Player Tracking and Stroke Recognition in Tennis Video

Terence Bloom
School of Electrical Engineering and Telecommunications,
The University of New South Wales,
Sydney, NSW 2052, Australia.

Andrew P. Bradley
Centre for Sensor Signal and Information Processing (CSSIP),
The University of Queensland,
St Lucia, QLD 4072, Australia.

Abstract

In this paper we present an investigation into the computer vision problem of tracking humans in digital video. The investigation domain is digital tennis footage and the aim is to track the tennis player and recognise the strokes played. The motivation behind this investigation is to eventually automate the task of digital tennis footage annotation so that metadata, such as the time codes and a description of the strokes played, are automatically appended to the video. This then enables a number of compelling applications, from simple search facilities for the home viewer, to more complex analysis tools suitable for a tennis coach. The system developed solves the problem of tennis player tracking and stroke recognition using relatively simple, and well known, image processing operations constrained by an a priori *knowledge of the image capture conditions, the background scene, and the application domain.*

Keywords: Player Tracking; Video Annotation; Metadata; Digital Tennis Video**.**

Introduction

Automatic human tracking is a computer vision problem. It is the problem of getting a computer to analyse the digital video stream of a scene, detect when a person enters that scene, and then to subsequently track that person's movement through the scene. There have been many investigations into this research topic and for various applications: From systems designed to track a single person and find their body parts, e.g., Pfinder [1] or W^4 [2], to systems designed to track multiple people and monitor their interactions, e.g., Computers Watching Football [3].

A computer capable of tracking humans would be useful for a wide range of applications from automated tracking for security or TV cameras, to a vision-based human-computer interface. A different type of application for this technology is to automate the task of annotating digital video with metadata that denotes the presence, and a description of the actions, of people in the video. With the increasing use of digital multimedia there is a corresponding increase in the need for tools that enable the fast and efficient indexing, querying, and browsing of multimedia databases. Emerging standards, such as the MPEG-7 audiovisual format, will support this concept by providing a standard language for metadata description schemes for multimedia content.

Annotating digital tennis footage with metadata such as time codes and a description of the strokes played would provide easy access to digital tennis footage in a video archive. A tennis coach, for example, would benefit from this by being able to easily retrieve training footage of certain strokes of his/her protégé in order to track their improvement over time, or more importantly, to analyse match footage of an opponent to identify their weaknesses. Another possible application for tennis player tracking and stroke recognition is the automation of statistics tallying to provide match commentators or home viewers with direct access to current match statistics.

The System

The aim of this research was to develop a computer vision system for tracking tennis players and recognising their strokes. This aim is the tactical precursor to the automatic annotation of digital tennis footage with metadata. The approach taken was to build an entirely software-based system and to work 'off-line' with digital video in the standard uncompressed AVI format. In this way, the work was focused on algorithm development rather than an efficient real-time implementation.

Some specific assumptions were made in order to be able to realistically achieve the aim. The most significant assumption is that the camera is essentially fixed in position with no panning or zooming used (although camera jitter is accounted for). Further, the only people on the court are assumed to be the two players and their shadows are considered to be part of the player. Finally, the tracking system requires some initialisation in the form of capturing the background scene with no players present. Identification of the key frames, i.e., the frames when a stroke is actually played (when the racquet makes contact with the ball), is supplied by monitoring the audio stream for the distinct sound of the impact. In addition, the raw footage is manually edited into the individual points (rallies)

and the system only attempts to track a single forecourt player.

As the tennis domain is very predictable, these assumptions are clear-cut, and can be relaxed one at a time by adding complexity to the system at a later stage.

The Algorithm

The algorithm is broken down into three units:

1. *Player Finding*: The player to be tracked is identified using a model of the background scene, standard image processing operations, and various *a priori* size and colour constraints.
2. *Player Tracking*: The player is tracked from one frame to the next by utilising the size and position of the player being tracked and their movement between consecutive frames.
3. *Stroke Recognition*: The key frames are identified as those frames where the racquet makes contact with the ball. These key frames are then further analysed to classify the tennis stroke as either: A forehand or backhand ground stroke, a volley, a smash, or a serve. Stroke recognition is performed using a three-stage algorithm based on the player's position in the court, finding the position of the player's racquet, and finally by finding the racquet arm of the player.

The structured and predictable nature of the tennis domain lends itself to exploitation similar to the notion of "closed-worlds" as applied to the football domain by Intille & Bobick [3]. That is, the visual processes that will be used to find the players are tailored using tennis domain knowledge and information that has already been learned about the player from previous processing. Furthermore, the complexity of the actual tracking problem is greatly reduced by some simple domain knowledge, e.g., that there are only two moving players, in two spatially distinct regions of the frames, i.e., we can expect them not to interact at all.

Player Finding

The traditional conceptual approach for tracking humans is to build and store a model of the scene, and then segment the humans in video by watching for variations from the scene model. In the tennis domain, the scene is the tennis court and the humans are the tennis players. A reference frame of the tennis court without any players present can be thought of as the simplest possible background model. Given a reference frame and a single frame from the tennis footage, the aim of the player finding unit is to derive, by visual means, the coordinates of the bounding box that encapsulates the player being tracked. Looking for large variations from the background, confirmed by the expected size and colour of the player, are used to achieve this.

The basic method for identifying large regions of variation is to first low-pass filter both the background frame and the current frame to remove random noise and reduce spatial detail. Then the pixel-by-pixel absolute difference is taken between them. The histogram of the difference image will exhibit a large dominant mode near zero (representing the static background pixels), and a smaller mode at a higher level (primarily representing the variation due to the player). Thus, by thresholding the difference image we can separate the player from the background by using a global threshold value found from the following equation:

$$threshold = mean + standard\ deviation / PSI.$$

Here PSI (Player Size Index**)** is a constant that is assigned based on the expected size of the player in the frame. In this way, the required percentage of pixels will be assigned to the player region after thresholding. The PSI has a limited number of values that relate to possible player sizes, e.g., when in the forecourt, on the baseline, or appearing or disappearing from the scene. If the player falls into a particular size range, the corresponding PSI is designated to that frame and used as the initial estimate for the next frame. The actual PSI values are calibrated manually for a given set-up.

Once the binary image has been obtained via thresholding morphological opening and closing operations, with square (4 x 4) structuring elements, are then used to remove small areas of noise. Then a connected component analysis is performed on the resulting image to identify the binary large objects (blobs) present. Finally, the blobs found are iterated through to find the one that is of the correct dimensions to be the player.

When the player is found in the first frame, a colour sample is taken as the average of a 5 by 5 grid of pixels from the centre of the player bounding box. Therefore, when more than one candidate player blob is identified (often the case when the player and their shadow are identified as two separate blobs) colour matching can be used to confirm the player blob. Colour matching is implemented as the sum of the absolute difference in each colour channel – sometimes referred to as a city-block distance [6]. The blob with the minimum distance to the initial colour sample of the player is then considered to represent the player. The operation of the three steps involved in the player finding unit are illustrated in Figure 1.

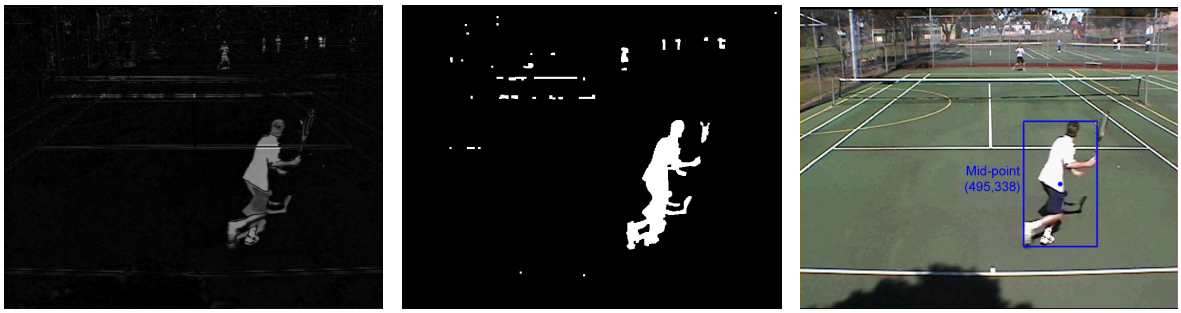

Figure 1: From left to right; the difference image produced from the pixel-by-pixel difference of a frame with the background reference frame. The difference image after thresholding and morphological filtering. The original frame with the player's position confirmed by identifying the bounding box of the blob that represents the variation in the difference image due to the player.

Player Tracking

Given the background scene image and a set of sequential frames of tennis footage, the aim is to be able to track the forecourt tennis player. That is, knowing at all times whether the player is present, and following their movement from frame to frame by maintaining the coordinates of their bounding box while they are present. The player finding unit is deployed to find the coordinates of the player's bounding box in each frame. While the player tracking unit, described here, applies robust inter-frame tracking techniques that further improves the reliability of the player finding unit.

Tracking requires a manual initialisation step for each point (rally) in the form of simply confirming the approximate position and size of the player found in the first frame. A flag indicating whether the player is present is then set accordingly, and maintained throughout. The player presence flag is only changed if the player disappears from, or re-appears into, view. The player is considered to have disappeared if the player finding unit reports the player as not being present in the current frame and the player bounding box had one side lying on the edge of the frame in the previous frame. Similarly, the player is said to have re-appeared if the coordinates of the player bounding box have one side lying on an edge of the frame and the player was not present in the previous frame.

The mid-point of the player bounding box is considered to be a good representation of the player's current position. Therefore, when the player is present, the bounding box found is confirmed by the player tracking unit if its mid-point has moved less than a pre-set distance between consecutive frames.

Stroke Recognition

Given the original set of blobs found by the player tracking unit for a key frame, i.e., a frame known to correspond to a tennis stroke, and the coordinates of the already confirmed player position, the aim of the stroke recognition unit is to make an elementary attempt at recognising the stroke played. Five different generic stroke types are defined in the current system, although there are many possible variations. Currently, the classified stroke types are: Forehand ground stroke; backhand ground stroke; volley; smash; and serve. On a conceptual level the player size, racquet orientation, and stroke timing can all be used to recognise these generic stroke types.

The stroke recognition unit utilises a three-stage algorithm. Initially, the player's relative size and position within the frame is used to distinguish the volleys from any other stroke (as a volley, by definition, is the only stroke played at the net). Hence, if the player's PSI is in the small size range and the player is at the net, the stroke is considered a volley. Next, if the stroke is not a volley, an attempt is made to find the position of the racquet head relative to the player's position. There are three distinct possibilities considered here: Above, to the right, or to the left, of the player. Assuming that we are tracking a right handed player and that we are looking at the forecourt player from behind, if the racquet head is found to the right of the player, the stroke must be a forehand ground stroke; if the racquet is found to the left of the player, the stroke must be a backhand ground stroke; and if the racquet is found above the player, the stroke must either be a smash or a serve. Finally, the timing of the stroke within a point is used to distinguish a serve from a smash; a serve being the only shot that is played at the beginning, or with the first few frames, of a point starting.

Figure 2: Left: The maximum possible racquet head size just fits into the square bounding box in red. A maximum allowable radius from the mid-point of the player is illustrated in blue. The green right-angled triangle formed from the mid-points of the player and the racquet is used to calculate the distance of the racquet from the player. Middle and right: A blob representing the racquet is found on the left of the player, thus the stroke is recognised as a backhand ground stroke.

The robust determination of the racquet position is the crux of reliable stroke recognition. This is achieved through further image processing on the already segmented binary difference image. It is expected that the disturbance created by the racquet head would often have been disconnected from the player's disturbance during previous visual processing. Thus, the first approach for finding the racquet is to look for a separate blob that could possibly represent it. As the time of a stroke is defined as the instant the racquet makes contact with the ball, it frequently happens that the disturbance of the racquet and the ball coincide, in fact, the ball's disturbance is often more solid and distinctive than that of the moving racquet. Hence, the racquet blob found is likely to be a somewhat blurred combination of the racquet, ball, and the court background.

The problem of finding the racquet blob is similar to that of finding the player blob. The full list of blobs present in the image is iterated, eliminating those that either contain less than a minimal number of pixels, or are too large (in either dimension) to be the racquet. These size criteria, which are easily established, discount most small noise blobs and other large disturbances (such as the net). In addition, to the size criteria, blobs beyond a pre-determined distance from the mid-point of the player can also be eliminated. These visual criteria are illustrated in Figure 2. The final stage for confirming a racquet blob is to take a 3 x 3 colour sample from the centre of each blob and to compare them to a known colour sample of a racquet. A blob is considered to be a match in colour only if all three channels are within a required range. If a blob is found to match, it is regarded as representing the racquet head. The mid-point of the racquet relative to the mid-point of the player is then used to classify the stroke played.

If none of the remaining blobs are found to be a good colour match, then it is assumed that either the racquet is contained within the player blob, or its disturbance is too distorted or discoloured. In this case, the stroke recognition unit undertakes further analysis of the player blob in order to find the arm holding the racquet. The player blob is first further morphologically closed through another iteration each of a dilation and erosion with a large, 7 x 7, structuring element. Then the player blob is skeletonised, using the Skeleton Zhou algorithm [4], to derive a 'stick figure' of the player, an example of which is shown in Figure 3. The major assumption used here is that the longest branch originating from a node in the top half of the player's skeleton (which discounts branches representing legs and shadows) represents the arm holding the racquet. Thus, the end-point is a good representation of racquet position relative to the player and so the stroke can be classified.

There were a few different approaches that could have been used for interpreting a player skeleton, such as the Hough Transform [5]. However, a more simplistic approach was used here and was found to be robust. The end-points of the skeleton are found by looking for pixels that have exactly one connected neighbour. Nodes are pixels with three or more connected neighbours. The length of the branch for each end-point is measured by counting pixels while iterating through to the closest node. If this node is in the upper half of the player, this branch is considered as possibly representing the player's arm holding the racquet. Once all of the branch lengths have been found, the longest branch is selected as the end-point representing racquet position.

The mid-point of the racquet blob, or end-point of the player skeleton, is therefore used to represent the position of the racquet. When compared to the mid-point of the player, using the heuristics described earlier, the stroke can be classified as a forehand, backhand, or overhead. The five generic strokes can thus, in the majority of cases, be recognised, especially when they are played clearly and in textbook style.

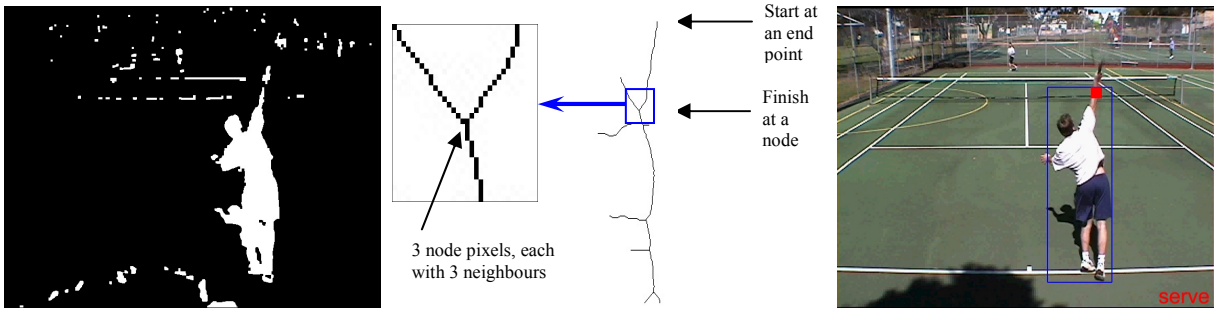

Figure 3: A frame where the racquet's disturbance was unable to be confirmed as all candidate blobs failed colour matching. The skeleton of the player is instead analysed in an attempt to find the arm holding the racquet. The top right branch was identified, and thus the position of the racquet relative to the player was found, helping identify the stroke as a serve.

Discussion

The stated aim of the work has been achieved, albeit under some restrictive assumptions. The forecourt player in the available footage was tracked quite robustly, even when exiting from, or re-entering into, view. The generic strokes were correctly recognised for the majority of examples tested. However, the algorithm indeed mis-recognised a stroke when either of the two assumptions broke down, i.e., that the colour sample used to match the racquet blob is accurate and representative; and that the skeleton's longest branch originating from the upper half represents the arm holding the racquet. However, as a first attempt the system served well in bringing to light the issues to be considered in future research.

It is believed that through further work to relax the necessary assumptions, the system would be capable of working well on real tennis training or even match footage. Thus, the algorithms described here could form the core technology of a system used for automatic tennis archive annotation. Significant further work is also required to improve the computational performance of the system so that it works in real-time. However, the methods used and described here are well studied and so there are a number of hardware and software solutions already available for this task.

The system's limitations are easily identified, and thus able to be addressed. To achieve more dynamic and reliable tracking, accounting for shadows, tracking both of the players, and handling non-players should be addressed. To be able to work with match footage the system should be able to work under camera movement, such as pan and zoom, and quickly and adaptively build up the required background scene image. To be more useful for annotation and statistics tallying, the system would be required to be capable of selecting the best key frame of when a stroke is played and preferably be able to distinguish between a greater variety of strokes than the five used in this study.

Conclusions

The point of departure for this research was the knowledge that current computer vision technology can be used to analyse a digital video stream to find a human moving in a scene. One possible application of this technology is to perform the useful, real-world, task of automatically annotating digital video streams with metadata that can then be used to search or summarise the video content. This research has demonstrated that for video footage of a single forecourt tennis player, captured and analysed under certain constraints, it is possible to solve this computer vision problem and hence to develop these types of applications. This was achieved by breaking the problem down into conceptual parts, solving these parts one by one by the application of relatively simple image processing techniques and some domain specific knowledge.

References

[1] 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, 7, pp.780-785, 1997.

[2] I. Haritaoglu, D. Harwood, & L. Davis, W^4: Who? When? Where? What? A real time system for detecting and tracking people, *International Conference on Face and Gesture Recognition*, Nara, Japan, 222-227, 1998.

[3] S. S. Intille & A. F. Bobick, Visual tracking using closed-worlds, Massachusetts Institute of Technology, Media Lab Perceptual Computing Group Technical Report No. 294, 1994.

[4] C. Quek & G. S. Zhou, A novel single-pass thinning algorithm and an effective set of performance criteria, *Pattern Recognition Letters*, 16:1267-1275, 1995.

[5] R. C. Gonzalez & R. E. Woods, *Digital Image Processing*, Addison-Wesley Publishing Company, Inc, 1993.