the system allows other features to be added easily. Unlike most BN systems that are used for classification, there is no need to retrain our BN as more people enter the environment. A limitation of the system is that it currently does not work in real time. This is mainly due to the computations involved in the face detector; however, by using other efficient face detectors, the system can be speeded up considerably. Also, though this system will be useful along corridors in buildings or in other situations where the person is walking towards the camera, this assumption needs to relaxed for more general applications.

References
Figure 1: The feature extraction process.

Figure 2: Bayesian Network structure used.

Figure 3: Scene A  Scene B  Scene C  Scene D
Figure 4: Example of a person in the entrance scene.

Figure 5: Layout of the virtual environment.

<table>
<thead>
<tr>
<th>Person</th>
<th>Time (units)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A C D</td>
</tr>
<tr>
<td>2</td>
<td>A B D</td>
</tr>
<tr>
<td>3</td>
<td>A C D</td>
</tr>
<tr>
<td>4</td>
<td>A B D</td>
</tr>
<tr>
<td>5</td>
<td>A B D</td>
</tr>
<tr>
<td>6</td>
<td>A C D</td>
</tr>
<tr>
<td>7</td>
<td>A C D</td>
</tr>
<tr>
<td>8</td>
<td>A B D</td>
</tr>
<tr>
<td>9</td>
<td>A C D</td>
</tr>
<tr>
<td>10</td>
<td>A B D</td>
</tr>
<tr>
<td>11</td>
<td>A C D</td>
</tr>
</tbody>
</table>

Table 1: Simulated path of people in Experiment I.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Correct Match Placement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st</td>
</tr>
<tr>
<td>Clothing Texture</td>
<td>84.94%</td>
</tr>
<tr>
<td>Clothing Invariance Color</td>
<td>76.99%</td>
</tr>
<tr>
<td>Face</td>
<td>57.51%</td>
</tr>
</tbody>
</table>

Table 2: Intermediate matching result accuracies.