TWO VIDEO ANALYSIS APPLICATIONS USING FOREGROUND/BACKGROUND SEGMENTATION

[1]Z. Zivkovic, [1]M. Petkovic, [1]R. van Mierlo, [1]M. van Keulen, [1]F. van der Heijden, [1]W. Jonker, [2]E. Rijnierse

[1]University of Twente, The Netherlands, [2]AVV (Dutch Ministry of Transport), The Netherlands

INTRODUCTION

A huge amount of video material is produced daily: television, movies, surveillance cameras etc. As the amount of the available video content grows, higher demands are placed on video analysis and video content management. A general review of the image based content indexing is given in Smeulders et al (1). The video indexing is reviewed for example in Brunelli et al (2).

Probably the most frequently solved problem when videos are analyzed is segmenting a foreground object from its background in an image. After some regions in an image are detected as the foreground objects, some features are extracted that describe the segmented regions. These features together with the domain knowledge are often enough to extract the needed high-level semantics from the video material. In this paper we present two automatic systems for video analysis and indexing. In both systems the segmentation of the foreground objects is the basic processing step. The extracted features are then used to solve the problem. The first system (described in the next section) is a traffic video analysis system. The foreground objects that need to be detected are the vehicles on a highway. Usually, there is a huge gap ("semantic gap") between the low-level features extracted from the foreground objects and the high-level concepts. However, for this domain it was possible to manually map the extracted features to the events that need to be detected (high-level concepts) in a simple way. The second system (section 3 of this paper) analyzes videos of tennis games. It is difficult to manually generate the mapping from the features to the high-level concepts. Therefore we exploited the learning capability of Hidden Markov Models (HMMs) to extract high-level semantics from the raw video data automatically.

Although very specific, the two applications have many elements that are important for any surveillance/ monitoring system.

TRAFFIC VIDEOS

One of the many tasks of the Dutch Ministry of Transport is construction and maintenance of dual carriage highways in The Netherlands. AVV is an advisory organ that among others gathers statistical data on road usage for traffic management and flow control. Nowadays, the video camera equipment is mainly used for dynamic goals, i.e. keeping control on what is going on. The recorded videotapes are usually re-used every 24 hours. Workers in the traffic control centers do not have time and capability to analyze the video material for gathering the statistical data.

Problem definition

In the project "Secure Multimedia Retrieval" (SUMMER) attention has been given to the possibilities of automatically extracting statistical data from traffic video material. A number of events of interest for the AVV were defined. The task was to automatically detect these events.

Object detection and feature extraction

A video camera was mounted above a highway, somewhere around the middle of one side of the highway (see figure 1). One camera monitors one side of the highway. The camera points in the direction of the traffic. The difficulties with the vehicles occluding each other are reduced if the camera is placed very high. However, mounting the camera too high was not practical. Therefore, the position of the camera was a compromise. The angle of the camera with the respect to the road was also a compromise. When the camera is for example parallel to the road we can see a long part of the road but the accuracy is decreased and there is a large amount of occlusion between the vehicles. The view angle of the camera lens is another parameter that needs to be chosen. The detailed analysis of the choices we made is given in van Mierlo (4).

Since the camera was static and the background scene was not changing rapidly we used a standard adaptive background subtraction method described in (4). See figure 1d and 1e for the results.

The camera intrinsic parameters (focal length, lens distortion parameters etc.) were estimated using a calibration object and the technique from Zhang (5). Some standard dimensions of the road were known. The 3D model was fitted to the road to estimate the camera extrinsic parameters (position and orientation). See figure 1a and 1b. The 3D model is also used to segment the image into the driving lanes – figure 1c. We assumed that the road is a flat surface which is usually true for a small part of a highway.

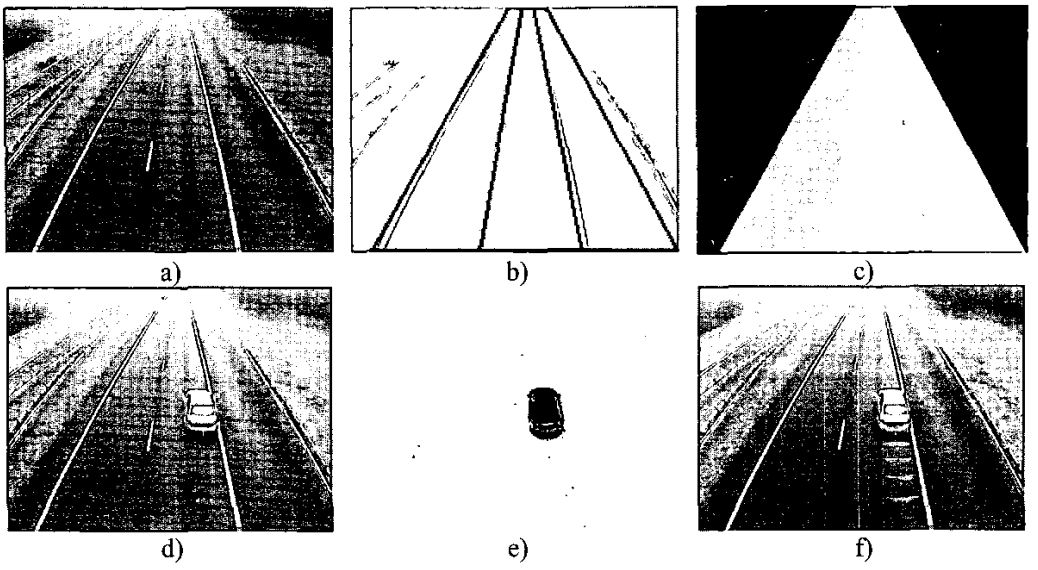

FIGURE 1 - Vehicle segmentation and feature extraction: (a) Original image; (b) Fitted 3D model; (c) Segmented road: (d) Original image and a vehicle: (e) Segmented foreground: (f) Detected and tracked bumper of the vehicle.

After the background subtraction, it is straightforward to detect the vehicles that are entering the scene. Furthermore we also estimated the speed of the vehicles. The bumpers of the vehicles were detected and tracked (Zivkovic and van der Heijden (7)) – figure 1f. We assumed that the bumper is approximately at the same height from the road surface for all the vehicles. Therefore, since the camera was calibrated, we could estimate the speed of the vehicles. For details see (4).

Results

The following events are detected: unnecessarily left lane driving, slow traffic on left lane, slow traffic on right lane, fast traffic on left lane, fast traffic on right lane, left lane passing, right lane passing, lane change, vehicle on emergency lane, truck on left lane.
The clips containing the events are automatically cut out from the raw material and compressed. The clips and the statistical data have been stored in a database and a user interface has been built to access the data and the video images in combination with administrative data AVV has stored in separate databases, such as data on accidents, traffic jams, road details, etc.
The traffic monitoring was previously analyzed a number of times, for example in Beymer et al. (6). One of our contributions is in solving the occlusion problem by carefully choosing the camera parameters and position. Further, a demonstration system is built to show the integration of different information sources from different databases with the results from the traffic video analysis system. There was a predefined set of events to detect. In the preliminary experiments the system was able to detect the events from 4 hours of video under different weather and traffic conditions. For details see: www.cs.utwente.nl/~summer.

TENNIS GAME VIDEOS

The project "Digital media warehouse system" (DMW) aims to advance scalable solutions to content-based retrieval technique in large multi-media databases (see: www.cs.utwente.nl/~dmw).

Problem definition

A case study is done and the limited domain of tennis game videos was analyzed to demonstrate the extraction and querying of high-level concepts from raw video data. The aim was to recognize different tennis strokes from the ordinary TV broadcast tennis videos.

Object detection and feature extraction

From the whole video of a tennis game we use the game shots when the camera is observing the whole field as in figure 2a. These shots can be automatically extracted from the video using a number of global image features and some heuristics.
First step is to segment the player from the background. The camera was not always static so the problem was solved in a different way. The initial segmenting is by

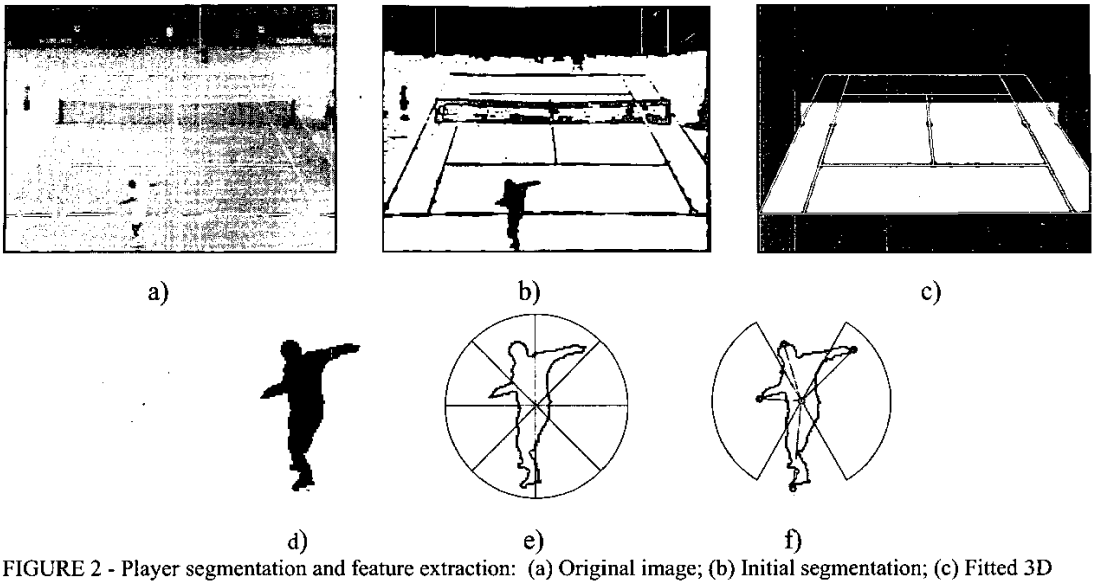

FIGURE 2 - Player segmentation and feature extraction: (a) Original image; (b) Initial segmentation; (c) Fitted 3D model; (d) Final segmentation; (e) Pie features; (f) Skeleton features

detecting the dominant color of the scene, figure 2b. The remaining lines are removed using a 3D model of the scene that was fitted to the scene automatically. Details are given in Zivkovic et al (8).

Further, we extract some specific features characterizing the shape of the segmented player's binary representation. The moving light display experiments from Johansson (10) demonstrated that people are able to recognize human activities provided by relatively little information (motion of a set of selected points on the body). Additional motivations for our approach are some recent results in recognising human activities from their binary representations Fujiyoshi and A. Lipton (11) and Rosales and Sclaroff (12). Besides the standard shape features such as orientation (f_1), and eccentricity (f_2), we extract the following features:

- The position of the upper half of the mask with respect to the mass center ($f_{3\text{-}4}$), its orientation (f_5), and the eccentricity (f_6). These features describe the upper part of the body that contains most of the information.
- A circle is centered at the mass center as shown in the Fig. 2e. For each segment of the circle, we count the number of pixels in the mask ($f_{7\text{-}14}$). This can be seen as a general approximate description.
- The sticking-out parts ($f_{15\text{-}16}$) are extracted by filtering and finding local maximums of the distance from a point on the contour to the mass center. Only certain angles are considered as indicated in Fig. 2f.

The player position was not used since it was not leading to any improvement according to the experiments.

Results

Features\Experiment	1*a*	1*b*	2
$f_{1\text{-}4}$	82.4	79.3	75.8
$f_{1\text{-}6}$	84.6	82.4	80.5
$f_{1\text{-}2,\,5\text{-}6}$	81.5	78.6	76.1
$f_{1\text{-}2,\,15\text{-}16}$	89.3	88.1	87.2
$f_{1\text{-}16}$	86.4	82.1	79.3
$f_{2\text{-}4,\,15\text{-}16}$	91.2	88.7	88.3
$f_{7\text{-}14}$	85.4	77.8	78.1
$f_{7\text{-}16}$	93.1	87.0	86.4

TABLE 1 - Recognition results (%)

Parts of video showing different tennis strokes were manually extracted. Automatic extraction is also possible using the audio, see Petkovic (3) for further elaboration. We used first order left-to-right discrete HMMs with 4 to 48 states. We used the k-means algorithm to divide the feature space into the discrete states - codebook. We tried various codebook sizes in the range of 8-80 symbols.

The first experiment we conducted had two goals: (1) determine the best feature set and (2) investigate the person independence of different feature sets. Hence, we have performed a number of experiments with different feature combinations. In order to examine how invariant they are on different players (both male and female ones), two series of experiments have been conducted: 1*a* and 1*b*. In the series 1*a*, we used the same player in the training and evaluation sets, while in 1*b* HMMs were trained with one group of players, but strokes performed by other players were evaluated. In both cases, the training set contained 120 different sequences, while the evaluation set contained 240 sequences. To be able to compare our results with Yamato at al (13), we selected the same six events to be recognized: forehand, backhand, service, smash, forehand volley and backhand volley. In each experiment, six HMMs were constructed - one for each

type of events we would like to recognize. Each stroke sequence was evaluated by all HMMs. The one with the highest probability was selected as the result (parallel evaluation). In order to find the best HMM parameters, a number of experiments with different number of states and codebook sizes were performed for each feature combination.

The recognition accuracies in Table 1 (% of correctly classified strokes using parallel evaluation) show that the combination of pie and skeleton features ($f_{7\text{-}16}$) achieved the highest percentage in the experiment 1*a*. The recognition rates dropped in experiment 1*b* as expected, but the combination of eccentricity, the mass center of the upper part, and skeleton features ($f_{2\text{-}4,\ 15\text{-}16}$) popped up as the most person independent combination, which is nearly invariant on different player constitutions. The optimal result with this combination of features was achieved with the codebook size of 24 symbols and HMMs with 8 states. Compared to (13), we achieved an improvement of 20% (experiment 1*b*) mostly due to improved, more informative, and invariant features (in the first place the novel skeleton features and then the pie features). The improvement we achieved is certainly more significant taking into account that we used TV video scenes with a very small player shape compared to the close-ups used in (13).

In the second experiment, we investigated recognition rates of different feature combinations using 11 different strokes: service, backhand slice, backhand spin, backhand spin two-handed, forehand slice, forehand spin, smash, forehand volley, forehand half-volley, backhand volley, and backhand half-volley. The training and the evaluation set remained the same as in experiment 1*b*, only the new classification was applied. Although some strokes in this new classification are very similar to each other (for example volley and half-volley or backhand slice and spin), the performance (Table 1, last column) dropped only slightly. The majority of false recognitions remained the same as in experiment 1. Nearly 65% comes from forehands recognized as backhands and vice versa, as well as from forehand-volleys recognized as forehands and vice versa.

CONCLUSIONS

In general, to be able to completely understand the video material, computers need to achieve visual competence near the level of a human being. This is still far beyond the state of the art. Nevertheless, for particular applications it is possible to design systems that create the appearance of high-level understanding. A basic video analysis step is segmentation of the foreground objects from the background. The usefulness and importance of this step is illustrated in this paper. For two domains, traffic videos and tennis game videos, we presented here two video indexing systems that automatically extract the high-level concepts from the video using the segmented images.

ACKNOWLEDGEMENTS

The SUMMER project was a national project carried out with a subsidy from the Dutch Ministry of Economic Affairs. The DMW project was sponsored by the Dutch Telematica Institute. We would like to thank all the people participating in the both projects.

REFERENCES

1. A.W.M.Smeulders, M.Worring, S.Santini, A.Gupta, and R.Jain, 2000, "Content based image retrieval at the end of the early years", IEEE Tr.PAMI, 22(12), 1349-80

2. R.Brunelli, O. Mich, C. M. Modena, 1999, "A Survey on Video Indexing", J. of Visual Communication and Image Representation, 10, 78-112

3. M. Petkovic, "Content-based Video retrieval Supported by Database Technology", PhD thesis, University of Twente.

4. R.J van Mierlo, 2002, "Video Analysis for Traffic Surveillance", MSc. Thesis, University of Twente

5. Z. Zhang, 2000, "A flexible new technique for camera calibration", IEEE Tr.PAMI, 22(11), 1330-1334

6. D.J. Beymer, P. McLauchlan, B. Coifman and J. Malik, 1997, "A real time computer vision system for measuring traffic parameters", In Proc. CVPR.

7. Z.Zivkovic, F.van der Heijden, 2002, "Better Features to Track ", In Proc. ICPR, Canada.

8. Z.Zivkovic, F.van der Heijden, M.Petkovic, W.Jonker, 2001, "Image processing and feature extraction for recognizing strokes in tennis game videos", In Proc. 7thASCI Conference, The Netherlands

9. M. Petkovic, Z. Zivkovic, W. Jonker, 2001, "Recognizing Strokes in Tennis Videos Using Hidden Markov Models", In Proc. IASTED Int. Conf. Visualization, Imaging and Image Processing, Spain

10. G. Johansson, 1973, "Visual perception of biological motion and a model for its analysis". *Perception and Psychoph.* 14(2), 210-211

11. H. Fujiyoshi and A. Lipton, 1998, "Real-time Human Motion Analysis by Image Skeletonization", In Proc. IEEE Workshop on Applic. of Comp. Vis., 15-21

12. R.Rosales and S.Sclaroff, 2000, "Inferring Body Pose without Tracking Body Parts", In Proc. CVPR

13. J. Yamato, J. Ohya, K. Ishii, 1992, "Recognizing Human Action in Time-Seuential Images using Hidden Markov Model", In Proc. CVPR, 379-385