Recognition of Tennis Strokes using Key Postures

Damien Connaghan, Ciarán Ó Conaire, Philip Kelly, Noel E. O'Connor
CLARITY: Centre for Sensor Web Technologies
Dublin City University, Dublin 9, Ireland

E-mail: connagha@eeng.dcu.ie oconaire@eeng.dcu.ie
kellyp@eeng.dcu.ie oconnorn@eeng.dcu.ie

***This work is supported by Science Foundation Ireland under grant 07/CE/I1147.**

Abstract — **In this paper we describe an approach for automatic recognition of tennis strokes using a single low cost camera. Professional tennis is played at high speed so the ability to classify tennis strokes on camera is hindered by the rapid movement of the players. We have developed an accurate recognition system which can automatically index tennis strokes from video footage. We aim to evolve this system so that meta data, such as time codes and descriptions of the strokes played, can be automatically indexed for a training session or a match. This level of indexing would provide an excellent foundation for the development of next generation sports coaching systems. The aim of this paper is to coarsely classify the main stokes played in tennis, i.e. a serve, forehand or backhand. The proposed approach is evaluated against a real-world dataset, obtained from elite players in a competitive training match.**

Keywords — **action recognition, tennis**

I Introduction

As video cameras become cheaper and easier to install, it becomes possible to consider their use in many domains, where until recently it would have been considered impractical. One such area where there is an increasing interest in the use of low budget infrastructure is in sports coaching, where technology can give an athlete a competitive advantage. In collaboration with Tennis Ireland [1], the national governing body for the sport of tennis in Ireland, we have developed the TennisSense system at their coaching headquarters [2]. The system can collect large volumes of video but indexing this video into meaningful segments is very time consuming for tennis coaches. It is therefore paramount that this system can automatically index the video into meaningful segments. Accurate shot classification provides a ideal foundation for video indexing as annotating the shots played will lead to finer grained video indexing.

A major research challenge is to create an accurate tennis stroke recognition system, which can automatically annotate a variety of tennis strokes. This recognition system should identify key postures in each tennis stroke and then use those postures to classify one stroke from another e.g. classify serves from forehands and backhands. It should be noted that we do not perform classification at the finer granularity level, such as a flat serve, top spin serve, slice serve, top spin-slice serve or a kick serve. In this paper we report on our first development of such a system.

We recorded three sessions from three high ranked tennis players, one of which is the current senior rank one player in Ireland. Each session contained a series of serves, forehands and backhands. Using the data, we successfully trained a recognition system to classify an input stroke as a serve or not. This was achieved by calculating the distances of the input stroke against a training set consisting of serves. Previous work in this area involves using Motion History Images and Motion Energy Images to articulate what motion the target is engaged in [3]. Using motion images however is not well suited to a scenario where the actions, rotations and movements of the tennis players torso and limbs are rapid and numerous, so in our work we extend this previous approach.

It must also be highlighted that this success was accomplished on a single low cost IP camera, which offers great potential for the development of low budget sports coaching tools.

II Related Work

In our initial experiments to recognise tennis strokes, we created Motion History Images (MEI) and Motion Energy Images (MHI) of all the strokes as described in [3]. The process of creating motion images extracts the player as foreground from each binary image in a given sequence and joins all the foreground regions together into a single frame to represent the players movements over a given action. However we found that due to the movement and rotation of the player, the timing and precise movement information was lost. For this reason we decided to extract the maximum amount of information from each frame to make a more informed decision.

Another approach uses broadcast video to classify tennis strokes [4]. Zhu et. al. were able to recognise player actions based on motion analysis. To achieve this, a relationship was established between the movements of different body parts and the regions in the image plane. However, this approach only recognises two basic actions, a left swing and a right swing, whereas we want to classify serves from backhands and forehands, so this approach did not suit our application.

III Real-time player extraction

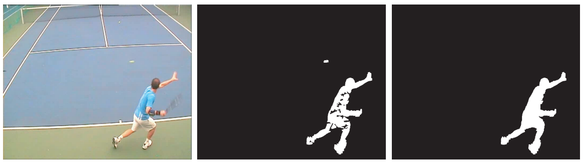
a) *Background subtraction*

Background subtraction is a well known technique for recognising moving foreground regions in computer vision as used in [5] [6] [7] and many more. This technique assumes a static camera is used and that image features, such as colour intensity or edge gradient information, of foreground objects differ to that of the background.

A basic method of background subtraction is to detect pixels belonging to foreground objects by determining if the difference between pixels in the current frame, f_i , and the corresponding pixels in a previous image consisting of a static background, b_i , are above a user defined threshold t. A pixel, (x, y), is declared as foreground if

$$|f_i(x,y) - b_i(x,y)| > t \qquad (1)$$

otherwise it is declared as background. In this work t is chosen empirically and background subtraction was used to create a binary image which contained the player as foreground. Since we intend to use a binary image of the player as input to the next stage of our algorithm, we simply applied further processing techniques on the image after background subtraction to smooth the foreground target. These techniques are discussed in the following section.



(a) Raw Frame (b) Background Subtraction (c) After morphological open function

Fig. 1: Processing the video frames and smoothing the foreground results in a clean result.

b) *Foreground Post-Processing*

The first pre-processing step is to clean up any noise within the foreground. This is done using a morphological open function [8]. This removes any small holes in the target and smooths out any noise . This approach also helps to inflate small features such as moving arm or leg segments.

It was also necessary to remove the tennis ball from the foreground. This was achieved by searching the binary image for any small blobs and removing them if they were smaller that a predefined threshold. The processing steps are illustrated in Figure 1.

IV Identifying Key Postures for the Training data

Each tennis player has his/her own style for each stroke, but there are also some common characteristics among different styles. One such example of a common characteristic between different styles of serve, is that the point of contact between the ball and racket occurs at an altitude greater than the height of the player.

By visual observation, we studied a variety of shots and identified three key player poses among all styles of serve, backhands or forehands. Figure 2 shows the three key postures common to all serves. Similar key postures were identified for backhands and forehands.

For each stroke in the training set we manually selected the three key postures to train the recognition system. This process was applied to serves, backhands and forehands.

V Developing a Recognition Framework

a) *Training Data*

After the video data was captured, we converted the entire video sequence into individual images. From the images, the strokes were manually segmented into serves, forehands and backhands for each player captured. Automatic segmentation is of course possible and is targeted for future work.

For a given stroke, we had an array of binary images which make up the shot played by the tennis player. The number of frames which can represent a tennis stroke vary in length but within this array exists the key postures which can be used to identify the type of stroke played as shown in Figure 2.

b) Statistical Comparisons of Strokes

For each binary image in the group that makes up a stroke, we compute statistical descriptions of these images using 7 Hu moments [9]. Hu moments are known to offer reasonable shape discrimination in a translation and scale invariant manner. Once we have the Hu moments for each image in the input sequence, we can classify the input stoke played.

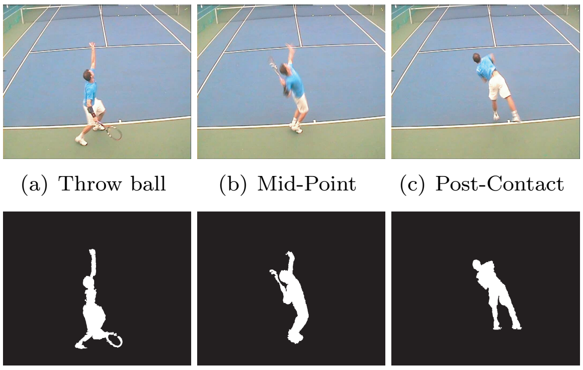
When the Hu moments are calculated, we analyze the values through a collection of serves to find the best key postures for representing a given tennis stroke such as a serve. The most suitable key postures would be the most common HU Moments across a collection of serves.

To measure the similarity of the Hu moments we used Mahalanobis distance [10]. This gives us the distance between the mean of the training set and the input frame. Mahalanobis distance is based on relationships between variables by which different patterns can be identified and analyzed. This metric determines similarity of an unclassified sample set to a classified one. It differs from Euclidean distance in that it takes into account the relationships of the data set and is scale-invariant. For this reason, Mahalanobis distance is ideal for our system, as we intend to classify tennis strokes performed by different players.

Each frame in the input vector is compared to the 3 key posture sets via Mahalanobis distance. A similarity matrix is constructed from these comparisons where row 1 consists of all the input vector comparisons to key posture 1 of the training set. Similarly rows 2 and 3 consist of comparisons to key postures 2 and 3 respectively. To get the best match for each key posture, we find the shortest path through each row of the matrix. A few simple rules are applied in that once the lowest match for key posture 1 was identified at position K, we assumed that the closest match to key posture 2 could only exist at location $K+1$ or greater.

After analyzing the moments from a series of serves, we identified a suitable threshold value, which could be used to classify the stroke in the training set from dissimilar input strokes. By calculating the mean of the distances for the training set of serves, we set the maximum distance threshold for a candidate input serve at 12.5.

We experimented with different sizes of training sets but 20 produced the best results. This is because the larger the size of the training results, the greater the variance in the tennis stroke. We found that ten strokes from two players gives a good representation of a stroke.



(a) Throw ball (b) Mid-Point (c) Post-Contact

Fig. 2: Key postures of a tennis serve, top row contains raw frames and bottom row contains the corresponding foreground.

VI Experimental Results

a) Visual Sensing

The video infrastructure consisted of a single IP camera with pan, tilt and zoom (PTZ) capability. The camera used was positioned behind the center of the baseline on the court and is part of the TennisSense system [2]. Nine different camera positions were tested, however capturing the players motions from behind, proved to yield significantly better recognition results. The camera is an AXIS 215 PTZ camera, which is positioned 2.8 meters above the court and has a high zoom functionality, as well as physical pan and tilt. The high zoom is useful for focusing on the player from behind the court baseline.

b) Data Capture

For the initial data capture sessions we consulted with the tennis coaches on a suitable format for capturing an adequate representation of tennis serves, forehands and backhands. It was advised that a player should spend five minutes warming up in order to find their true rhythm and this would then provide us with accurate data. For this experiment, data was captured from three right handed elite players.

Twenty serves from each player were recorded and each player was located left of center behind the baseline whilst serving. Each player was then fed thirty balls which were returned by a forehand or a backhand. These returns were played from a variety of locations on the court to create a realistic scenario. In total we had sixty minutes of data from three players, which contained sixty serves, over sixty forehands and forty backhands.

c) Recognition Process

To recognise an input action we built a training set consisting of twenty serves, ten from each player.

The first test involved the recognition of twenty input strokes. These input strokes contained a mixture of forehands and previously unseen serves from three players and the results can be viewed in Table 1, rows 1-20. Serves from two of these players were used to build the training set and the third player was unclassified, however none of the training data was reused as testing data. Inputs twenty to thirty in Table 1 are all backhands which are compared to a training set of twenty serves.

For each input in Table 1, the shortest distance to key posture one, two and three are recorded in columns KP 1, KP 2 and KP 3 respectively. The 'Result' column displays what the recognition systems output was and the 'Stroke Played' column displays what type of stroke was actually played.

To illustrate the results in Table 1, we will explain how the system obtained the results for input 9. After all the frames were processed, the Hu moments of each image in the input stroke were generated. Using the Mahalanobis distance we calculate how close each frame was to key posture 1. The closest distance to key posture 1 was 19.8. Likewise the closest frames to key posture 2 and 3 were calculated and the distances 16.1 and 36.1 were recorded respectively. These three values have a mean of 24.0 which is above the serve threshold of a maximum 12.5. A mean of 24 gives a high confidence that this input is not a serve and therefore the system has recognised the input stroke as a forehand. 'Stroke Played' displays what the actual stroke was and in this case it was a forehand.

Rows 20 to 30 are all backhand inputs so if the stroke played is not recognised as a serve then it will be classified as a backhand. As can be seen from inspecting the results in Table 1, a high level of accuracy has been obtained in identifying tennis serves from forehands or backhands.

Stroke Number	KP 1	KP 2	KP 3	Result	SP
1	4.4	3.2	7.7	S	S
2	6.8	3.4	6.5	S	S
3	6.6	6.3	4.7	S	S
4	14.6	15.4	117.3	F	F
5	10.0	21.9	5011	F	F
6	42.4	3	179	F	F
7	7.7	7.5	2.6	S	S
8	5.9	9.2	6.6	S	S
9	19.8	16.1	36.1	F	F
10	3.0	4.5	14.3	S	S
11	9.9	10.0	12.3	S	S
12	8.8	5.7	11.6	S	S
13	19.4	9.6	6.9	S	S
14	25.0	17.3	3.6	F	F
15	5.9	7.3	3.8	S	S
16	2.7	7.7	24.3	S	S
17	31.3	11.1	22.2	F	F
18	44.3	7.4	24.0	F	F
19	16.7	64.4	297	F	F
20	6.5	2.6	2.7	S	S
21	40.9	15.2	602.34	B	B
22	16.2	16.1	4.6	S	B
23	39.4	8.9	32.8	B	B
24	29.6	11.0	117	B	B
25	16.9	10.9	57.7	B	B
26	18.0	17.1	63.2	B	B
27	11.0	6.0	195	B	B
28	32.7	16.2	173	B	B
29	38.2	12.5	21.6	B	B
30	38.2	15.9	28.8	B	B

Table 1: Inputs 1-20 are a mixture of forehands and previously unseen serves and inputs 20-30 are backhands only. This recognition system identifies if the stroke played is a serve. KP = Key Posture. SP = Stroke Played

VII Conclusions and Future Work

The experimental analysis, although performed on a relatively low number of players, indicates a high level of accuracy, as illustrated by Table 1.

However, as a first attempt the system served well in bringing to light the issues to be considered for future research. At present we cannot classify different types of strokes played. It will be a future challenge to recognise a topspin serve from a slice serve, for example or a backhand from a forehand. Given that low cost cameras are being used here it will also be interesting to see how effective multiple cameras will be in helping to solve the underlying research challenges poised by identifying different types of tennis strokes.

The key postures were manually identified in this paper by inspecting the frames and visually identifying similar frames across a multiple of serves from different players. Whilst the similarity in Hu moments was used to verify the similarity of these key postures, a more advanced approach would be to build a system that can identify the best match for key postures and thus remove the need for a visual inspection. Given the computational overhead involved in executing this task and the time involved it was out of the scope of this paper for now, but it would be a paramount requirement in the future progress of this automated recognition system.

References

[1] Tennis Ireland. `http://www.tennisireland.ie`.

[2] D. Connaghan et. al. A sensing platform for physiological and contextual feedback to tennis athletes. In *BSN*, pages 224 – 229, 2009.

[3] J.W. Davis and A.F. Bobrick. The representation and recognition of action using temporal templates. In *CVPR*, pages 928 – 934, 1997.

[4] G. Zhu et. al. Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. In *ACM Multimedia*, pages 431 – 440, 2006.

[5] C Ó Conaire, P. Kelly, D. Connaghan, and N. O'Connor. Tennissense: A platform for extracting semantic information from multi-camera tennis data. In *DSP*, 2009.

[6] P. Kelly, N. E. O'Connor, and A.F. Smeaton. Robust pedestrian detection and tracking in crowded scenes. *Image and Vision Computing Journal*, 27(10):1445 – 1458, 2009.

[7] T. Bloom and A.P. Bradley. Player tracking and stroke recognition in tennis video. In *VSSN*, pages 93 – 94, 2006.

[8] J. Gil and R. Kimmel. Efficient dilation, erosion, opening, and closing algorithms. In *Pattern Analysis and Machine Intelligence*, pages 1606 – 1617, 2002.

[9] M. Hu. Visual pattern recognition by moment invariants. In *IRE Trans. Information Theory*, volume 8, pages 179 – 187, 1962.

[10] S. Xiang, F. Nie, and C. Zhang. Learning a mahalanobis distance metric for data clustering and classification. *Pattern Recognition*, 41(12):3600 – 3612, 2008.