MODELING HIGH LEVEL STRUCTURE IN SPORTS WITH MOTION DRIVEN HMMS

Niall Rea, Rozenn Dahyot and Anil Kokaram.

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

ABSTRACT

In this paper we investigate the retrieval of dynamic events that occur in broadcast sports footage. Dynamic events in sports are important in so far as they are related to the game semantics. Thus far, the temporal interleaving of camera views has been used to infer these types of events. We propose the use of the spatio-temporal behaviour of an object in the footage as an embodiment of a semantic event. This is accomplished by modeling the evolution of the position of the object with a Hidden Markov Model (HMM). Snooker is used as an example for the purpose of this research. The system firstly parses the video sequence based on the geometry of the content in the camera view and classifies the footage as a particular view type. Secondly, we consider the relative position of the white ball on the snooker table over the duration of a clip to embody semantic events. A colour based particle filter is employed to robustly track the snooker balls. The temporal behaviour of the white ball is modeled using a HMM where each model is representative of a particular semantic episode. Upon collision of the white ball with another coloured ball, a separate track is instantiated.

1. INTRODUCTION

Retrieval and summarisation of sport events have received increasing interest in recent years. This has been motivated by the commercial value of certain sports and due to the demands by broadcasters for systems that ease the burden of annotating vast quantities of live and archived sports video material. Thus far, these systems have typically relied on high-cost, manually obtained annotations in the form of captions, embedded text and transcripts. Automatically derived content-based features have recently been used to supplement the existing metadata in some of these systems [1].

Retrieval is a non-trivial task in general and is made even more difficult by the so-called semantic gap between machine and client. The most natural way for a user to query a corpus is by way of semantics. Therefore, in order to create a successful semantic based retrieval system it is necessary to understand the user context. This can be accomplished by restricting the domain being addressed. Low level content features can then be mapped to high-level semantics by applying certain domain rules and constraints.

As important events in sports typically occupy short time periods, it is logical to parse the footage at an event level. This offers the prospect of creating meaningful summaries while eliminating superfluous activities. This could prove beneficial for sport services on low-bandwidth devices.

A common approach used to infer semantic events in sports footage is to model the temporal interleaving of camera views inherent in the footage with a HMM [2]. Televised snooker coverage, however, does not provide sufficient freedom to use such a model based on camera views. In this paper, we propose a novel approach for semantic event recognition in sports whereby the spatio-temporal behaviour of the white ball is considered to be the embodiment of a semantic event. In the appropriate camera view, the white ball is tracked using a colour based particle filter [3]. 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 and pot detection. A separate ball track is instantiated upon detection of a collision and the state of the new ball can be monitored. The evolution of the white ball position is modeled with a discrete HMM. Models for each event are trained using six different subjective human perceptions of the events in terms of the evolving position of the white ball. The footage is parsed and the important events are automatically retrieved.

2. SYSTEM FOR PARSING SNOOKER FOOTAGE

Broadcast snooker footage exhibits many similar characteristics to most other televised sports. The finite number of fixed camera views are arranged in such a way as to cause the viewer to become immersed in the footage while trying to convey the excitement of the game to a mass audience. The typical views used are those of the full-table, close-ups of the player or crowd, close-ups of the table and an assortment of views of the table from various angles.

The most important view in the footage can be considered to be that of the full table. Analysis on 30 minutes of

Work sponsored by Enterprise Ireland Project MUSE-DTV (**M**achine **U**nderstanding of **S**ports **E**vents for **D**igital **T**elevision).

televised footage shows it to occupy approximately 60% of the total coverage. This overview of the entire table allows balls to be tracked and pots to be detected. It is therefore necessary to ensure that the camera views can be classified with high precision.

2.1. Overview of the system

In [4] we proposed a novel technique for parsing snooker footage 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 televised sports which exhibit strong geometrical properties in terms of their playing areas. An outline of the system's feature extraction and view recognition modules are shown in figure 1. Correctly labeled full table views are passed to an event processor where tracking and pot detection are conducted.

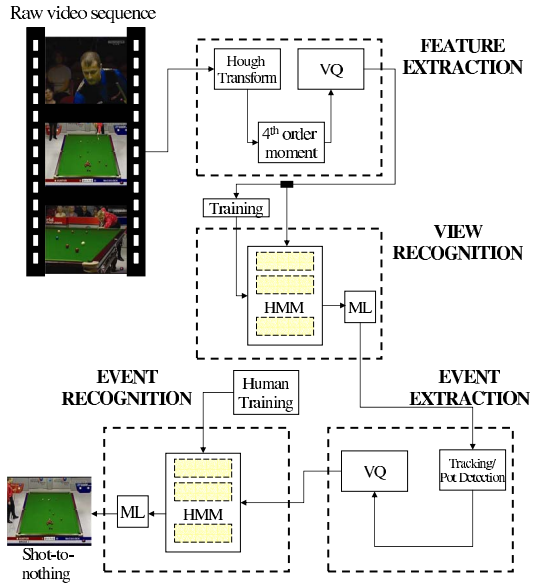

Fig. 1. System for parsing broadcast snooker footage.

2.2. Events of interest 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 'plays' we consider are characterised by the spatio-temporal behaviour of the white ball.

Shot-to-nothing: The white ball is hit from the top of the table, traverses the table, a colour is potted, and returns back to the top. Figure 2 (a) illustrates this type of shot.

Break building: 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. Figure 2 (b) illustrates this type of shot.

Conservative play: Similar to the shot-to-nothing, except a coloured ball will not be potted when the white navigates the full length of the table.

Snooker escape: 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.

In the case of all these events a '**miss**' could be flagged if a collision is not detected.

3. MOTION EXTRACTION

It was observed that the track drawn out by the white ball over a shot characterises an important event. If the temporal evolution of the white ball position can be modeled, semantic episodes in the footage can be classified.

The proposed approach is similar to those methods used in handwriting recognition [5]. 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 [4], 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. It is 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 a rest. This information is available directly from the ball tracker.

3.1. Ball tracking

The tracker used in this work is similar to that implemented in [3]. The objects to be tracked however are significantly smaller (approximately 100 pels in size). 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. Computation of the distribution in this way is expensive but does produce significantly better results for smaller objects.

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:

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

$\rho\left(x_k{}^{(n)}\right)$ 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 $\sigma^2 = 0.1$.

Tracking performance is good [4]. Some obvious problems are experienced however. For example, when a 'flush' collision between two balls of the same colour occurs, the ball in advance will be tracked. This could be remedied by applying the algorithm outlined in 3.2 to all balls. If a ball is near a pocket and the player walks in front of the camera blocking the ball from view, a pot will be detected. Application of ad-hoc rules will prevent such incorrect inferences.

(a) Shot-to-nothing. (b) Break building.

Fig. 2. Example of tracking results.

3.2. 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.).

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

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 $v_k^x/v_p^x, v_k^y/v_p^y$ are examined to isolate changes in different directions. A collision is therefore detected when the following condition is satisfied.

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

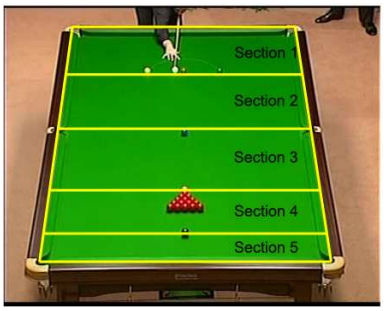

Fig. 3. Spatial segmentation of the table into 5 sections.

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 exhibited (last condition).

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.

4. MOTION UNDERSTANDING USING HMM

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. The temporal behaviour of the vertical position of the white alone could therefore be considered to exemplify 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 3). 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 ball. The observation sequence, O, is therefore the sequence of evolving table sections.

4.1. Classification of events

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. 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 2.2 were formed in terms of the temporally evolving table coding sequence of the white ball. The models used are shown in figure 4 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.

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

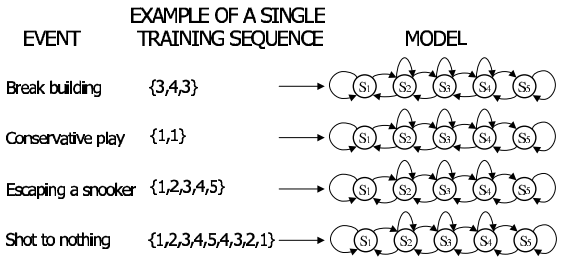

Fig. 4. Models of the events and an example of training.

5. RESULTS

Experiments were conducted on two footage sources (*F1*,*F2*) from different broadcasters of 16.2 and 21.5 minutes in duration. 20 occurrences of the events to be classified were recognised in *F1*, of which 11 were break-building, 6 conservative, 1 miss, 2 shot-to-nothing and 0 snooker escapes. 27 events occurred in *F2* of which 16 were break-building, 8 conservative, 1 shot to nothing and 2 snooker escapes. Results of the event classification are shown in table 1 where P and R are precision and recall. A problem arose in the classification of a conservative play as a miss in *F2*. Light contact was made by the white with another ball and a collision was not detected, inferring a miss. Both snooker escapes in *F2* were classified as conservative plays, the ball speed could however be used to improve this retrieval. Two shot-to-nothings in *F1* were classified as conservative plays but using the fact that the ball was potted, the correct event was inferred.

Event type	*F1* (P)	*F1* (R)	*F2* (P)	*F2* (R)
Shot-to-nothing	100%	100%	100%	100%
Break-building	100%	90.9%	100%	100%
Snooker escape	N/A	N/A	0%	0%
Cons. play	85.7%	100%	75%	75%
Miss	100%	100%	N/A	N/A

Table 1. Event classification results

6. FUTURE WORK

In this paper we considered the dynamic behaviour of an object in a sport to be the embodiment of events in the game. Modeling this low level feature allows important episodes to be automatically retrieved from the footage. Results obtained are promising using the most relevant 60% of footage. Future research will involve extracting features from the remaining footage and mapping them to high level concepts. Augmenting the feature set with more tracking information could improve the retrieval further. We are also investigating the possibility of generating game summaries where the excitement of each match could be gauged by the frequency of the different events.

7. REFERENCES

[1] C. Djeraba, "Content-based multimedia indexing and retrieval," *IEEE Multimedia*, vol. 9, no. 2, pp. 52–60, 2002.

[2] 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)*, 2002.

[3] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Colour based probabilistic tracking," in *European Conference on Computer Vision 2002 (ECCV 2002)*, 2002.

[4] H. Denman, N. Rea, and A. C. Kokaram, "Content based analysis for video from snooker broadcasts," *Journal of Computer Vision and Image Understanding (CVIU): Special Issue on Video Retrieval and Summarization*, 2003.

[5] J. J. Lee, J. Kim, and J. H. Kim, "Data-driven design of hmm topology for on-line handwriting recognition," in *The 7th International Workshop on Frontiers in Handwriting Recognition*, 2000.