Video Annotation for Content-based Retrieval using Human Behavior Analysis and Domain Knowledge

Hisashi Miyamori Shun-ichi Iisaku

Communications Research Laboratory
Ministry of Posts and Telecommunications
4-2-1, Nukuikitamachi, Koganeishi, Tokyo, 184-8795, JAPAN
miya@crl.go.jp

Abstract

This paper proposes an automatic annotation method of sports video for content-based retrieval. Conventional methods using position information of objects such as locus, relative positions, their transitions, etc. as indices, have drawbacks that tracking errors of a certain object due to occlusions cause recognition failures, and that representation by position information essentially has limited number of recognizable events in the retrieval. Our approach incorporates human behavior analysis and specific domain knowledge with conventional methods, to develop integrated reasoning module for richer expressiveness of events and robust recognition. Based on the proposed method, we implemented content-based retrieval system which can identify several actions on real tennis video. We select court and net lines, players' positions, ball positions, and players' actions, as indices. Court and net lines are extracted using court model and hough transforms. Players and ball positions are tracked by adaptive template matching and particular predictions against sudden changes of motion direction. Players' actions are analyzed by 2-d appearance based matching using the transition of players' silhouettes and hidden markov model. The results using two sets of tennis video is presented demonstrating the performance and the validity of our approach.

1. Introduction

Recently, the amount of video information has been rapidly increasing in various fields. Visual databases, video browsing, and video surveillance are expected to play an important role as major applications in the field of academic education, medical sciences, and training for artists, dancers, and sports players, etc. in the future. In such an environment, indexing, retrieval, and summarization will be the key function to achieve efficient browsing, which means that the video annotation will become increasingly important. The standardization activity is also scheduled to begin as MPEG-7, which is mainly focused on video description method for content-based retrieval.

Previous approaches to video annotation for content-based retrieval have mainly focused on visual features such as color, shape, texture, and motion[1]-[4]. Although they have their advantages to be applied to a wide variety of generic video, those visual features have a common drawback that it can represent only low-level information, Thus, it becomes very difficult for such features to represent high-level information as most users need when retrieving appropriate video segments.

In order to solve the problem that those generic indices can only represent low-level information, there have been several researches that try to make specific content retrieval more realistic by introducing their domain knowledges[5]-[8]. These researches mainly use position information of objects in a scene such as locus, relative positions, their transitions, etc., and analyze them in a specific domain like soccer, basketball, tennis, etc., in order to relate them to particular events which correspond to high-level information.

Since these approaches utilize domain knowledge, they can represent high-level information efficiently if all the necessary objects are successfully extracted. However, previous works still have a problem that the recognition easily becomes unstable due to partial tracking errors of objects or lack of necessary information. There is also an essential restriction to recogniz-

able events in the retrieval, since they only use position information in the analysis.

This paper proposes an automatic annotation method of sports video for content-based retrieval using human behavior analysis module and specific domain knowledge. Our approach incorporates human behavior analysis and specific domain knowledge with conventional methods, to develop integrated reasoning module for robust recognition. The advantage of our approach is that it can increase the number of different recognizable events in the retrieval and that it can realize robust recognition which is less affected by partial lack of information or identification errors.

The rest of the paper is organized as follows; In section 2, the system overview is presented. In section 3, the extraction of court and net lines using court model and hough transforms are illustrated. Players' and ball trackings by adaptive template matching are described in section 4 and 5, respectively. The analysis of players' actions are depicted in section 6, and the retrieval using profiles is shown in section 7. Several results on real tennis video are presented in section 8, and the conclusion is summarized in section 9.

2. System Overview

Figure 1 shows the block diagram of the proposed system. Our test domain is tennis. Our goal is to retrieve video segments containing tennis actions such as *forehand stroke, backhand volley, smashing, etc.* from the whole video of tennis matches.

Video annotation usually comprises of two steps; (1) shot partitioning of raw footage of input video, and (2) context identification based on certain model on each video segment. The first shot segmentation step is beyond the scope of this paper.

Tennis video generally includes various scenes like close-up shots focusing on each player, judges, spectators, etc., but most typical shots include tennis court, which are shooted from a diagonally upper position from the ground. In this paper, we also assume that the input is such shots including tennis court, which are already selected, for example, by certain color-based selection approach[8].

Annotation is processed by the following steps. First, court and net lines are extracted using court model and hough transforms on each frame. Then, player's positions are tracked considering the extracted court and net lines. Ball position is tracked using special prediction modes. Players' behavior is identified using players' shape changes extracted from inside the window centered at the players' position. These position and action information become the input to the integrated reasoning module, which outputs the action indices in the database. Position information is also stored as low-level indices.

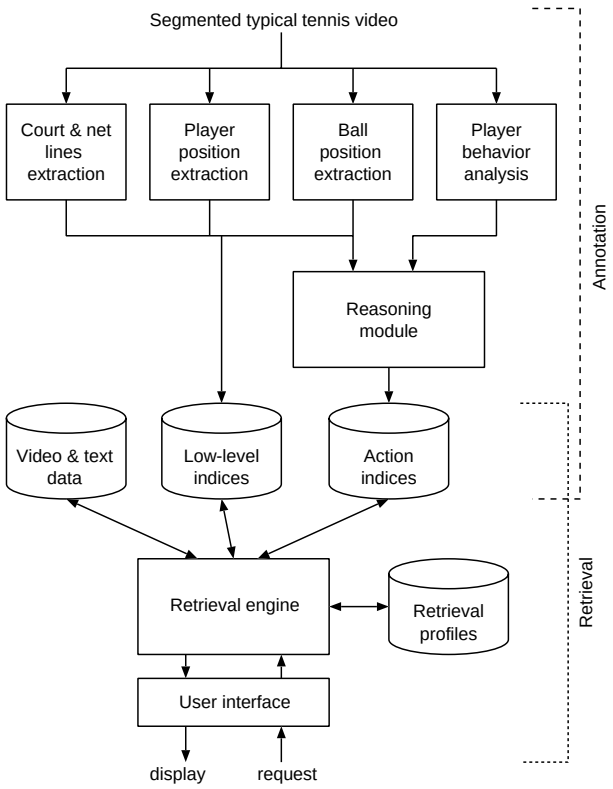

Figure 1. System overview

Content-based retrieval is realized by calculating the similarity using those indices based on the retrieval profiles which correspond to the input keyword. Retrieval profiles mean the definition of evaluation steps which relates retrieval keywords to some action and low-level indices. The profiles enable the retrieval of complicated formation plays such as *"The server puts the service in the opponent's backside, and then, the volleyer hits the returned ball"*.

Each operation of the annotation is explained in details in the following sections.

3. Extraction of Court and Net Lines

Extraction of court and net lines utilizes the specification of tennis court as the domain knowledge.

We define court feature points $P_{c_1}, \cdots, P_{c_{14}}$, court lines $L_{c_1}, \cdots, L_{c_9}$, and net feature points $P_{n_1}, \cdots, P_{n_3}$, net lines L_{n_1}, L_{n_2}, as shown in figure 2, and refer them in this paper. Extraction starts on court lines, and then, followed by net-line extraction.

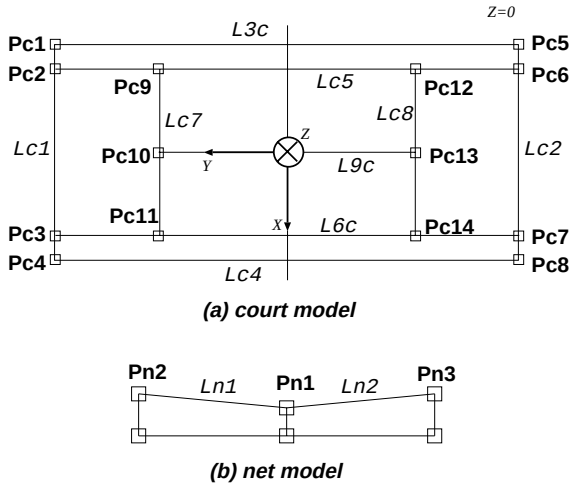

Figure 2. Court and net model

Each court line is decided by connecting two feature points at its both sides. The feature points are detected by the following steps:

1. At $t = 0$, initial feature points $P_{c_i}(0)$ are given as input. Then, each court line $L_{c_i}(0)$ is transformed onto the hough plane. After the peak detection, set the windows $W_{c_i}(0)$ of the size w_{th}, w_{ro}, centered at each peak on the hough plane.
2. At $t = t$, perform AND operation of the binary image $B(t)$ from the original image, and the adjacent regions of each court line $L_{c_i}(t-1)$, to generate partial binary image $B_c(t)$ including only adjacent regions of court lines (referred to as *court-line binary image* in this paper).
3. Transform each adjacent regions onto the hough plane, perform peak detections in each window $W_{c_i}(t-1)$, to update the feature points $P_{c_i}(t)$.
4. Transform court lines $L_{c_i}(t)$ onto the hough plane, and update the detection windows $W_{c_i}(t)$. Return to step 2.

Note that some feature points which go out of the screen due to camera pannings, etc., is also updated by the estimation using the connectivity knowledge of the court model (figure 3). Our assumption is that the middle-area feature points $P_{c_i}(t)$ (i=9,10,12,13 or 10,11,13,14) always stay in the visible area. By the same reason, some initial feature points can be skipped at step 1.

Net-line extraction basically has the same steps as court-line extraction.

Initial feature points $P_{n_i}(0)$ are given at $t = 0$, and set the windows $W_{n_i}(0)$ for each net lines $L_{n_i}(0)$ on the hough plane. At $t = t$, subtract the court-line binary image from the original binary image, to generate $B_n(t) = B(t) - B_c(t)$, referred to as net-line binary image. $B_n(t)$ is transformed onto the hough plane, followed by peak detection inside the windows, resulting the update of feature points $P_{n_i}(t)$.

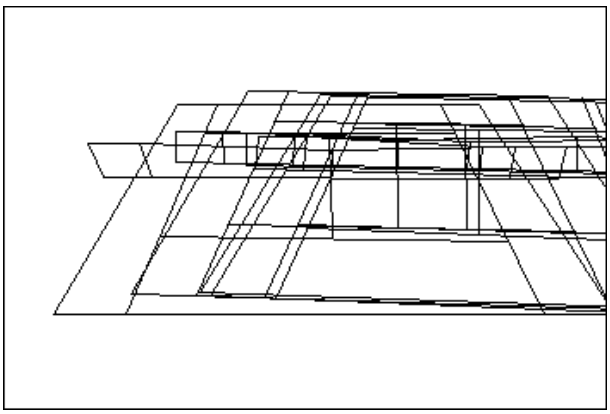

Figure 3. Tracking of court and net lines

4. Extraction of Players

Players' positions are extracted by adaptive template matching.

First, initial positions are detected as follows;

1. At $t = 0$, generate residue images $D_1(t), D_2(t)$ between images of the s-frame distance from the current frame;
   $D_1(t) = I(t) - I(t-s)$, $D_2(t) = I(t+s) - I(t)$
2. Perform AND operation between these two residues to make $D(t)$.
   $D(t) = D_1(t) \wedge D_2(t)$
   Fatten $D(t)$ by morphological dilation operation.
3. Put labels on each connected areas in $D(t)$. In order to exclude speckles, observe the areas over several sequences. Afterwards, let the center of the largest connected area $r(t)$ be the initial position of the player.

Then, the players tracking is performed by the following steps;

1. At $t = t$, set the template $T(x, y)$ of the size $w_x \times w_y$ at the center of the player's position.
2. Calculate the absolute difference between the template $T(x, y)$ and the candidate region $W(x, y)$ of the next frame inside the searching range.

$$R(x, y) = \sum_{i=1}^{w_x} \sum_{j=1}^{w_y} |W(i+x, j+y) - T(i, j)| \quad (1)$$

Note that we exclude from the absolute difference calculation, the area of court and net lines which is already derived using the domain knowledge in the previous section[9].

3. Select the position (x, y) which has the minimum value of $R(x, y)$ as the player's position at the next frame. Update the template by the $R(x, y)$. Repeat step 2, 3 until the last frame.

5. Extraction of the Ball

When dealing with the ball extraction, the following points should be considered[10].

- The size of the ball is small.
- The size of the ball sometimes changes due to perspectives and motion blurs,
- The direction of the ball is suddenly changed by ball hittings.
- The positions of the ball is sometimes invisible due to overlapped or hidden by court and net lines, players, etc.

Therefore, it would not be appropriate to apply position extraction of players directly to the ball extraction. We introduce special predictions which are switched according to the distance between the ball and the players. The ball extraction is done as follows;

1. Generate a template $T_b(x, y)$ of the size $b_x \times b_y$ which includes the ball beforehand.
2. Detect all candidate position of the ball which matches to the template $T_b(x, y)$ in the surrounding area of the players' positions.
3. Repeat matching for the following several frames to find candidates which moves radially from the player position center. Continue this process, and let the final single candidate be the detected ball position. Go to step 4.
4. Perform template matching using $T_b(x, y)$. Search range is centered at the predicted position from the current position using the previous displacement of the ball. Our assumption is that the ball locus can be approximated as a straight line in a short period. If the distance between the ball and the players becomes smaller than the threshold TH_d, go to step 2, otherwise repeat step 4.

Without ball predictions at step 4, tracking often fails when the ball is overlapped or hidden by the court and net lines. Ball tracking can be successfully continued to track on small objects, using the prediction modes used here.

6. Analysis of Player's Behavior

Players' behaviors are analyzed based on silhouette transitions, in order to get rid of the difference of player's clothes. We employ hidden markov model and 2-d appearance based model to identify behavior category.

First, player's silhouette is extracted as follows;

1. Generate color clusters using $r(t)$. The $r(t)$ denote the labeled points inside the player's position window used in section 4.
2. For the points not labeled inside the player's position window, label it as player's area if it is in the range of color distance TH_c from each cluster. The color clusters of past few frames can be utilized for stable labeling.
3. Subtract court and net lines from the labeled area. Afterwards, scan the adjacent area of court and net lines, and interpolate the subtracted area by drawing perpendicular line against the court and net lines.
4. Normalize the labeled area by the size $S_x \times S_y$, to make the silhouette image $S(t)$.

Silhouette transitions indicate specific features corresponding to human behaviors. Prepare the continuous silhouette images of several behavior patterns as training data, expand them to certain eigenspace, and store several high-ranked parameters which show unique changes to each behavior pattern, beforehand. Then, input silhouette data is expanded to the same eigenspace, find the nearest trained pattern using the selected parameters, to identify behavior category. The starting and ending points of each behavior are given when the distance between the opponent player and the ball becomes more and less than the threshold TH_p, respectively, in this paper.

The generic behaviors can be represented by combining the identified behavior IDs, each of which has the starting and ending points.

We define *foreside swing, backside swing, over-the-shoulder swing* as the behavior categories. Expansion to eigenspace is done by KL transform[11] in this paper.

7. Application to Content-based Retrieval

In the previous sections, we have extracted four basic components; (1) *court and net lines*, (2) *player's position*, (3) *ball position*, and (4) *player's behavior*.

When described as indices, player and ball positions are described as mapped position onto the court model using the results (1) and (2),(3), respectively. Player's behavior is defined as a time period between starting

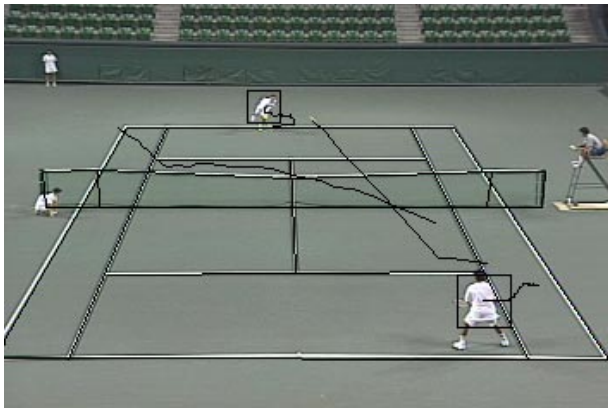

Figure 4. Tracking of players and the ball

Figure 5. Typical silhouettes of "over-the-shoulder swing"

point and ending point which has the behavior ID derived in the previous section.

Retrieval profiles define the evaluation condition which correspond to each retrieval keyword, and are described using one or more object behaviors, their temporal occurrence orders, distance between objects, arrangement of objects, etc.

Once the retrieval keyword is given, the scenes which satisfy the evaluation condition in the profile is searched, and presented on the user's display.

8. Results

Let us consider the following three representation methods of tennis actions, for comparison with the conventional methods.

method 1 player positions only[8].

method 2 player and ball positions.

method 3 method 2 plus player's behaviors.

Method 1 may be possible for rough classification, but often tend not very expressive in representation. Also, it usually ends up in ambiguous descriptions, since it only refers to player's positions on the court. Take the following representation for example; "*Both players stay around baselines during a certain period*". This description cannot reasonably discriminate *baseline-rally* from *the scene that the player is about to serve*, because players stay around the baselines in either scene.

Method 2 is considered to give more accurate representations of tennis actions. For instance, *"forehand stroke"* can be described in the following way;

- Let starting and ending point be the moment when the ball cross back and forth over the upper net, respectively.
- If the following conditions are satisfied, the input is identified as forehand stroke. At the moment when the ball is in the nearest position to the player,
  - the player stays around a baseline,
  - the ball is on the foreside of the player.

Likewise, we defined the definitions for several tennis actions, and performed content-based retrieval. Table 1 shows the retrieval results. The table shows that

Table 1. Retrieval result of tennis actions

keyword			# of occur.	errors	omissions
stoke	fore	b	60	3 (5.0)	0 (0.0)
		t	68	18 (26.4)	3 (4.4)
	back	b	84	1 (1.2)	0 (0.0)
		t	69	12 (17.4)	4 (5.8)
volley	fore	b	4	1 (25.0)	1 (25.0)
		t	3	1 (33.3)	0 (0.0)
	back	b	1	0 (0.0)	0 (0.0)
		t	1	0 (0.0)	0 (0.0)
service	b		46	0 (0.0)	0 (0.0)
	t		48	0 (0.0)	0 (0.0)

(Inside the parenthesis shows the ratio against the total number of occurrences. "b","t" mean "bottomside" and "topside" of the court, respectively.)

method 2 can reasonably recognize tennis actions to some extent, although the frequency of occurrence in the table is rather impartial, due to the video sequence used here. Typical errors occur when;

case 1 foreside and backside (right and left) is recognized the other way around,

case 2 the player hits smashes, and

case 3 the player hits lobs.

These errors indicate that the accurate identification becomes difficult using only position information.

Method 3 introduces the behavior information in addition to the positions. In this case, the *"forehand stroke"* can be defined as follows;

- Let starting and ending point be the moment when the ball cross back and forth over the upper net, respectively.
- If the following conditions are satisfied, the input is identified as forehand stroke. At the moment when the ball is in the nearest position to the player,
  - the player stays around a baseline,
  - the player behavior ID is *"foreside swing"*.

Recognition results show that the method 3 can successfully recognize the actions in case 1 (figure 4). Method 2 sometimes becomes unstable due to occlusion by the players, whereas the method 3 works rather robust in the sense that it observes the human behavior categories during a certain period.

Also, *"smash"*, which cannot be appropriately represented by method 2, can be described using the behavior ID of "*over-the-shoulder swing*". The action is identified in a successful manner in the retrieval (fig 6).

Note that *"lobbing"* cannot be identified by method 3, since it only employs two-dimensional ball positions. The estimation of three-dimensional ball position and extra process to deal with the ball going out of and returning to the screen may be needed for further study.

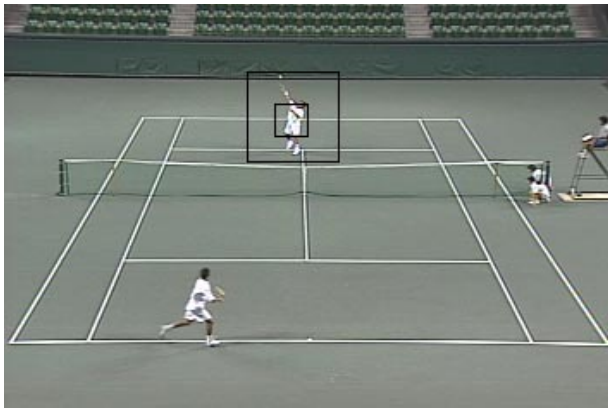

Figure 6. Retrieved action of "smash"

(The small window shows the template size used in player tracking, whereas the large window indicates the silhouette size after the labeling.)

9. Conclusion

Video annotation method for content-based retrieval using human behavior analysis and specific domain knowledge is proposed. After the court and net models are developed, court and net lines, player's positions, and ball positions are extracted on real tennis video. Player's behaviors are also analyzed using its silhouette images. Retrieval conditions are defined as profiles which correspond to each retrieval keyword, and are represented by combining previously extracted components to identify tennis actions. Retrieval results show that the proposed method can successfully identify several tennis actions which cannot be described, or which cause errors, by conventional methods. Tests applied to more video data remain for further research.

References

[1] S.Ravela, et. al.: "Retrieving images by similarity of visual appearance", In the Proc. of the IEEE Workshop on Content Based Access of Images and Video Databases, CAIVD'97, pp.67-74, 1997

[2] M.Flickner, et. al.: "Query by image and video content: the QBIC system", IEEE Computer Magazine, pp.23-32, 1995

[3] A.Nagasaka, Y.Tanaka: "Automatic video indexing and full-video search for object appearances", IPSJ Trans. Vol.33, No.4, pp.543-550, 1992

[4] A.Akutsu et. al.: "Video indexing using motion vectors", In SPIE Proc. Visual Communication and Image Processing '92, pp.522-530, 1992

[5] Y.Gong, et. al: "Automatic parsing of TV soccer programs", Proc. Int'l Conf. on Multimedia Computing and Systems, pp.167-174, 1995

[6] D.D.Saur, Y-P.Tan, S.R.Kulkarni, P.J.Ramadge: "Automated analysis and annotation of basketball video", Storage and Retrieval for Image and Video Databases V, SPIE-3022, pp.167-187, 1997

[7] T.Kawashima, K.Yoshino, Y.Aoki: "Qualitative image analysis of group behavior", CVPR, pp.690-693, 1994

[8] : G.Sudhir, J.C.M.Lee, A.K.Jain"Automatic classification of tennis video for high-level content-based retrieval", Proc. of IEEE Workshop on Content-Based Access of Image and Video Databases, CAIVD'98, 1998

[9] S.S.Intille, A.F.Bobick: "Visual tracking using closed-worlds", MIT Media Laboratory Perceptual Computing Section Technical Report No.294, 1994

[10] Y.Ohno, J.Miura, Y.Shirai: "Tracking players and a ball in soccer games", CVIM, Tech.report of IPSJ, 114-7, pp.49-56, 1999

[11] T.Watanabe, C.W.Lee, A.Tsukamoto, M.Yachida: "Method of real-time gesture recognition for interactive system", ICPR, pp.473-477, 1996