Detecting Automobiles and People for Semantic Video Retrieval

Rene Visser   Nicu Sebe   Michael S. Lew

LIACS Media Lab
Leiden University, 2300 CA Leiden
The Netherlands

Abstract

This paper describes a method for detecting automobiles and people in streaming or archived video. Our video object tracking system is based on Kalman filter updating of an active contour over the video sequence. We use the sequential probability ratio test (SPRT) to classify the moving objects. Results are shown of a real video sequence from a busy city intersection.

1 Introduction

In several visual information retrieval systems, face detection is used toward allowing queries on a semantic level. In our work, we describe a method for detecting automobiles and people in complex real scenes. The detection of people is also different than face detection in that the face detectors generally require face regions of at least 16x16 pixels. In our test videos, the face region is often 5x5 pixels or smaller; or the face may be turned away from the camera. Thus, we detect the entire body of the person, not the face.

Detecting particular objects in video is an important step toward semantic understanding of visual imagery. For example, in content based retrieval, the ability to detect people and automobiles gives the option of advanced queries such as "Find a video clip which contains a crowded area or a fast moving car."

In this paper, we give a summary of the current work in video segmentation for moving objects followed by the implemented tracking algorithm. The novel aspect of our work is described in Sections 4 and 5 where we use the sequential probability ratio test for classifying the moving objects and show some results. Conclusions are given Section 6.

2 Background

The segmentation of moving objects is an important problem in image sequence analysis and in the problem of video retrieval (i.e. [2]). Several systems have approached this problem such as [1,3-14]. A real-time system for tracking people, called Pfinder ("Person finder"), is proposed in [8]. First, a model of the scene is built by observing the scene when no person is present. For each pixel, the mean color value and the covariance of the associated distribution is determined. Then when a person enters the scene, the system begins to build up a model of that person. This is done by first detecting a large change in the scene, and then building up a multi-blob model of the person over time.

There are other systems which attempt to detect semantic categories in visual imagery such as ImageScape [3] which used information theory for detecting semantic concepts such as sky, stone, water, people, etc. This work is different in that it uses the SPRT for the concept detection.

3 Video Object Tracking

In order to find the moving objects we maintained a weighted average of the scene to be used as the

1051-4651/02 $17.00 (c) 2002 IEEE

background reference frame. Subtracting the current frame from the background reference frame results in blobs of moving objects. We described the blobs using an eigenvector/value decomposition of the region within the blob. A blob is thus represented by a vector of n values and n parameters; and the (x,y) origin parameters.

We applied an approach similar to [1] which used Kalman filtering (Appendix) to maintain the object identity over the video sequence and to optimize the tracking process of each blob. Figure 1 displays an example of blob segmentation where the left image is the original frame and the right image contains the segmented person.

The tracking method is important in that it is the first stage of segmentation. Any errors which appear in the tracking stage will also propogate to the classification stage described next.

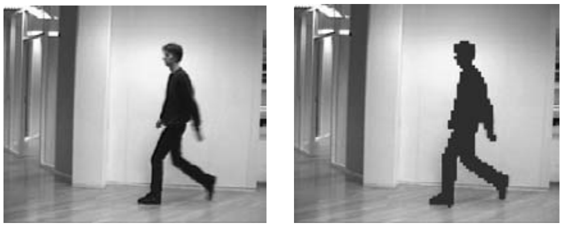

Figure 1. Frame from sequence (left); Segmented blob using background subtraction and Kalman Filtering (right).

4 Sequential Probability Ratio Test

We turned to the statistical literature for the sequential probability ratio (SPRT) test, which we use to classify the blobs as people, automobiles, or unknown. SPRT is a statistical test which uses a variable number of measurements and demonstrates how the number of measurements can be traded off for lower error probabilities.

Let $\boldsymbol{y_m}$ be a vector of m measurements $y_1, y_2, \ldots y_m$ on an object to be classified as either w_1 or w_2. SPRT is defined such that we classify $\boldsymbol{y_m}$ according to

$$L(\boldsymbol{y_m}) = p(\boldsymbol{y_m}|w_1)/p(\boldsymbol{y_m}|w_2) \tag{4.1}$$

$$\text{If } L(\boldsymbol{y_m}) > A \text{ then choose } w_1 \tag{4.2}$$

$$\text{If } L(\boldsymbol{y_m}) < B \text{ then choose } w_2 \tag{4.3}$$

If $L(\boldsymbol{y_m})$ falls between A and B then take another measurement.

where A and B are thresholds with A>B.

It is well known that SPRT is optimal in the sense that it minimizes the number of observations necessary to achieve error probabilities e_1 and e_2 for the classes, w_1 and w_2.

In the case where the measurements are independent, the test can be shown to be

$$L(\boldsymbol{y_m}) = \prod_k p(\boldsymbol{y_k}|w_1)/p(\boldsymbol{y_k}|w_2) \tag{4.4}$$

where k varies from 1 to m.

Furthermore, we can relate the thresholds A and B to the error probabilities, e_1 and e_2. After derivation the result is

$$A \leq (1-e_1)/e_2 \tag{4.5}$$

$$B \geq e_1/(1-e_2) \tag{4.6}$$

which provides a direct way of selecting the thresholds to achieve desired error probabilities.

In our system, the vector of n eigenvectors for a blob is considered to be a single measurement. As the blob is tracked over the video sequence, it gives us a new measurement per frame of the video sequence. Specifically, y_1 corresponds to the eigenvectors of the blob in frame 1; y_2 corresponds to the eigenvectors of the blob in frame 2, etc. $p(\boldsymbol{y_m}|w_i)$ was estimated from training sets using the distance in eigenfeature space.

In the SPRT process, we assigned w_1 to the class of *people* and w_2 to the class of *automobiles*. If $L(\boldsymbol{y_m})$ fell between A and B after all the possible frames were used, then the blob was labeled as class *unknown.*

In summary, our system works as follows

(1) Video segmentation and tracking similar to [1].
(2) Blob region representation using eigenvector decomposition [15].
(3) Blob region classification into 3 classes: people, automobiles, or unknown using SPRT.

5 Results

It is notable that our system uses off-the-shelf components which can be found in a typical computer/electronics store. We expect our system to be challenged by the following sources of noise:

- low resolution/detail images
- color and lens distortion
- loss of brightness and contrast from the video capture process.
- artifacts from the video tracking/segmentation process.
- block compression artifacts inherent in MPEG-1

For our tests, we used 6 video sequences of city street intersections. Each sequence was 5 minutes (at 25 fps) in length and captured using a PAL camcorder. The video was extracted to 1.5 Mbps MPEG-1 digital format using an ATI All-In-Wonder Rage 128 card. The frame resolution was half PAL resolution.

Moreover, the system is expected to function in situations where the size of the moving object is quite small (less than 20 pixels wide).

On a PIII-800 Mhz computer, our system was able to capture, track, segment, and classify all of the blobs at a rate of 17 fps.

In Figures 2 through 5, we display several examples of the results on a city street intersection test video clip. In each figure, the image on the left shows the current estimated background image and the right image displays where it found moving people (white box - green on a color display) and automobiles (black box - red on a color display). An example of a mismatch is shown in Figure 6 where two people were segmented as a single blob by the tracking/segmentation algorithm and were misclassified as an automobile by the SPRT algorithm.

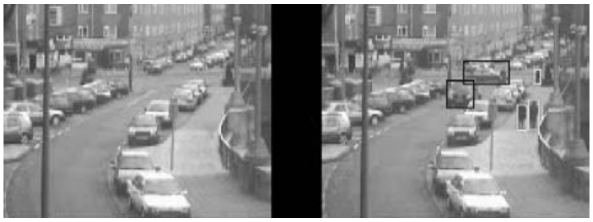

Figure 2. Example of detecting automobiles (black rect.) and people (white rect.).

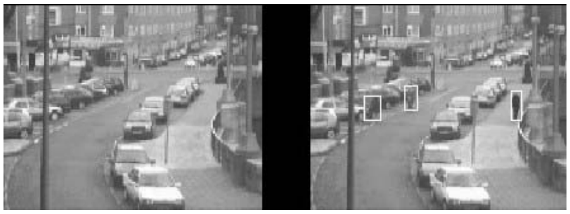

Figure 3. Example of detecting people

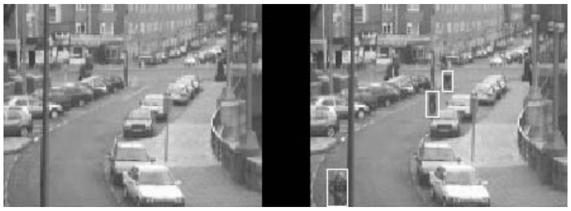

Figure 4. Example of detecting people

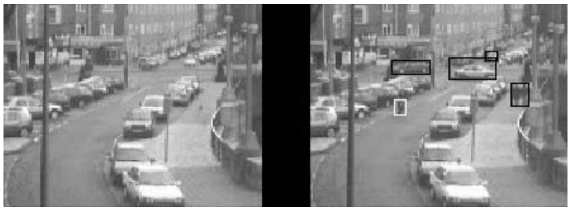

Figure 5. Example of a 2 people mistaken for an automobile.

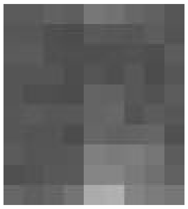

Figure 6. Example of a mismatch. This image contains two people walking next to each other. Our system classified it as an automobile. Even for a human, it is challenging to classify due to the low resolution of the image.

6 Discussion & Conclusions

In the previous sections we have described a system for detecting automobiles and people from video sequences. The novel aspect of this paper is in applying the sequential probability ratio test to the problem of classifying the blobs. The SPRT is naturally suited for this task because it was designed with sequential measurements/classification in mind.

In particular, each frame in the video containing the blob adds additional evidence toward the classification process. As more frames are processed, it can be shown that the error probabilities decrease.

From the tests, our system had misdetection rates of 0.09 and 0.07 for people and automobiles, respectively. Considering that the size of the blobs was very small (often near 15 pixels in width), the classification system was remarkable accurate.

The most common misclassification occurred when groups of people walked close by each other. This caused the tracking system to segment the group as a single object. Addressing this problem would require improving the tracking/segmentation stage or trying a different tracking algorithm.

Regarding the application of content based retrieval, we think the system has promising results and the people/automobile classifications could be used to augment a video retrieval system.

Future work is planned toward integrating the object tracking/classification system into a video retrieval system for the WWW.

References

[1] A.M. Baumberg, “Learning Deformable Models for Tracking Human Motion”, PhD thesis, The University of Leeds, School of Computer Studies, UK, October 1995.

[2] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, D. Zhong, “VideoQ: An Automated Content Based Video Search System Using Visual Cues”, *ACM Multimedia ‘97 Proceedings*, pp. 313-324, (Seattle, Washington, USA), November, 9-13, 1997.

[3] M. Lew, “Next Generation Web Searches for Visual Content," *IEEE Computer*, November, pp. 46-53, 2000.

[4] M. Irani, P. Anandan, “A Unified Approach to Moving Object Detection in 2D and 3D Scenes”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 6, pp. 577-589, June 1998.

[5] R. J. Qian, M. I. Sezan, K. E. Matthews, “Face Tracking Using Robust Statistical Estimation," *Proc. IEEE International Conference on Image Processing*, , vol. 1, pp. 131-135, 1998.

[6] H. S. Sawhney, S. Ayer, “Compact Representations of Videos Through Dominant and Multiple Motion Estimation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 8, pp. 814-830, August 1996.

[7] C. Stauffer, W. E. L. Grimson, “Adaptive background mixture models for real-time tracking”, *Proc. Computer Vision and Pattern Recognition*, (Fort Collins, Colorado), June 23-25, 1999.

[8] C. R. Wren, A. Azarbayejani, T. Darrell, A. P. Pentland, “Pfinder: Real-Time Tracking of the Human Body”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 780-785, July 1997.

[9] S. A. Niyogi, E. H. Adelson, “Analyzing and Recognizing Walking Figures in XYT”, *Proc. Computer Vision and Pattern Recognition*, pp. 469-474, (Seattle, WA), June 21-23, 1994.

[10] S. M. Smith, J. M. Brady, “ASSET-2: Real-Time Motion Segmentation and Shape Tracking”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 17, No. 8, pp. 814-820, August 1995.

[11] M. Betke and N. Makris, "Fast Object Recognition in Noisy Images Using Simulated Annealing." *Proceedings of the Fifth International Conference on Computer Vision*, pp. 523-530, June 1995.

[12] M. Betke, E. Haritaoglu and L. S. Davis, "Real-Time Multiple Vehicle Detection and Tracking from a Moving Vehicle." *Machine Vision and Applications*, July 2000.

[13] I. Haritaoglu, D. Harwood and L. Davis, "W4: Real-Time Surveillance of Peole and Their Activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 809-830, August 2000.

[14] Y. Ivanov and A. Bobick, "Recognition of Visual Activities and Interactions by Stochastic Parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 852-872, August 2000.

[15] A. Pentland, B. Moghaddam and T. Starner, "View-Based and Modular Eigenspaces for Face Recognition," IEEE Conf. on Computer Vision and Pattern Recognition, 1994.

Appendix

Kalman Filtering (see [1])

The Kalman equations used for the tracking were:

Time update equations:

$$\hat{x}_{k+1}(-) = A_k \hat{x}_k(+) \qquad \text{(A.1)}$$

$$P_{k+1}(-) = A_k P_k(+) A_k^T + Q_k \qquad \text{(A.2)}$$

Measurement update equations:

$$P_k^{-1}(+) = P_k^{-1}(-) + H_k^T R_k^{-1} H_k \qquad \text{(A.3)}$$

$$K_k = P_k(+) H_k^T R_k^{-1} \qquad \text{(A.4)}$$

$$\hat{x}_k(+) = \hat{x}_k(-) + K_k\left(z_k - H_k \hat{x}_k(-)\right) \qquad \text{(A.5)}$$

1051-4651/02 $17.00 (c) 2002 IEEE