CLASSIFICATION AND REPRESENTATION OF SEMANTIC CONTENT IN BROADCAST TENNIS VIDEOS

N. Rea, R. Dahyot and A. Kokaram

Department of Electronic and Electrical Engineering.
Trinity College Dublin, Dublin 2, Ireland.
oriabhan@tcd.ie

ABSTRACT

This paper investigates the semantic analysis of broadcast tennis footage. We consider the spatio-temporal behaviour of an object in the footage as being the embodiment of a semantic event. This object is tracked using a colour based particle filter. The video syntax and audio features are used to help delineate the temporal boundaries of these events. For broadcast tennis footage, the system firstly parses the video sequence based on the geometry of the content in view and classifies the clip as a particular view type. The temporal behaviour of the serving player is modelled using a HMM. As a result, each model is representative of a particular semantic episode. Events are then summarised using a number of synthesised keyframes.

1. INTRODUCTION

Automatic retrieval of high-level content from broadcast TV footage has occupied the video processing research community for several years. In this time, broadcast sport has proved to be a domain which has received much attention as a result of being amongst those TV channels showing the fastest growth across EU member states [1].

Retrieval of high-level events from broadcast sports is a non-trivial task. Correlating those actions which occur in the footage with notions of the viewer is known as "bridging the semantic gap" - an ubiquitous problem in retrieval tasks. One way of circumventing this gap is to confine the retrieval problem to individual domains where the user context can be more easily addressed. Low-level features can be mapped to high-level concepts by applying well understood domain rules. This approach has yielded good results in broadcast coverage of soccer [2], baseball [3], tennis [4] and snooker [5] amongst others.

In this paper, we call on our past experience in snooker event classification [5]. In order to prove that this approach can be generalised to other sports, we attempt to use the same framework for high-level content retrieval from broadcast tennis footage. The system initially parses the footage based on the geometrical and colour content. In the appropriate view, a colour based particle filter is employed to track the players as they move about the court. We then assume that the spatio-temporal behaviour of the serving player can be considered to embody a semantic event. This behaviour is modelled by a discrete HMM. Models for each event are created by training the HMMs using human interpretation of the events. Three footage sources (*Pierce*, *Malisse* and *Hewitt* of 2949, 4114 and 12009 frames in duration) are parsed and events in each are classified and summarised.

Work sponsored by Enterprise Ireland Project MUSE-DTV (**M**achine **U**nderstanding of **S**ports **E**vents for **D**igital **T**elevision), CASMS (**C**ontent **A**ware **S**ports **M**edia **S**treaming) and EU-funded project MUSCLE (**M**ultimedia **U**nderstanding through **S**emantics, **C**omputation, and **LE**arning FP6-507752)

2. RECONSTRUCTING THE GLOBAL VIEW AND SHOT-LEVEL PARSING

Most sports footage is typified by non-action sequences interspersed with action. These high-value semantic events span short durations at irregular intervals. In order to extract such events it firstly becomes necessary to determine where in the footage these events are most likely to occur. We assume that the main semantic content can be extracted from the global view. In tennis this is the view in which live play typically takes place - the entire court can seen and the players can be tracked. Figure 1 shows a typical example.

Typically the playing surface colours contribute toward a large proportion (typically $\geq 70\%$ in tennis) of the overall colour distribution in the global view. In this view, peaks in the individual distributions correspond to the playing area. An adaptive threshold is used for segmentation. It employs a greedy algorithm which can adapt to the colour content. The idea is to select $r\%$ of the histogram centred on the mode. This is achieved by integrating bins to the left and right of the mode (i.e. select the right bin over the left if $\mathcal{H}(i+1) > \mathcal{H}(i-1)$ where $\mathcal{H}$ is the histogram and i is a particular range of bin values), until the threshold is met.

The HSV colour space is used to segment the tennis court. As white has a high brightness and low saturation, pixels with values greater than the range produced by applying the algorithm to the brightness component are considered to be non-court surface pixels. Those values less than the range returned by the saturation component also contribute to the court lines. The binary 'and' operation between the two thresholded colour spaces retrieves the court lines (figure 1) and players.

The geometry of a tennis court is similar to that of a snooker table [5]. By exploiting the Radon transform (which is robust to outliers such as the players) of the segmented image and the fact the diagonals of a trapezoid intersect at its centre, relevant lines

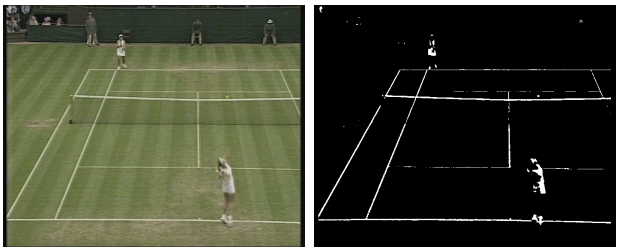

Fig. 1. Segmented court lines using the greedy histogram.

can be retrieved (e.g. tram lines). Due to the dynamic nature of the game however, tennis footage exhibits a great deal of horizontal translational camera motion as the camera pans to follow the main action on court. As a consequence, horizontal camera translation in image space results in a vertical translation of the projections in Radon space. Furthermore, since the camera capturing the global view is fixed at the centre of the court, camera panning will cause a perceived rotation of the lines about the fixed location of the camera filming the action. This induces a horizontal translation of the projection in Radon space.

The geometry can be reconstructed without the need for motion compensation by exploiting the symmetry of the court in Radon space and assuming that at least 2 'vertical' lines and 4 or 5 horizontal lines are always seen. To simulate the 'hidden' lines, a frame is extracted where the ρ value (the perpendicular distance to the line from the origin, located at the centre of the image) to each of the vertical lines is equal (from figure 2, where $\rho_1 \approx \rho_8 \approx \rho_o^*$ and $\rho_2 \approx \rho_9 \approx \rho_i^*$). This is referred to as the 'centre frame' and is used to calibrate the dimensions of the court in the presence of pan. By calculating the drift of the relevant peak in the current frame from its location in Radon space in the centre frame, the perpendicular distance to the corresponding hidden tram line can be approximated as:

$$\begin{aligned} \rho_o^{(t)'} &= \rho_o^* + (\rho_o^* - \rho_o^{(t)}) \\ \rho_i^{(t)'} &= \rho_i^* + (\rho_i^* - \rho_i^{(t)}) \end{aligned} \tag{1}$$

Where $\rho_o^{'}$ and $\rho_i^{'}$ are the distances to the hidden outer and inner

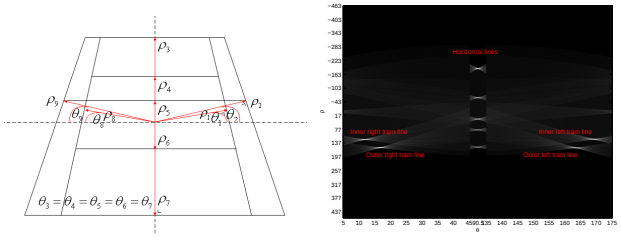

Fig. 2. Illustration of the tennis court geometry. Left to right: A schematic of a tennis court with parameters (ρ, θ) for each line; The Radon transform of the schematic.

tram lines respectively and $\rho_o^{(t)}$ and $\rho_i^{(t)}$ are the distances to the outer and inner visible tram lines in the current frame t.

In a similar fashion to estimating the values for the ρ parameters, θ values (the angle between the horizontal axis and the orthogonal of the appropriate line) for the right hand tram lines are approximated using equations 2 and the left hand tram lines are approximated using equations 3, where θ_o^* and θ_i^* are the values of θ in the centre frame.

$$\begin{aligned} \theta_o^{(t)'} &= \theta_o^* + (\theta_o^* - \theta_o^{(t)}) \\ \theta_i^{(t)'} &= \theta_i^* + (\theta_i^* - \theta_i^{(t)}) \end{aligned} \tag{2}$$

$$\begin{aligned} \theta_o^{(t)'} &= 180 - \theta_o^* + (180 - \theta_o^* - \theta_o^{(t)}) \\ \theta_i^{(t)'} &= 180 - \theta_i^* + (180 - \theta_i^* - \theta_i^{(t)}) \end{aligned} \tag{3}$$

The resulting angles $\theta_i^{(t)'}$ and $\theta_o^{(t)'}$ are those of the corresponding tram lines on the opposite side of the court, where i stands for inner and o outer. θ_o and θ_i are the angles of the lines in the current frame t. Fully reconstructed global views are shown in figure 3.

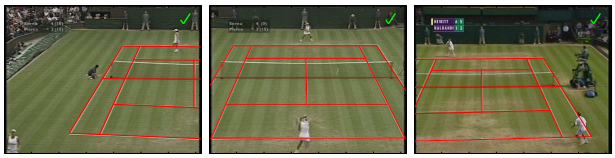

Fig. 3. Court line location detection robust to camera movement. The interpolated lines (shown in red) are overlaid on the existing white lines.

3. PLAYER TRACKING

An event in tennis is embodied by the motion of the serving player in the time period between the detection of the first racquet hit (racquet hits are detected using the method outlined in [6]) and a transition from a global view to a non-global view.

A particle filter is employed to track the players as they move around the court. A set of candidate particles, $\{(q_t^{(n)}, w_t^{(n)})|n = 1...N\}$ (where $q_t^{(n)}$ and $w_t^{(n)}$ are the state information and weights of N particles respectively) is diffused around the projected player location using a deterministic second order auto-regressive motion-model and a stochastic Gaussian component.

Since the (real world and scene) geometry of the playing surface is known and the height and width can be calibrated for each player (here, height is taken as 1.8m and width is 0.5m), a deterministic estimation of the size of the candidate regions can be approximated using the perspective distortion information. This ensures that the correct region size is always used and there is less influence in the candidate colour histograms due to background pixels.

The likelihood used attempts to match a target model of the player's HSV colour distribution (which is created in the first frame of the event) with each of the candidates. A Bhattacharyya distance measure is used to calculate the similarity between candidate histograms, ρ, and the target, ξ. It is in turn used to weight the sample set. Likelihoods are computed as:

$$p(\rho_t^{(n)}|q_t^{(n)}, \xi) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(1-h[\rho(q_t^{(n)}),\xi])}{2\sigma^2}} \tag{4}$$

Where $h\left[\rho,\xi\right] = \sum_{j=1}^{m}\sqrt{\rho^{j}\xi^{j}}$. Furthermore, to encourage better tracking, the known background colour can be used to formulate a likelihood ratio, $w_t^{(n)}$, where κ is the measured background histogram specified by the user.

$$w_t^{(n)} = \frac{p(\rho_t^{(n)}|q_t^{(n)},\xi)}{p(\rho_t^{(n)}|q_t^{(n)},\kappa)} \tag{5}$$

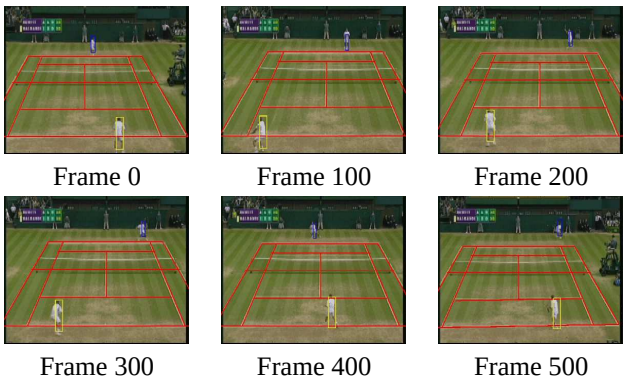

Fig. 4. Player tracking of a rally event. Both players are tracked over a sequence of 500 frames. More tracking results can be viewed at [7].

4. EVENT CLASSIFICATION

Analogous to the approach used for on-line handwriting recognition [8], active regions in the tennis court may be identified as in figure 5. The trajectory of a player is therefore encoded by the traversal through the regions. High-level events such as: Aces (λ_1), Faults (λ_2), Double Faults (λ_3), Serve and volleys (λ_4) and Rallies (λ_5) can therefore be classified similar to the recognition of the strokes of a stylus tracing out characters.

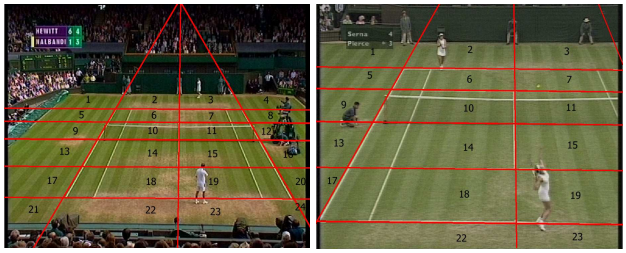

Fig. 5. Spatial segmentation of the tennis court into 24 sections.

Modelling the temporal behaviour of the serving player is accomplished using a first order HMM. The model topology is derived from the observations, $\mathcal{X}_1$, reflecting the nature of the target patterns. Each vertical column of the court (figure 5) ($N = 4$) is represented by a state and observations are the three cells in each column (since the serving player can only be on one side of the court i.e. $K = 12$). A model λ, can be defined for each of the competing events.

The Baum-Welch algorithm is used to find the maximum likelihood model parameters that best fit the training data. As the events are well understood in terms of the geometrical layout of the court, the models can be trained using human understanding. Training data is collated via a GUI where 6 users specified 5 tracks for each event in terms of the temporally evolving court coding sequence of the serving player.

Each semantic episode can then be classified by finding the model that results in the greatest likelihood of occurring:

$$\hat{\lambda} = \arg\max_{1\leq c\leq \mathcal{C}}[P(\mathcal{X}_1|\lambda_c)], \qquad \mathcal{C} = 5 \text{ events.} \tag{6}$$

To augment the retrieval, further features are extracted from the footage. The observations therefore take the form:

$$\mathbf{X} = \{\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4\} \tag{7}$$

- $\mathcal{X}_1$ is the spatially encoded trajectory of the serving player.
- $\mathcal{X}_2$ is a binary value which registers the first racquet hit which occurs in each global view.
- $\mathcal{X}_3$ is a binary value representing which side of the court the player is on for their detected service (left 0, right 1).
- $\mathcal{X}_4$ is a binary value which is 1 if the time between the first and second racquet hits in a global view is greater than 5 seconds.

The models used and binary operations are shown in figure 6.

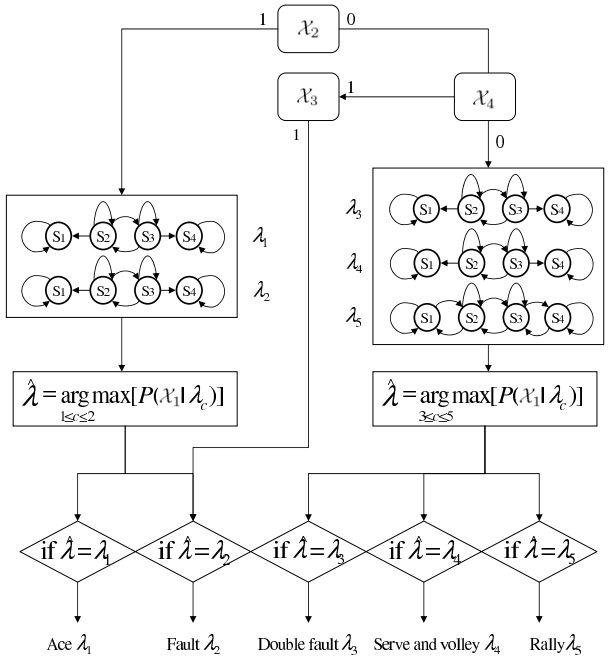

Fig. 6. Event HMMs with binary classifiers.

The tabulated confusion matrix (table 1) shows two misclassified rallies. In *Hewitt*, one of the rallies is quite short in duration. This is analogous to the motion behaviour of the player for a double fault event. A similar problem occurs in *Pierce* where a short

rally exhibits a comparable track to a fault where the player remains in one state. The final misclassification results from a fault being classified as a rally. In *Malisse*, the single observed ace event is a game point winner. The camera follows the player for a short duration as he returns to his seat. His motion is similar to that of an attacking serve and volley play, and is classified as such.

Event	λ_1	λ_2	λ_3	λ_4	λ_5	Missed	**Total**
λ_1	**2**	0	0	1	0	0	3
λ_2	0	**7**	1	0	0	0	8
λ_3	0	0	**2**	0	0	0	2
λ_4	0	0	0	**2**	0	0	2
λ_5	0	1	1	0	**17**	0	19
Total	2	8	4	3	17	0	34

Table 1. Confusion matrix for event classification in all tennis footage

5. SUMMARISATION

A keyframe is generated for the time between each pair of successfully detected racquet hits. Therefore, depending on the event type and duration of event, a number of keyframes are required for summarisation. The keyframes consist of a synthesised representation of the court and players. Given that the locations of the court lines are known, they can be compensated for camera induced translation and rotation. A motion history (the player position is sampled on each fifth frame) of the players depicts their behaviour during the racquet hits. Since the ball is difficult to track, its trajectory is approximated as the line that connects the locations of the two players when they hit the ball. A motion history illustrates the velocity profile of the ball. A summary of 2 keyframes from a rally event is shown in figure 7. This kind of primitive graphical summary could be sent to low bandwidth devices (such as in the form of an MMS message to a 2.5G compliant device) or to browse events in a game without having to view the entire sequence.

6. FINAL COMMENTS

In this paper high-level events in tennis were automatically classified by modelling the spatio-temporal behaviour of the serving player. Events were then summarised using a number of keyframes composed of a synthesised tennis court onto which a representation of the behaviour of the players and ball was overlaid.

7. REFERENCES

[1] European Audiovisual Observatory, "The impact of transfrontier broadcasting services on television markets in individual member states," Transfrontier Television in the European Union: Market impact and selected legal aspects, March 2004, URL: `http://www.obs.coe.int/online_publication/transfrontier_tv.pdf.en`.

[2] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala, "Soccer highlight detection and recognition using hmms," in *IEEE International Conference on Multimedia and Expo (ICME '02)*, August 2002, vol. 1, pp. 825–828.

[3] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden markov models," in *Proceedings of the International Conference on Image Processing (ICIP '02)*, September 2002, pp. 609–612.

[4] E. Kijak, P. Gros, and L. Oisel, "Temporal structure analysis of broadcast tennis video using hidden markov models," in *SPIE Storage and Retrieval for Media Databases*, January 2003, pp. 289–299.

[5] N. Rea, R. Dahyot, and A. Kokaram, "Modeling high level structure in sports with motion driven hmms," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, May 2004, vol. 3, pp. 621–624.

[6] R. Dahyot, A. C. Kokaram, N. Rea, and H. Denman, "Joint audio visual retrieval for tennis broadcasts," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, April 2003, vol. 3, pp. 561–564.

[7] "Sigmedia website," URL: `http://www.mee.tcd.ie/~sigmedia/research/indexing/sport_indexing.php`.

[8] J. J. Lee, J. Kim, and J. H. Kim, "Data-driven design of hmm topology for on-line handwriting recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 15, no. 1, pp. 107–121, 2001.

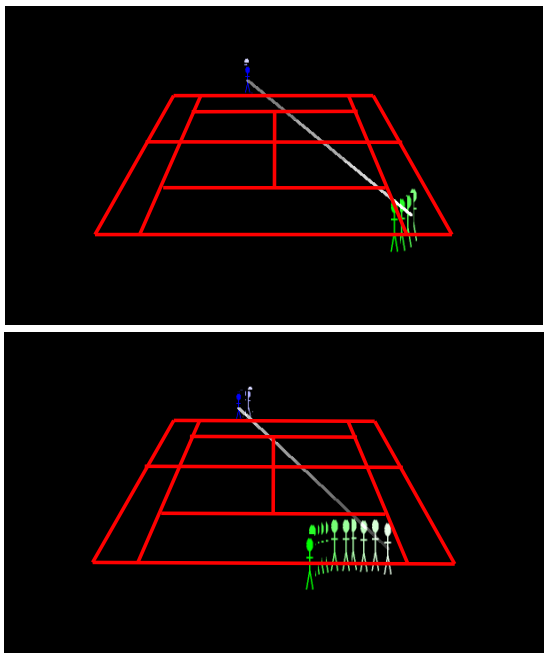

Fig. 7. Keyframes from a rally event. Blue serving (top). Green returns (bottom). Increasing saturation (player) / brightness (ball) over time.