Semantic Event Detection in Sports through Motion Understanding. *

N. Rea°, R. Dahyot†° and A. Kokaram°

° Electronic and Electrical Engineering Department,
University of Dublin, Trinity College Dublin, Ireland.
† University of Cambridge, Trumpington Street,
Cambridge CB2 1PZ, United Kingdom.
oriabhan@tcd.ie

Abstract. In this paper we investigate the retrieval of semantic events that occur in broadcast sports footage. We do so by considering the spatio-temporal behaviour of an object in the footage as being the embodiment of a particular semantic event. Broadcast snooker footage is used as an example of the sports footage for the purpose of this research. The system parses the sports video using the geometry of the content in view and classifies the footage as a particular view type. A colour based particle filter is then employed to robustly track the snooker balls, in the appropriate view, to evoke the semantics of the event. Over the duration of a player shot, the position of the white ball on the snooker table is used to model the high level semantic structure occurring in the footage. Upon collision of the white ball with another coloured ball, a separate track is instantiated allowing for the detection of pots and fouls, providing additional clues to the event in progress.

1 Introduction

Research interests in high-level content based analysis, retrieval and summarisation of video have grown in recent years [1]. A good deal of the interest has been focused on the detection of semantic events that occur in sports video footage [2, 3]. This has been fueled primarily by the commercial value of certain sports and by the demands of broadcasters for a means of speeding up, simplifying and reducing the costs of the annotation processes. Current techniques used for annotating sports video typically involve loggers manually accounting for the events taking place [1]. The existing manually derived metadata can be augmented by way of automatically derived low level content-based features such as colour, shape, motion and texture [4]. This enables queries against visual content as well as textual searches against the predefined annotations allowing for more subjective queries to be posed.

As humans operate at high levels of abstraction and since the most natural means for the lay person to query a corpus of data is through the use of semantics, it makes

* 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 MOUMIR (**MO**dels for **U**nified **M**ultimedia **I**nformation **R**etrieval).

sense to develop algorithms that understand the nature of the data in this way. In order to do so, it becomes necessary to restrict the algorithms to a unique domain. These constraints enable low-level content based features to be mapped to high-level semantics through the application of certain domain rules.

The necessity for automatic summary generation methods for sports is highlighted by the fact that the semantic value of the footage spans short durations at irregular intervals. The remainder of the footage is generally of no consequence to the archetypal viewer (i.e. views of the crowd, breaks in play). Interesting events occur intermittently, so it makes sense to parse the footage at an event level. This offers the prospect of creating meaningful summaries while eliminating superfluous activities.

A common approach used to infer semantic events in sports footage is accomplished by modeling the temporal interleaving of camera views [5]. This is typically carried out using probabilistic modeling techniques such as HMMs or NNs. This inherent temporal structure of some broadcast sports is not however, evident in snooker footage. Thus, a model based on evolving camera views can not be used for the purposes of this research. Other works use deterministic methods [6], but are limited in some regards with respect to the adaptivity of the models to changes in playing conditions. In this paper, we propose a novel approach for the detection of semantic events in sports whereby the spatio-temporal behaviour of an object is considered to be the embodiment of a semantic event. For the case of snooker, in the appropriate camera view, the white ball is tracked using a colour based particle filter [7]. Parzen windows are used to estimate the colour distribution of the ball as it is a small object relative to the rest of the image. The implementation of the particle filter allows for ball collision detection and ball pot detection. A separate ball track is instantiated upon detection of a collision and the state of the new ball can be monitored. Detection of such events augment the HMM decision process by providing a binary classifier where uncertainty is present. The evolution of the white ball position is modeled using a discrete HMM. Models are trained using six subjective human perceptions of the events in terms of their perception of the evolving position of the white ball. The footage is parsed and the important events are automatically retrieved.

2 Shot classification

Similar to other sports, the finite number of fixed camera views used in broadcast sports footage are arranged in such a way as to cause the viewer to become immersed in the game while trying to convey the excitement of the match to a mass audience. In snooker, the typical views used are those of the full-table, close-ups of the player or crowd, close-ups of the table and assorted views of the table from different angles.

For the purpose of this research we consider the most important view to be that of the full table. Analysis on 30 minutes of televised footage from three different broadcast sources shows it to occupy approximately 60% of the total coverage duration. In this view all balls and pockets on the table are visible, enabling ball tracking and pot detection. It is therefore necessary to ensure that the camera views can be classified with high precision.

Shot classification is accomplished using the method outlined in [8]. The footage is parsed at a clip level based on the geometrical content of the camera views. This approach does not require extraction of 3D scene geometry and is generic to broadcast sports footage which exhibit strong geometrical properties in terms of their playing areas. The temporal evolution of the derived feature is modeled using a first-order

discrete HMM, allowing the views to be correctly classified. The system for parsing snooker footage is illustrated in figure 1. The relevant full table shots are passed to an event processor where tracking, pot detection and foul detection are performed.

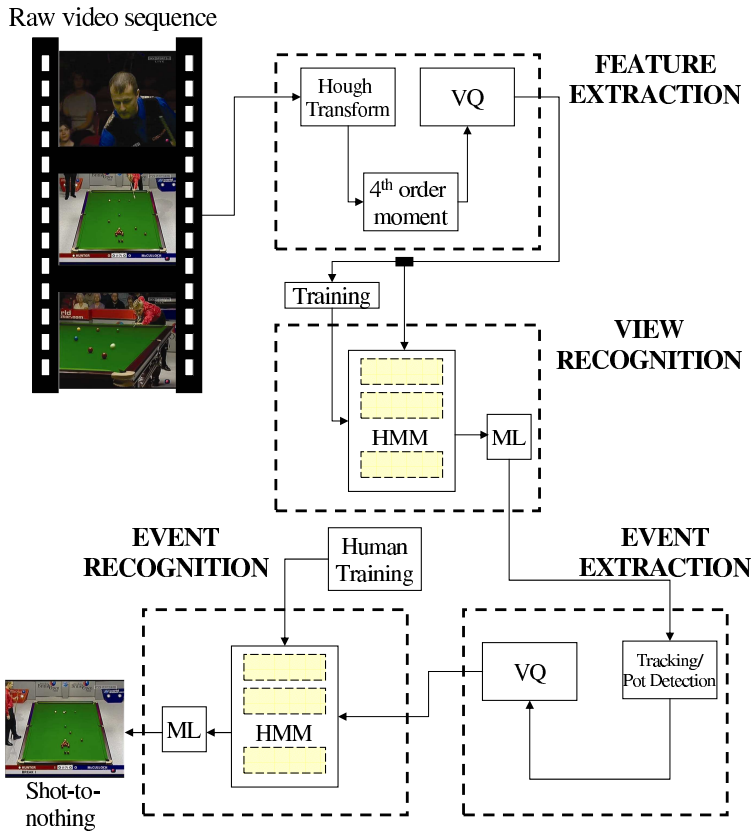

Fig. 1. System for parsing broadcast snooker footage.

3 Event classification

It was observed that the track drawn out by the white ball over the duration of a player's shot characterises an important event. If the spatio-temporal evolution of the white ball's position can be modeled, semantic episodes in the footage can be classified. We must firstly define the events of interest that occur in snooker in terms of the spatio-temporal evolution of the position of the white ball and how pots and fouls affect the event semantics.

3.1 Events of interest in snooker and game heuristics

In snooker, players compete to accumulate the highest score possible by hitting the white ball and potting the coloured balls in a particular sequence. The coloured balls vary in value from one (red) to seven (black), so different strategies must be employed to gain and maintain control of the table. The occurrence of a ball pot or foul (the white ball not colliding with a coloured ball at all) will affect the viewer's perception of the event in hand. Priori domain knowledge makes use of these events, allowing a set of heuristics to be established which are used to evaluate the current maximum likelihood classification upon detection of a foul or a pot. This is illustrated in figure 3.

The 'plays' we consider are characterised by the spatio-temporal behaviour of the white ball as follows (where C is the event number) and are affected by the state of the coloured ball (pot/no pot) and the white ball (foul/no foul).

Break-building: $C = 1$. As the player attempts to increase his score he will try and keep the white ball in the center of the table with easy access to the reds and high valued balls. If a pot has been detected, the player is attempting to build a high break ($C = 1$) (figure 2). In the unlikely event of one of the balls not being potted, the white ball will probably be in a position such that the remaining balls will be eminently 'potable'. This is called an 'open table' event ($C = 5$).

Conservative play: $C = 2$. Similar to the shot-to-nothing, except a coloured ball will not be potted when the white navigates the full length of the table. If this model is chosen as being the most likely, and a pot is detected, a shot-to-nothing will be inferred ($C = 4$). This is because the ball will be in an area where it might prove difficult for a player to pot the next coloured ball in the sequence.

Escaping a snooker: $C = 3$. If the player is snookered (no direct line of sight to a ball) he will attempt to nestle the white amongst the reds or send the white ball back to top of the table. If a pot is detected following the classification of a snooker escape, the heuristics will infer a break-building event ($C = 1$). As the only goal of the player will be to escape the snooker without conceding a foul or an open table if a ball is potted, it simply serves as a bonus.

Shot-to-nothing: $C = 4$. The white ball is hit from the top of the table, traverses the table, and returns back to the top of the table. If a pot is detected, the pot heuristics will infer a shot-to-nothing ($C = 4$) (figure 2). If there is no pot, the spatio-temporal evolution of the white ball position will show that the player is attempting to return the white ball to the top of the table. A conservative play event, ($C = 2$), could therefore be inferred as he is making the next shot as difficult as possible for his opponent.

In all of these cases a foul by the white, flagged by a non-instantiated second track, or if the white is potted will result in a foul ($C = 6$) being inferred. Play will then be transfered to the opposing player.

It was also observed that a snooker escape event is characterised by a cut from the full-table view to a close up view of the ball about to be hit. This occurs while the white ball is still in motion. If the velocity of the white ball, $V > 0$, a snooker escape is inferred (figure 3).

3.2 Motion extraction

The proposed approach is similar to those methods used in handwriting recognition [9]. The position of the input device in these systems is easily obtainable through a stylus/pad interface. In the case of snooker however, the exact position of the white ball is

not so readily available. Having located the full table views in the footage [8], a robust colour based particle filter is employed in order to keep track of the position of the white ball in each frame and simultaneously track the first ball hit.

Localisation of the white ball: Events within clips are found by monitoring the motion of the white ball. As there is no camera motion in the full table view, the white is initially located by finding the brightest moving object on the table as it first starts moving. The semantic episode begins when the white ball starts moving and ends when it comes to rest. The implementation of the particle filter trivialises the accretion of these velocity values.

3.3 Ball tracking

The tracker used in this work is similar to that implemented in [7]. The objects to be tracked however are significantly smaller (approximately 100 pels in size). We use the HSV colour space for our colour based probabilistic tracker. In order to facilitate an increase in resolution by selecting a small object relative to the size of the image, the colour distribution needs to be extended for both target and candidate models. Parzen windows are used to estimate the distribution of the hue and saturation components while the luminance component is quantised to 16 bins to reduce the effect of the lighting gradient on the table.

A target model of the ball's colour distribution is created in the first frame of the clip. Advancing one frame, a set of particles is diffused around the projected ball location using a deterministic second order auto-regressive model and a stochastic Gaussian component. Histograms of regions the same size as the ball are computed using the particle positions as their centers. A Bhattacharyya distance measure is used to calculate the similarity between the candidates and the target which is in turn used to weight the sample set, $X = \left\{ \left(x_k^{(n)}, w_k^{(n)} \right) | n = 1...N \right\}$, where N is the number of particles used. The likelihood of each particle is computed:

$$w_k^{(n)} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\left(1-\sum_{j=1}^{m}\sqrt{\rho\left(x_k{}^{(n)}\right)^j \xi^j}\right)}{2\sigma^2}} \tag{1}$$

$\rho\left(x_k{}^{(n)}\right)$ is the histogram of the candidate region at position x_k for sample n, ξ is the target histogram and m is the number of histogram bins and $\sigma^2 = 0.1$.

3.4 Collision detection

A ball collision is detected by identifying changes in the the ratio between the current white ball velocity v_k and the average previous velocity v_p (defined below, where d is the frame where the white starts its motion).

$$\mathbf{v}_p = \frac{1}{(k-2)-d} \left(\sum_{i=d}^{k-2} \mathbf{v}_i \right) \tag{2}$$

If the ball is in the vicinity of the cushion, a cushion bounce is inferred and d is set to the current frame. Ratios in the x and y velocity components $v_k^x / v_p^x, v_k^y / v_p^y$

Fig. 2. Tracking and table sections. Left to right: Shot-to-nothing; Break building; Spatial segmentation of the table.

are analysed to isolate changes in different directions. A collision is inferred when the condition in equation 3 is satisfied.

$$h_k = \left\{\left(\frac{|\mathbf{v}_k^x|}{|\mathbf{v}_p^x|} < 0.5\right) \wedge \left(\frac{|\mathbf{v}_k^y|}{|\mathbf{v}_p^y|} > 0.5\right)\right\} \vee \left\{\left(\frac{|\mathbf{v}_k^y|}{|\mathbf{v}_p^y|} < 0.5\right) \wedge \left(\frac{|\mathbf{v}_k^x|}{|\mathbf{v}_p^x|} > 0.5\right)\right\} \vee \left\{\left(\frac{|\mathbf{v}_k^y|}{|\mathbf{v}_p^y|} < 0.5\right) \wedge \left(\frac{|\mathbf{v}_k^x|}{|\mathbf{v}_p^x|} < 0.5\right)\right\} \tag{3}$$

The condition therefore flags an event when velocity changes by 50%. The form of the decision arises because the physics of colliding bodies implies that at collision, changes in velocity in one direction are typically larger than another except in the case of a 'flush' collision where a reduction of $< 50\%$ in both directions is exhibited.

Pot detection: Distinguishing between correct tracking and the loss of a 'lock' can be accomplished by using a threshold on the sum of the sample likelihoods, L_l. If the cumulative likelihood at time k, $L^k > L_l$ a correct lock is assumed, and the ball has been found. If $L^k/L^{k-1} < 0.5$, the ball has been potted.

3.5 Spatial encoding of the table

The dimensions of the table, the positions of the balls and their values dictate the flow of the play to be mostly along the long side of the table (or without loss of generality, along the vertical). The temporal behaviour of the vertical position of the white alone could therefore be considered to embody a semantic event. Using the fact that diagonals of a trapezoid intersect at its center, the table can be divided into 5 sections at the coloured ball's spot intervals (figure 2). Initially, the table is divided by intersecting the main diagonals, retrieving the center line. Sub division of the two resulting sections retrieves the pink and brown lines, and so on. The starting and end positions of the white ball alone do not sufficiently represent a semantic event. The model must be augmented by the dynamic behaviour of the white ball. The observation sequence, O, is the sequence of evolving table sections.

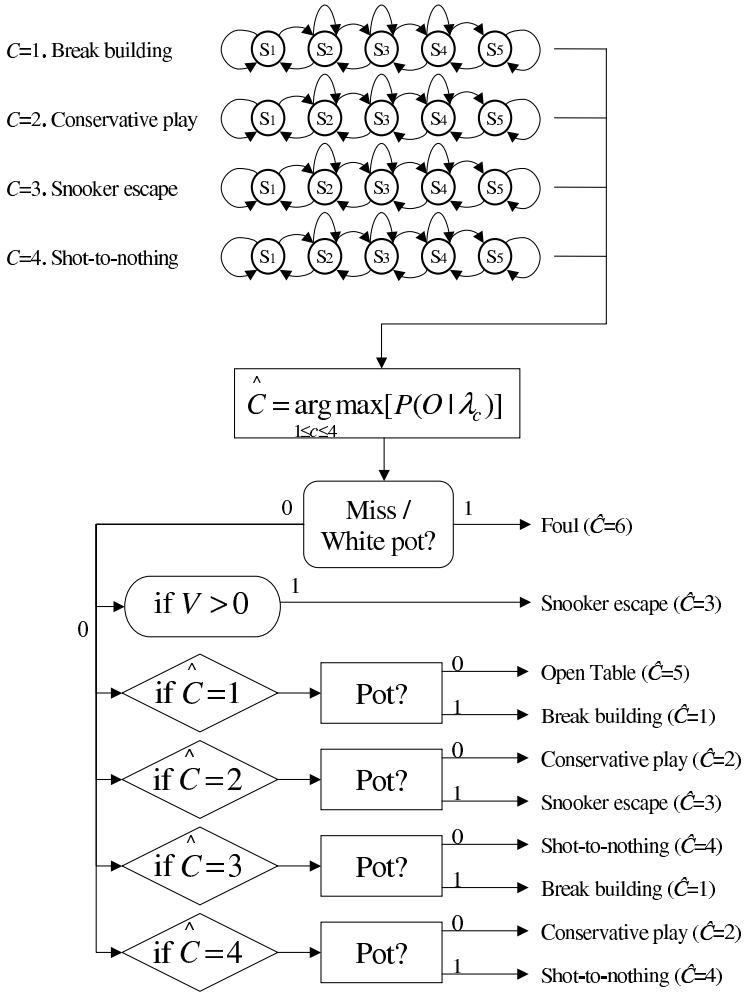

Fig. 3. Event HMMs with pot and foul classifiers.

3.6 Establishing the model topology

Modeling the temporal behaviour of the white ball in snooker is accomplished using a first order HMM. HMMs have been shown to be one of the most efficient tools for processing dynamic time-varying patterns and allow a rich variety of temporal behaviours to be modeled. The model topology is derived from the observations, reflecting the nature of the target patterns. A left-to-right/right-to-left topology is chosen to model the motion of the white ball for each event, revealing the structure of the events in state form. Each section is represented

by a state in the HMM where the state self-transitions introduce time invariance as the ball may spend more than one time-step in any one section.

Knowing the number of states (or sections of the table), $N = 5$, and discrete codebook entries, $M = 5$, a model λ, can be defined for each of the competing events. A succinct definition of a HMM is given by $\lambda_c = (A_c, B_c, \pi_c)$, where c is event label. The model parameters are defined as: A, the state transition probability matrix, B, the observation probability matrix, and π a vector of initial state probabilities.

The Baum-Welch algorithm is used to find the maximum likelihood model parameters that best fit the training data. As the semantic events are well understood in terms of the geometrical layout of the table, the models can be trained using human understanding. Six separate human perceptions of the events listed in section 3.1 were formed in terms of the temporally evolving table coding sequence of the white ball. The models used are shown in figure 3 with an example of a single training sequence. The models are initialised by setting $\pi^n = 1$ where $n = O_1$.

Each semantic episode can then be classified by finding the model that results in the greatest likelihood of occurring according to equation 4.

$$\hat{C} = \arg \max_{1 \leq c \leq \mathcal{C}} [P(O|\lambda_c)], \qquad \mathcal{C} = 4 \text{ events.} \tag{4}$$

4 Results

Experiments were conducted on two footage sources (*F1*,*F2*) from different broadcasters of 17.5 and 23.2 minutes in duration. 21 occurrences of the events to be classified were recognised in *F1*, of which 11 were break-building, 6 were conservative plays, 2 shot-to-nothings, 1 open-table, 0 snooker escapes and 1 foul. 30 events occurred in *F2* of which there were 16 break-building, 8 conservative plays, 1 shot-to-nothing, 2 open tables, 2 snooker escapes and 1 foul. The classification results 1 are assessed by computing the recall (R) and the precision (P).

$$R = \frac{A}{A+C} \quad P = \frac{A}{A+B} \tag{5}$$

A is the number of correctly retrieved events, B the number of falsely retrieved events and C the number of missed events.

In *F1* the only misclassification was that of a shot-to-nothing being classified as a break building event. In *F2* a problem arose in the classification of two conservative plays. One was misclassified as a foul due to light contact being made by the white with a coloured ball and a collision was not detected, while the second was misclassified as an open table event.

5 Discussion

In this paper we have considered the dynamic behaviour of an object in a sport as being the embodiment of semantic episodes in the game. Modeling the temporal

Event type	*F1* (P)	*F1* (R)	*F2* (P)	*F2* (R)
Break-building ($C = 1$)	91.67%	100%	94.12%	100%
Conservative play ($C = 2$)	100%	100%	100%	75%
Snooker escape ($C = 3$)	N/A	N/A	100%	100%
Shot-to-nothing ($C = 4$)	100%	50%	100%	100%
Open Table ($C = 5$)	100%	100%	66%	100%
Foul ($C = 6$)	100%	100%	50%	100%

Table 1. Event classification results.

evolution of the low level feature in this way allows important episodes to be automatically extricated from the footage. Results obtained are promising using the most relevant 60% of footage. Augmenting the feature set with more tracking information could improve the retrieval further. We are currently attempting to use the same process to classify semantic episodes that occur in broadcast tennis footage. Furthermore, we are investigating the possibility of generating game summaries where the excitement of each match could be gauged by the frequency of different events.

References

1. Bertini, M., Bimbo, A.D., Nunziati, W.: Semantic annotation for live and posterity logging of video documents. In: Visual Communications and Image Processing (VCIP 2003). (2003)
2. Kijak, E., Gros, P., Oisel, L.: Temporal structure analysis of broadcast tennis video using hidden markov models. In: SPIE Storage and Retrieval for Media Databases. (2003) 289–299
3. Assfalg, J., Bertini, M., Bimbo, A.D., Nunziati, W., Pala, P.: Soccer highlight detection and recognition using hmms. In: IEEE International Conference on Multimedia and Expo. (2002)
4. Djeraba, C.: Content-based multimedia indexing and retrieval. IEEE Multimedia **9** (2002) 52–60
5. Chang, P., Han, M., Gong, Y.: Extract highlights from baseball game video with hidden markov models. In: Proceedings of the International Conference on Image Processing (ICIP '02). (2002)
6. Ekin, A., Tekalp, A.M.: Automatic soccer video analysis and summarization. In: International Conference on Electronic Imaging: Storage and Retrieval for Media Databases. (2003) 339–350
7. Perez, P., Hue, C., Vermaak, J., Gangnet, M.: Colour based probabilistic tracking. In: European Conference on Computer Vision 2002 (ECCV 2002). (2002)
8. Denman, H., Rea, N., Kokaram, A.C.: Content based analysis for video from snooker broadcasts. Journal of Computer Vision and Image Understanding (CVIU): Special Issue on Video Retrieval and Summarization **92** (2003) 141–306
9. Lee, J.J., Kim, J., Kim, J.H.: Data-driven design of hmm topology for on-line handwriting recognition. In: The 7th International Workshop on Frontiers in Handwriting Recognition. (2000)