A MULTI-CAMERA VISUAL SURVEILLANCE SYSTEM FOR TRACKING OF REOCCURRENCES OF PEOPLE

Thang V. Pham∗ Marcel Worring Arnold W.M. Smeulders

ISLA, Informatics Institute
University of Amsterdam
Kruislaan 403, 1098 SJ, the Netherlands
{vietp, worring, smeulders@science.uva.nl}

ABSTRACT

This paper describes a software system to track the reoccurrences of objects in multi-camera visual surveillance. Specifically, it is designed for after-event tracking of people to aid in a typical investigation of events occurring in a certain location at a certain time. This is a nontrivial problem because of several aspects that influence the appearance of scenes and people, such as changes in viewpoints, lighting conditions, shadow, occlusion, and weather conditions. Another challenge, which is the focus of this paper, is to integrate different required components into a complete working system, namely (i) motion detection, (ii) object classification, (iii) object modeling and matching, and (iv) interactive retrieval and visualization. We have designed and implemented a robust system consisting of state-of-the-art technologies in each component. We performed experiments with the system on a real-life dataset gathered from 12 street surveillance cameras over two hours in a city area. The experiments showed promising results in retrieving the reoccurrences of four target subjects.

Index Terms— Object tracking, object detection, object classification, visual surveillance.

1. INTRODUCTION

The amount of surveillance videos is increasing. Finding useful information from a large collection of surveillance videos is becoming unmanageable by human. This paper aims at building an object retrieval system for such a large collection.

We concentrate on the problem of finding re-occurrences of human in videos. The principle application is to assist the user - e.g. police - to quickly identify other occurrences of the subject of interest. We consider the after-event scenarios. For example, the police are informed of certain incidence. They go back to the videos in the area of interest around certain time, and analyze the data.

∗The author is now with the OncoProteomics Laboratory, VU University Medical Center, the Netherlands.
Fig. 1. The operational scheme of the proposed system.

2. SYSTEM

Figure 1 depicts the architecture of the proposed system. There are four main modules: motion detection (in 2.1), person classification (in 2.2), person matching (in 2.3), and user interface (in 2.4). A set of videos is to be processed to extract all appearances of people therein. These appearances are stored in the object repository for interactive browsing and retrieval. The module for person classification requires an external set of training examples of people.

2.1. Motion detection

Amongst many existing methods [8, 9, 10, 11], we chose a simple adaptive technique that models the background pixel by pixel independently.

Let \( I^t_x \) denote the three-dimensional RGB color vector of pixel at image location \( x \) in frame \( t \), \( t \geq 0 \). For \( t > 0 \), a background model \( B^t_x \) is updated

\[
B^t_x = \alpha B_x^{(t-1)} + (1 - \alpha) I^t_x
\]

with \( B^0_x = I^0_x \), and \( \alpha \) control the adapting rate. At time \( t \), a pixel location \( x \) is classified to be background if there is a small difference between the current image and the main-

\[
||I^t_x - B^t_x||_2^2 < \beta \tag{2}
\]

where \( \beta \) is a threshold.

In our system, \( \alpha = 0.15 \) and \( \beta = 0.05 \). These values were set by visual examination.

The result of this background subtraction is a binary image of moving object. Image rectangles containing more than 20% of the foreground pixels are considered an object candidate, which will be further tested by the object classification module.

We note that no other preprocessing such as temporal or spatial smoothing is performed.

2.2. Object classification

This section describes the histogram of oriented gradients in [3] for human detection. It is derived from the SIFT descriptors which has been shown to perform well for image matching [2].

The image (frame) is divided into a grid of overlapping cells and blocks. The advantage of using features derived from an image block is that one can perform local normalization, which is essential for good performance. In this paper, each 32x64 window is divided into 3x7 overlapping blocks of 16x16. Features are derived from each block independently and subsequently concatenated to create a feature vector.

Each block is further divided into 2 \( \times \) 2 cells; each cell is of 8 \( \times \) 8 pixels (see Figure 2).
Each descriptor is computed as follows. First, the gradient vector is computed at each pixel. For color images, the gradients are computed separately for each channel, and the one with maximal norm is selected as the pixel’s gradient vector. The gradients are computed with a simple $[-1, 0, 1]$ kernel.

In the next step, the gradient vectors within each $8 \times 8$ cell are accumulated in a histogram of orientation. Here, we used nine orientations, resulting in a 9-vector $u$ for each cell.

Finally, the descriptor $v$ at each location is formed by concatenating the vectors in each $2 \times 2$ block, $v = [u_1, u_2, u_3, u_4]$ and normalized

$$v = \frac{v}{||v||} \quad (3)$$

In short, each block is a vector with 36 components. Thus, in total each object is represented by a vector of $3 \times 7 \times 36$ components. The reader is referred to [3] for a discussion on the various parameter values used here.

These descriptors are used in support vector machine learning [12] to construct a human classifier.

### 2.3. Object matching

There is some novelty in our approach to object matching. It can be seen that the image rectangles containing human appearances are dominated by the background. As a result, object matching based on image information in the whole complete rectangle would fail. Instead of solving the difficult problem of segmenting the articulated object from the background [13], we exploit the information from motion detection.

Figure 3(a) shows examples of the motion detection result. As can be seen, there are noises and shadow around the human appearance in general. Thus, we expect that by averaging over these motion detection results, we would be able to focus on the most likely region containing object appearance. Figure 3(a) is the average. By keeping only about 30% of the most likely pixels that will appear in the object appearance, we effectively create an object mask.

Subsequently, we use RGB color histogram of the pixels that are not masked out for object matching.

### 2.4. Browsing and visualization system

An important component of the system is the user interface, which facilitates interactive browsing, querying and visualization of the video collection and search results. Figure 4 shows a screenshot of the user interface.

On the left is a general control panel where the user can switch between camera views, storing search results, and loading existing search results.

The main window is composed of five parts A, B, C, D, and E. Part A is the camera view with overlaying of human detection result (thus, it may be used to evaluate the human detection component in 2.2 visually).

Part B is the camera topology view. When used together with time information, this spatial information is helpful in establishing object trajectory over multiple cameras.

Part C contains the list of all human detected in 2.2 sorted roughly by the order of appearance. Thus, it can be used as an object-based browsing tool for the video player in part A. It also has an object selection tool for similarity searching.

Part D is the list of objects that are similar to the object selected in C (see 2.3). It also has the object selection tool for similarity searching and storing of results.

Finally, E is the list of stored results. One can add to this list using the object selection tool in C or D, and remove objects from this list.

In short, we have built a compact user interface that has essential features for the interactive object-based browsing and searching of the video collection.

### 3. EXPERIMENTS

#### 3.1. Setup

With the assistance of the police, we were able to acquire real-life surveillance videos in a designated part of a city. There were four subjects walking around the area for almost two hours. To increase the variation in the dataset, the subjects intentionally changed clothes with different colors and textures at times. They also entered and left shops and buildings a number of times. Each subject is equipped with a global positioning device to aid in the construction of the ground-truth (not yet implemented in the current paper). The trajectories of the subjects can be seen in figure 5.

There are 12 cameras; and we gathered the video data over approximately two hours (thus, there are 24 hours of surveillance video to be processed in total). The dataset contains several sources of variation. Most notably are the presence

---

**Fig. 3.** (a) Examples of motion detection result and background modeling. (b) Average of all detection results (including false alarms). An object mask is created by keeping 30% of the most frequently occurring pixels.
of a large number of pedestrians participating in the traffic, weather (the presence of wind and cloud changes the appearance of background drastically in a short time), occlusion due to traffic and building, changes in the colors and texture of the clothes of the subjects, the presence and disappearance of parked cars (thus, the assumption of static scenes is violated). In addition, the area where we collected data was under active surveillance. This means that in some cases the cameras are manually operated by police controllers. Typical actions are zooming in/out, panning and tilting. As a consequence, the camera views change quickly during these manipulations. In short, while there are certainly sources of variations that are not captured (such as rain and nighttime vision with street lighting), the dataset contains many challenges for analysis of surveillance videos.

3.2. Experiments and results

Due to the lack of ground-truth, most of our evaluation is carried out by visual examination. We will discuss the results of motion detection, object classification, and object matching.

The motion detection module works reasonably well. We rarely observed misdetection of human presence due to motion detection. Note that we did not perform any postprocessing such as shadow removal because in our design the classification step would eliminate the effects. Also, because of the adaptive nature of the algorithm, it can handle scene changes such as the parking and departing of cars. On the other hand, it has a negative effect when pedestrians stop for a long time (several minutes), they will gradually be merged into the background. This, however, does not create a serious problem because typically they are detected when entering and leaving the static positions.

For the human classification module, we performed experiments with the detection size. The training examples provided in [3] are of size $64 \times 128$, which are quite large for our application. One approach to detect humans in lower resolution is to blow up the image frame. However, this approach comes with a much larger computational cost. We opted for another approach, which resizes the training examples and subsequently builds object classifier based on these smaller size examples. Figure 6 shows the ROC curves for three different sizes. One can observe that a size of $16 \times 32$ would degrade the accuracy considerably. Our system employs a detection size of $32 \times 64$. But it should be noted that it is desirable to be able to detect objects at a lower resolution (in
For the object matching module, our experiment showed that without object masking, the quality of matching is very poor due to the domination of background.

There are two observations regarding the contribution of object matching. The first is that it helps eliminate a large number of false alarms. The consequence is that we can raise the number of true detection in the human classification phase while accepting more false alarms. Finding a good balance, however, is not trivial. The second observation is that interaction plays a very important role, especially when simple object descriptor does not cover the complete object appearance. In the screenshot in figure 4 part E (the list of result), one can see that the appearance of the front and the back are very different. It is impossible to get all such appearance in one search. Nevertheless, thanks to interaction, they still can be retrieved.

Finally, for the general task of searching for reoccurrences of human, we performed the search for each person in approximately 20 minutes. We were able to retrieve appearance of the four subjects during the two hours. Nevertheless, qualitative results are not yet available because of the lack of a ground-truth.

4. DISCUSSIONS

We have presented a system for searching reoccurrences of human appearance. We value the completeness of the system from preprocessing to interactive searching and visualization. We were able to retrieve the appearance of four subjects in 12 surveillance videos over two hours.

We consider this as a stepping stone for further improvements. For motion detection, temporal and spatial filtering certainly helps in reducing the processing load down stream and increasing accuracy.

For object classification, one problem is that of training examples. Currently, we use the dataset provided by [3]. This dataset, however, is not quite representative for the data we have. The examples are mostly captured on the same image plane, not from high above as in surveillance videos. We expect that methods that gather the dataset online automatically provide a promising direction [4]. Furthermore, temporal filtering as object tracking [14] might be useful in eliminating false positives.

There are two ways to improve object matching. Firstly, one might employ representation invariant to certain known imaging conditions. Secondly, a more ambitious goal is a part-based representation covering the whole aspect graph of object appearance.

Acknowledgments

This research is supported by MultimediaN. We would like to thank the Netherlands Forensic Institute (NFI) and especially Jurrien Bijhold for their contributions to the project.

5. REFERENCES


