Image Segmentation and Feature Extraction for Recognizing Strokes in Tennis Game Videos

Z.Zivkovic, F.van der Heijden

Faculty for Electrical Engineering, University of Twente,
P.O. Box 217, 7500 AE Enschede, The Netherlands,
Email {z.zivkovic, f.vanderheijden}@el.utwente.nl

M. Petkovic, W. Jonker

Computer Science Department, University of Twente,
P.O. Box 217, 7500 AE Enschede, The Netherlands,
Email {milan, jonker}@cs.utwente.nl

Keywords: video analysis, video annotation, content-based video retrieval, action recognition

ABSTRACT

This paper addresses the problem of recognizing human actions from video. Particularly, the case of recognizing events in tennis game videos is analyzed. Driven by our domain knowledge, a robust player segmentation algorithm is developed for real video data. Further, we introduce a number of novel features to be extracted for our particular application. Different feature combinations are investigated in order to find the optimal one. Finally, recognition results for different classes of tennis strokes using automatic learning capability of Hidden Markov Models (HMMs) are presented. The experimental results demonstrate that our method is close to realizing statistics of tennis games automatically using ordinary TV broadcast videos.

1. INTRODUCTION

Computer using cameras to observe and interact with their environment is no longer just a fantasy [4]. The most important part for the interaction is the recognition people activities. This problem is addressed in this paper.

In general, to be able to completely understand their environment computers need to achieve visual competence near the level of a human being. This is still far beyond the state of the art. Anyway, for particular applications researchers were able to design systems that create the appearance of high level understanding. Similarly, we constrain ourselves to tennis games videos with the aim to recognise different tennis strokes using the classification from tennis literature [11]. An important application we had in mind is content-based video retrieval [1-3]. Being able to recognise human actions computers would enable automatic annotation of large video databases.

There is huge amount of work done in the area of human action recognition, making it today one of the hottest topics in computer vision society. An effective tool for modelling time-varying patterns are HMMs. They have attracted great attention in the speech recognition research community [19], but recently HMMs have found use in gesture and human action recognition. The first publication addressing recognition of human actions using HMMs [8] describes the application of discrete HMMs in recognition of six different tennis stroke classes in a constrained test environment. Other attempts to annotate the tennis matches were reported in [9,10]. Various constraints were present in small-scale experiments.

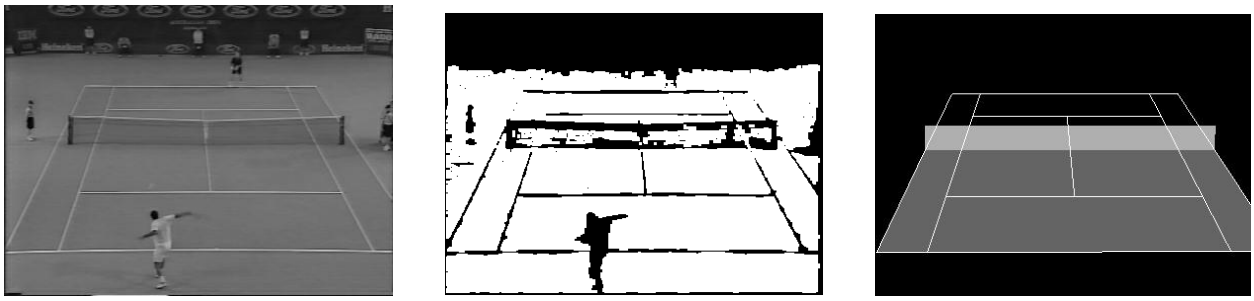
We define our problem as an optimization problem in the next section. Driven by domain knowledge we present an algorithm to solve our problem. First step in our approach is to segment the player from the background, which is supported by a robust player segmentation algorithm that is able to work even with low quality VHS video data. The algorithm is independent of type and colour of tennis court and allows us to run large experiments with videos from different tennis tournaments without changing the parameters of the algorithm. In the next step, we extract a number of different features from the player binary representation, believing that it is informative enough to reach our goals. Then, HMMs are trained using different feature combinations in order to find the optimal one for this particular application. Finally, experimental results are discussed and some conclusions are drawn.

2. PROBLEM DEFINITION

Our problem can be defined as an optimization problem. The optimization criterion on a high semantic level can be written as:

J= percentage of good classified events (tennis strokes)

Even for our specialized case of tennis strokes recognition, maximizing our optimization criterion over all possible tennis game videos present a hard to solve problem and it is even harder to prove the optimality of the solution.



a) Original image b) Initial segmentation c) 3D model after fitting to the lines

d) Final segmentation e) Pie features f) Skeleton features g) Influence function

Figure 1. Image segmentation and feature extraction

Anyway, using a number of heuristics and solving some sub-problems we try to move towards the desired maximum.

Famous Johansson's moving light displays [16] showed that people are able to recognize human activities provided by relatively little information (motion of a set of selected points on the body). Additionally, recent results showed that it is possible to recognise some human activities from their binary representations [17,18]. This, together with the common usage of HMMs for the action recognition, motivated us to approximate our problem with two sub-problems:

- segmenting player from the background.
- extracting some informative features from the player binary representation and training HMMs.

Our analysis and solutions for these two problems are described in the next two sections.

3. IMAGE SEGMENTATION

We begin from the game scenes (camera observing the whole field as in the Figure 1a). Using a number of global image features and some heuristics these scenes can be automatically extracted from the video but this is beyond the scope of this paper. Hindered by the low quality of the video and trying to make a clear approach not hampered by a large amount of domain knowledge heuristics needed, we constrain ourselves only to the player in the lower part of the image. Using our knowledge about the scene we develop a robust scheme for image segmentation that gradually leads us to the solution.

3.1. Robust dominant color extraction

It can be observed that the game scenes (Fig. 1a) have a dominant field color forming the largest cluster in the RGB color space. Inspecting several tennis surfaces and their scatter diagrams, we can conclude that field color distribution can be modeled by a 3-dimensional Gaussian:

$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, Cov) \qquad (1)$$

where $\mathbf{x}$ presents the 3-dimensional RGB color vector and $\boldsymbol{\mu}$ and Cov are the mean vector and the covariance matrix of the distribution.

Standard approach would be to use the Expectation Maximization (EM) algorithm [12] but it needs the pre-known number of clusters. A solution to this problem is the *minimum description length* principle [13]. We choose here for a much simpler approach. A simplified robust M-estimator [14] is used to estimate the parameters of the Gaussian discarding other color pixels as outliers.

We use a simple influence function presented in Figure 1g. Here d presents square root of the Mahalanobis distance:

$$d^2=(\mathbf{x}-\boldsymbol{\mu})^T\, Cov^{-1}\, (\mathbf{x}-\boldsymbol{\mu})$$

This reduces the algorithm to iterative repetition of the following two steps:

$$\boldsymbol{\mu}=\Sigma'\mathbf{x}/n$$

$$Cov= \Sigma'(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T/n,$$

where summing Σ' is done only for the n samples $\mathbf{x}$ whose Mahalanobis distance from the current mean estimate $\boldsymbol{\mu}$ and according to the current covariance estimate Cov, is less than c.

To prevent sudden changes we further low pass filter our estimates during the image sequence by:

$$\boldsymbol{\mu}(k)=\boldsymbol{\mu}(k-1)+\rho(\boldsymbol{\mu}'-\boldsymbol{\mu}(k-1))$$
$$Cov(k)=Cov(k-1)+\rho(Cov'-Cov(k-1)),$$

where $\boldsymbol{\mu}'$, Cov' present estimates from the current image and $\boldsymbol{\mu}(k)$ and $Cov(k)$ are the filtered estimates.

For the constant c we use c=4, ($4\text{-}\sigma$ *ellipsoid*) regarding the rest of the samples as outliers. The presented algorithm was tested on various tennis field types and various images using the same parameter c.

3.2. Algorithm

The segmentation algorithm is divided into three steps:

Step1: Using estimated statistics of the tennis field color we do the initial quadratic segmentation of the image as shown in Fig. 1b. We use Mahalanobis distance 4 as threshold. Using erosion and dilation morphological operations the player is detected as the largest compact region in the lower half of the image.

Step 2: The initial segmentation is also used, after thinning, to fit the tennis court lines model. The Gauss-Newton minimization procedure on Gaussian blurred images is used to fit the model. Multiple starting points are used for the initial sequence images together with the multi-scale approach using Gaussian pyramid [15] to increase the convergence area. The 3D model is constructed having in mind the visibility of the lines.

Step 3: After fitting the 3D model we get more knowledge about the scene at a higher semantic level (Fig 1c). This knowledge together with the initially segmented player is used to form the start values for robust estimation of the parameters of a number of Gaussian models (1) for different colors. Model parameters for the colors of the field, the lines and eventually the net are estimated using the data from some neighborhood of the initially detected player. These values are then used to refine the player segmentation. We do not expect the player to have the same color as the field or the net. On the other hand the players often have white clothes similar to the color of the field lines. Distinction here is done using our knowledge about the line positions. We discard all the structures that are thin in the line direction and are positioned according to the fitted model. Unfortunately this can result in removing isolated thin white color parts of the player, when they are aligned and overlapped with the field lines. Luckily, these situations are very rare.

4. FEATURE EXTRACTION

Features characterising the shape of the segmented player binary representation can be extracted in various ways. For our domain, general methods, like Fourier Descriptors (FD) or construction of linear subspaces using principal component analysis (PCA), are of very little help. Having a specific case of human figure in this particular application, we extract some specific parameters trying to maximize their informativeness. Except from the standard shape features such as orientation (f_1), and eccentricity (f_2), we extract the following features:

- position of the upper half of the mask with respect to the mass centre ($f_{3\text{-}4}$), its orientation (f_5), and eccentricity (f_6) - this describes the upper part of the body that contains most of the information.
- part of the mask contained in the circle cut out centred at the mass centre ($f_{7\text{-}14}$) as shown in the Fig. 1e - this can be seen as a general approximate description.
- similar to [17] we extract the sticking-out parts by filtering and finding local maximums of the distance from the point on the contour to the mass centre. Only certain angles are considered as indicated in Fig. 1f. We use only the indication if such a stick-out part is present on one of the sides which gives us two new features ($f_{15\text{-}16}$).

5. EXPERIMENTAL RESULTS AND DISCUSSION

For our experiments we used ordinary TV broadcast tennis videos with different players from different tournaments: Australia Open (Sampras-Agassi, Higins-Molik, Kournikova-Davenport, etc.), Swisscom Challenge, Vienna Open, etc. Training sequences are manually selected using a tool that has been developed for video annotation and pre-processing. We used first order left-to-right discrete HMMs with 4 to 48 states and codebook size from 4 to 80 symbols. In the training process, we conducted a number of iterations of the Baum-Welch algorithm with modified re-estimation formula for the training with multiple observation sequences [19]. In each experiment strokes were recognised by an isolated HMM as well as by multiple HMMs in parallel evaluation.

5.1. Experiment 1

In this experiment we aimed at two goals: (1) determine the right feature set and (2) investigate 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 including female ones, two series of experiments have been conducted: 1*a* and 1*b*. In the series 1*a* we used the same players in the training and evaluation sets, while in 1*b* HMMs were trained with only one player, but strokes performed by various players were evaluated. In both cases, the training set contained 40 different sequences, while the evaluation set contained 120 sequences.

To be able to compare our results with related approaches 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 recognise. In order to find the best HMM parameters, a number of experiments with different number of states and codebook size were performed for each feature combination.

Table 1. Recognition results

Feat.\Ex.	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%	83.1%	
$f_{2\text{-}4,\ 15\text{-}16}$	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

The recognition accuracies (% of rightly classified strokes using parallel evaluation in Table 1) 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 drop in experiment 1*b* as expected, but the combination of eccentricity, the mass centre of the upper part, and skeleton features ($f_{2\text{-}4,\ 15\text{-}16}$) pops up as the most person independent combination, which is nearly invariant on different player constitutions. The optimal result with this feature combination was achieved with the codebook size of 24 symbols and HMMs with 8 states.

Improvement of results in relation to related approaches comes from better training algorithm and mostly from improved, more informative and invariant features, in first place novel skeleton features and then pie features previously described.

5.2. Experiment 2

In this experiment, we investigated recognition rates of different feature combinations using regular classification of strokes from tennis literature [11]. There are 11 different strokes, namely: 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. After analysing of false recognized strokes, we found that only 15% comes from false recognition of similar strokes. The majority of false recognitions remained the same as in experiment 1. Nearly 65% comes from forehands recognized as backhands and opposite, as well as from forehand-volleys recognized as forehands and opposite. Having, for example, the ball position (an attempt is reported in [5]) would certainly make the distinction between mentioned strokes more robust and significantly increase the recognition rate.

6. CONCLUSIONS

In this paper, we have exploited the automatic learning capability of HMMs to extract high-level semantics from raw video data automatically. A set of novel features and a robust feature extraction scheme are introduced for the tennis domain. A number of experiments are done and the results showed that previously described skeleton features are of the greatest importance. However, they are also the most unstable due to segmentation errors that occur mostly in cases of quick arm movements. But that can be overcome by using high quality digital videos.

Analysis done on the real tennis video data has demonstrated that our approach can increase the number and the percentage of accurately recognizable events in comparison to mentioned methods. Nevertheless, experimental results with regular classification of tennis strokes show that our method is promising to realize statistics of tennis games automatically using normal TV broadcast videos. Extracting more information from our data, like the ball position or by using also the sound, would probably increase our recognition accuracy, which is the direction for our future work.

7. REFERENCES

[1] A. Del Bimbo, *Visual Information Retrieval*, Morgan Kaufmann, San Francisco, California, 1999.

[2] A. Yoshitaka, T. Ichikawa, "A Survey on Content-Based Retrieval for Multimedia Databases", *IEEE Transactions on Knowledge and Data Engineering*, 11(1), pp. 81-93, 1999.

[3] M. Petkovic, W. Jonker, "Overview of Data Models and Query Languages for Content-based Video Retrieval", *SSGRR International Conference*, L'Aquila, Italy, pp., 2000.

[4] R. Cipolla, A. Pentland eds., *Computer Vision for Human-Machine Interaction*, Cambridge University Press, 1998.

[5] G. Pingali, Y. Jean, A. Opalach, I. Carlbom, "LucentVision: Converting Real World Events into Multimedia Experiences" *IEEE ICME*, New York City, vol.3, pp. 1433 -1436, 2000.

[6] H. Jiang, A. Elmagarmid, "Spatial and temporal content-based access to hypervideo databases", *VLDB Journal*, 7(4), pp. 226-238, 1998.

[7] M. Petkovic, W. Jonker, “A Framework for Video Modelling”, In the Proc. of *International Conference on Applied Informatics*, Innsbruck, 2000.
[8] J. Yamato, J. Ohya, K. Ishii, “Recognizing Human Action in Time-Seuential Images using Hidden Markov Model”, In the Proc. of *IEEE CVPR*, pp. 379-385, 1992.
[9] H. Miyamori, S-I. Iisaku, “Video Annotation for Content-based Retrieval using Human Behaviour Analysis and Domain Knowlegde”, *IEEE AFGR*, pp. 320-325, 2000.
[10] G. Sudhir, J. Lee, A. Jain, “Automatic Classification of Tennis Video for High-level Content-based Retrieval”, *IEEE Workshop on Content-based Access and Image and Video Databases*, pp. 81-90, 1998.
[11] J. Yandell, *Visual Tennis*, Human Kinetics, 1999.
[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum-likelihood from incomplete data via the EM algorithm”, *J. Royal Statist. Soc*. Ser. B, 39, 1977.
[13] S. Ayer and H. Sawhney, Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding”, In *Proc. ICCV*, pp. 777-84, 1995.
[14] P. J. Huber, *Robust statistics*, John Wiley, 1981.
[15] A. Rosenfeld, ed., *Multiresolution image processing and analysis*, Springer-Verlag, 1984.
[16] G. Johansson, “Visual perception of biological motion and a model for its analysis”. *Perception and Psychoph*. 14(2), pp. 210-11, 1973.
[17] H. Fujiyoshi and A. Lipton, “Real-time Human Motion Analysis by Image Skeletonization” *IEEE Workshop on Applications of Computer Vision (WACV)*, Princeton NJ, pp.15-21, 1998.
[18] R. Rosales and S. Sclaroff, “Inferring Body Pose without Tracking Body Parts”, In *Proceedings IEEE CVPR*, 2000.
[19] S. Michaelson, M. Steedman, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, 1990.