

EVENT DETECTION BASED ON GENERIC CHARACTERISTICS OF FIELD-SPORTS

David Sadlier, Noel E. O'Connor

Centre for Digital Video Processing, Dublin City University, Ireland.

ABSTRACT

In this paper we propose a generic framework for event detection in broadcast video of multiple different field-sports. Features indicating significant events are selected, and robust detectors built. These features are rooted in generic characteristics common to all genres of field-sports. The evidence gathered by the feature detectors is combined by means of a support vector machine, which infers the occurrence of an event based on a model generated during a training phase. The system is tested across multiple genres of field-sports including soccer, rugby, hockey and Gaelic football and the results suggest that high event retrieval and content rejection statistics are achievable.

1. INTRODUCTION

Sports video analysis has received much attention in the area of digital video processing. Existing approaches can be broadly classified into two distinct categories – genre specific and genre independent analyses. Due to the dramatic variances in broadcast styles for different sports genres, much of the prior art concerns genre specific approaches. Genres that have been targeted include soccer [1-3], basketball [4], Formula-1 [5], baseball [6], cricket [7] and tennis [8]. All these works show that genre specific approaches typically yield successful results within the targeted domain.

However, central to all these works are complex algorithms, performing standalone modeling of specific events, based on intrinsically critical characteristic features. This limits their applicability to other sports. Previously published approaches that are generic in nature include [9] and [10], and in [11] and [12] achievements are actually shown to operate quite reasonably across multiple genres.

It is impractical to attempt to develop a unique solution that operates successfully across all genres of sports video. Thus in our approach we limit our scope to some extent, whilst not becoming too content specific. Our chosen domain is *field-sport* broadcasts, encompassing all sports genres that fall within this classification, e.g. soccer, American football, rugby, Australian rules, field hockey, Gaelic football etc. The motivation is that these sports all share common characteristics, which may be generically exploited in the analysis. These include:

- (i) Two opposing teams + referee(s)
- (ii) Enclosed playing area
- (iii) Grass pitch
- (iv) Field lines
- (v) Commentator voice-over

- (vi) Spectator cheering
- (vii) On-screen video graphic (scoreboard)
- (viii) Three well-defined styles of camera shot: global (main), zoom-in and extreme close-up
- (ix) Objectives concerned with territorial advancement and directing an object (e.g. ball) towards a specific target

We define an *event* as constituting a successful scoring incident e.g. a *goal* in soccer or hockey, a *try* or *conversion* in rugby, a *point* in Gaelic football etc. Event defining feature detectors are designed motivated by the common characteristics above. To ensure generality in the design of the proposed approach, a variety of field-sport video content was captured from various broadcasts sources. In all, 100-hours of content was captured in MPEG-1 format, which was subsequently divided into a training corpus (for developing model hypotheses), and a test corpus (for evaluation). Both were manually annotated in order to create a *ground truth*.

2. FIELD-SPORT EVENT CHARACTERISTICS

A subjective examination of the multi-genre training corpus indicated that 98% of all events are followed by an action replay. A further consistent feature is the lag time that immediately follows the occurrence of an event, before the cut to action replay. The director utilizes this *reaction-phase* to capture the responses of players and/or crowds to the significance of the event. Moreover, during the reaction-phase, the characteristics of the content typically include; (i) a close-up shot of the player(s) and/or relevant parties involved; (ii) a camera shot showing the crowd celebrating; (iii) an increase in audio activity (particularly in the voice band frequency range, corresponding to commentator vocals); (iv) activity in the on-screen graphics (scoreboard); and (v) a surge in near-field motion activity (NFMA) as the camera attempts to capture the intense celebratory behaviour of the scorer. From the data, very few reaction-phase durations are in excess of 25s, and this was adopted as a post-event critical-see-window (CSW). Based on an evaluation of the training data, the occurrences of these characteristics within the CSW are shown in Table-1.

The evidence also suggests that for a given event (e.g. *goal*, *try*, etc.), it is typical for the action to be situated at the end-zone

Table-1. Characteristics contained in CSWs

Characteristic	Occurrence
Close-up sequence	97%
Crowd sequence	73%
Audio level in excess of broadcast mean level	84%
Removal of scoreboard graphic during update	61%
NFMA in excess of broadcast mean level	76%

region of the playing field - 73% of all training corpus events conformed to this constraint.

3. CONTENT PRE-PROCESSING

It is desirable to incorporate a pre-processing phase, where the content is initially segmented and clearly irrelevant periods are initially rejected, prior to subsequent analysis.

3.1. Shot Boundary Detection

Hard shot-cut boundary detection is effectively a solved problem. In the training corpus, 94% of all boundaries are hard cuts and a standard histogram based approach [13] was employed to detect these.

3.2. Close-up Image Detection & Shot Filtering

Close-up image detection is performed by exploiting the two most salient characteristics of such images; (i) the presence of a face in the top-middle-centre region (i.e. the focus) of the frame, together with (ii) a torso in the bottom-middle region of the frame, occluding an arbitrary background – see Fig. 1; Image A1. Although colour-based object recognition may not be practical in many video scenarios, it is suitable for sports video, where colours are purposely used to differentiate players, and clearly defined rules constrain the action. As a result the colours of the playing surface and the players/referee shirts usually consist of one or two (striped) dominant colours. Our close-up detection model is based on the degree to which the image has (i) a skin-toned entity (i.e. a face) in a specific region of expectation of the frame, and (ii) a dominant coloured entity (i.e. a shirt) in another region of expectation. Regions of expectations were empirically defined from training corpus data.

As outlined in Section 2, it was estimated that 97% of all training corpus events included a close-up image sequence during the reaction-phase. Thus it was decided to employ this feature as the basis for a shot-retention condition. The stipulation requires that for a given shot to be retained, it must be followed by an instance of a close-up (I-frame) image within the post-event CSW.

4. FEATURE DETECTORS

4.1. Feature Detector 1: Crowd Image Detection

Crowd image detection is achieved by exploiting the fact that they are inherently uniform high-frequency images. In field-sports video content, the majority of images capture relatively sizeable monochromatic, homogeneous regions e.g. grassy pitch, player's shirt. Crowd image detection is achieved by exploiting the characteristic that, in the context of such a non-complex image environment, such images are relatively detailed – see Fig. 1. A discrimination between detailed and non-detailed pixel blocks is made by examining the number of non-zero frequency (AC) Discrete Cosine Transform (DCT) coefficients used to represent the data in the frequency domain. On this basis a feature set of I-frame crowd image confidence values $\{Fv_1\}$ is calculated:

$$\{Fv_1\} = [\text{Crowd Image Confidence}]_{\text{I-Frames}}$$



Fig. 1. Field-sports video images: A1: Close-up image; A2: Zoom-in; A3: Global perspective; B1-B3: Crowd Images.

4.2. Feature Detector 2: Speech-Band Audio Activity

A comprehensive discourse on how audio speech band energy may be estimated by examining MPEG bitstream scalefactor weights may be found via [14]. On this basis a feature set $\{Fv_2\}$ of audio speech band energy levels, estimated at 0.5s intervals, is yielded.

$$\{Fv_2\} = [\text{Audio Speech-Band Energy}]_{0.5s}$$

4.3. Feature Detector 3: On-Screen Graphics Tracking

4.3.1. Scoreboard Graphic Frame Location

For scoreboard text to be visible, there must be a strong luminance contrast between the foreground and background, a characteristic, which, during encoding, will demand an ample amount of AC-DCT luminance coefficients. This fact is exploited to pinpoint the graphic within the frame.

4.3.2. Scoreboard Presence/Absence Tracking

For a given broadcast, the scoreboard graphic location is generally static within the frame. Furthermore over the course of a broadcast, it tends to be on-screen for the majority of the duration. Therefore for the critical pixels constituting the graphic, the temporal mode values, computed across all I-frames of the broadcast, will be representative of the graphic's general appearance. Thus for a given image, a SAD operation between its critical pixels and the mode values should provide a good indication of whether the graphic is present/absent. This operation is performed in the luminance domain for all I-frames and the outputs used as another feature set $\{Fv_3\}$.

$$\{Fv_3\} = \text{SAD} [\text{Iframe}_N, \text{Mode}]_{\text{Critical Pixels}}$$

4.4. Feature Detector 4: Motion Activity Measure

Visual motion activity for each P-frame is estimated from the evidence conveyed by the motion vectors present in the MPEG video bitstream. The approach is similar to that given in [15], which is based on counting the number of macroblocks in the frame whose motion vector length is greater than a pre-selected 'zero'-threshold. However in our approach, the 'zero'-threshold value is chosen to be large enough as to ignore slow, smooth, far-field motion, while detecting jerky, uneven, near-field motion, i.e. the type of turbulent motion expected during celebratory moments. These values constitute feature set $\{Fv_4\}$.

$$\{Fv_4\} = [\text{Near-Field Visual Activity}]_{\text{P-frame}}$$

4.5. Feature Detector 5: Field Line Orientation

For events situated in the end-zone region of the playing field, the fixed position of the camera and resulting perspective is such that the field-end lines may only assume certain angles. Empirically, these were found to mainly lie within the interval 5° - 25° , relative to the camera. To exploit this, the playing fields are first segmented in the hue space. A Roberts edge detector [16] is applied, and then the Hough Transform [17] is used to extract field lines. This operation is performed for each I-frame and the angles of the most prominent line form feature set $\{Fv_5\}$.

$$\{Fv_5\} = [\text{Angle of Most Prominent Line}]_{I\text{-frame}}$$

5. FEATURE DATA AGGREGATION AND SVM

For a given shot, corresponding feature data is aggregated into a *shot feature vector* (SFV).

$$\text{SFV} = [Vcc_1, Vcc_2, Vcc_3, Vcc_4, Vcc_5]$$

The vector should convey event-critical information, i.e. for a given SFV, it is required that the individual vector component coefficient (Vcc) values, aggregated from the feature data sets, reflect the features' contribution to the overall probability that the content of the shot exhibits an event. Specifically, within the CSW from the end-boundary of each retained shot, Vcc1 is defined as the maximum I-frame crowd image confidence value; Vcc2 is defined as the maximum speech-band energy value; Vcc3 is defined as the maximum I-frame graphic activity value; Vcc4 is defined as the maximum P-frame visual motion activity confidence value.

$$\begin{aligned} Vcc_1 &= \text{Max } \{Fv_1\}_{CSW} & Vcc_2 &= \text{Avg } \{Fv_2\}_{CSW} \\ Vcc_3 &= \text{Max } \{Fv_3\}_{CSW} & Vcc_4 &= \text{Max } \{Fv_4\}_{CSW} \end{aligned}$$

$\{Fv_5\}$ is a feature data set representing the angles of the most prominent field lines for I-frame images. To quantify the confidence that a given shot culminates with the camera focused on field end-zone activity, the field-line angles of the last 6 I-frames preceding the shot-end boundary (*critical I-frames*) are examined. As the corresponding orientations are found to lie in the key interval, the confidence value increases accordingly, i.e.

$$\begin{aligned} Vcc_5 &= i \div 6 \\ \{i\} &= \# \text{ critical I-frame orientations within the key range} \end{aligned}$$

These five-dimensional SFVs constitute the input data for the classifier. An SVM (with radial basis function kernel) was chosen as the learning algorithm for these experiments.

6. EXPERIMENTAL EVALUATION

6.1. Training & Testing Phases

The training corpus consisted of 210 events – 70 events from soccer, 70 from rugby, and 70 from Gaelic football. For each individual event, the corresponding SFVs, coupled with the ground truth, were used to train the SVM, such that a generic field-sport event model was generated, inferred from the key feature patterns. The SVM was trained for wide-varying values of the regularization parameter, C .

The test corpus consisted of 60 events from soccer video, 80 events from both rugby and Gaelic football video, and 40 events

from hockey video. Each trained classifier was run on the SFVs of the test corpus content, such that corresponding shots are deemed either *eventful* or *non-eventful*. Consecutive shots with identical classification were grouped together yielding eventful or non-eventful *scenes*.

6.2. Results & Evaluation

The scene classifications were compared to that of the ground truth. Estimations for event retrieval ratios (a true-positive statistic) and content rejection ratios (a true-negative statistic) were computed. Fig. 2 presents a plot of content rejection ratio (CRR) against event retrieval ratio (ERR), for varying values of C , with respect to the Rugby content alone. The maximum ERR value achievable is 97%, i.e. the retrieval of 97% of all events in the game i.e. *tries*, *drop-goals*, *conversions*, and *penalties*. The corresponding CRR value for this particular value of C is 38%. This maximum ERR limit, and corresponding CRR value, are a consequence of the filtering performed during the preprocessing stage – the *preprocessor limit*. By varying C during the training phase, the resulting classification varies from the preprocessor limit towards a saturation limit of 78% CRR and 68% ERR. Similar statistical analyses are performed for the other field-sport genres. Fig. 3 plots CRR Vs ERR for retrieval of *goals* in Soccer, Fig. 4 plots CRR Vs ERR for retrieval of *goals* in Hockey, and Fig. 5 plots CRR Vs ERR for retrieval of *goals* and *points* in Gaelic Football.

For cross-genre performance comparison, Table-2 lists the preprocessor limit for each genre. Also, the CRR values for a sensibly chosen ERR *evaluation point* are juxtaposed. Clearly it is desirable to maintain both the ERR and CRR statistics as high as possible. However, in a real retrieval system, high recall is paramount since a user would be more likely to tolerate the inclusion of non-events as opposed to event omissions. Thus, the evaluation point is chosen to be the maximum CRR achievable for 90% ERR. From this table it is clear that in general, the preprocessor limits are satisfactory except in the case of Hockey where, although 52% of content was rejected, only 90% of events were retained from this phase. Clearly it is desirable that the ERR of this preprocessor limit be higher. Considering the performances at the evaluation point, the best overall performing genre was Soccer for which 74% CRR may be achieved at the evaluation point of 90% ERR. The poorest performing genre at

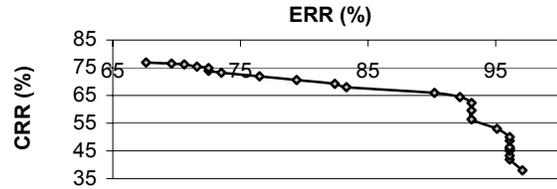


Fig. 2. Rugby content: ERR Vs CRR.

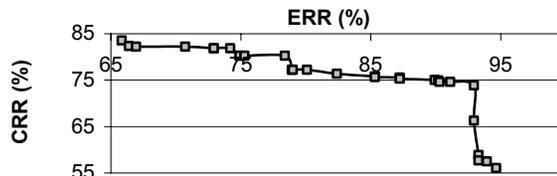


Fig. 3. Soccer content: ERR Vs CRR.

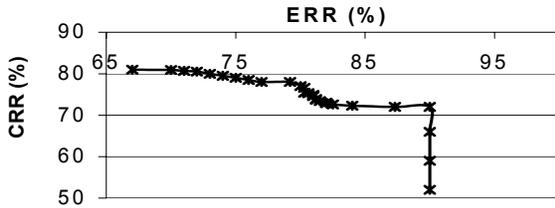


Fig. 4. Hockey content: ERR Vs CRR.

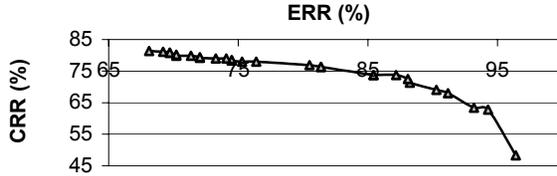


Fig. 5. Gaelic Football content: ERR Vs CRR.

the evaluation point was Rugby for which 65% CRR may be achieved. Hockey performed slightly better than Gaelic Football, with 72% CRR and 69% CRR respectively.

It is postulated that the reasons for the variance in individual genre performances may be related to the pace of the respective games. It was noted that at the evaluation point, it was the faster paced games, i.e. soccer and hockey, which outperformed the more slowly paced games of Gaelic football and rugby. Following a subjective examination of the content it was observed that the faster paced games tend to contain more live action. Therefore the video structure tends to be more defined, i.e. less play breaks. Conversely, broadcasts of a relatively slowly paced game such as Rugby, tend to include more contextual content e.g. more close-up shots, more crowd shots, more replays, more dissolves etc. i.e. a more sporadic abundance of the features critical to this analysis.

7. CONCLUSIONS & FUTURE WORK

In this paper we have outlined a generic framework for event detection in field-sports broadcast video. An event model is inferred from evidence from feature detectors, which are chosen such that they are recyclable across multiple sports genres within the field-sport domain. The techniques have been applied and tested across four distinct genres of field-sport video. A large experimental corpus, which was obtained from multiple broadcast sources, was utilized for the analysis. Compared to a manually annotated ground truth, it has been shown that both high event retrieval and content rejection statistics are achievable. It has further been described how the SVM can be tuned such that the classification may be biased to any point on

Table 2. ERR/CRR Preprocessor Limits & CRR Values at Evaluation Point

	Preprocessor Limits		Max CRR for 90% ERR
	ERR	CRR	
Rugby	97%	38%	65%
Soccer	95%	56%	74%
Hockey	90%	52%	72%
Gaelic F.	96%	47%	69%

the classification characteristic of the model. Future work will focus on quantifying the individual contribution of each feature, and an analysis of the effect of feature threshold selection on overall system performance.

8. REFERENCES

- [1] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala, "Soccer highlights detection and recognition using HMM's," Proc. *IEEE ICME 2002*, Lausanne, Switzerland.
- [2] A. Ekin, A.M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Processing*, June 2003.
- [3] R. Cabasson and A. Divakaran, "Automatic extraction of soccer video highlights using a combination of motion and audio features," in *Symp. Electronic Imaging: Science and Technology: Storage and Retrieval for Media Databases*, Jan. 2002, vol. 5021, pp. 272-276.
- [4] S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic detection of goal segments in basketball videos," Proc. *ACM Multimedia*, 2001, pp. 261-269.
- [5] M. Petkovic, V. Mihajlovic, M. Jonker, and S. Djordjevic-Kajan, "Multi-modal extraction of highlights from TV formula 1 programs," Proc. *IEEE ICME 2002*, Lausanne, Switzerland.
- [6] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden Markov models," Proc. *IEEE ICIP*, 2002.
- [7] M. Lazarescu, S. Venkatesh and G. West, "On the automatic indexing of cricket using camera motion parameters," Proc. *IEEE ICME 2002*.
- [8] R. Dayhot, A. Kokaram, and N. Rea, "Joint audio-visual retrieval for tennis broadcasts," in Proc. *IEEE ICASSP 2003*.
- [9] A. Hanjalic, "Generic approach to highlights extraction from a sport video," Proc. *IEEE ICIP 2003*.
- [10] C. Jianyun, L. Yunhao, L. Songyang, W. Lingda, "A Unified Framework for Semantic Content Analysis in Sports Video," Proc. *2nd International Conference on Information Technology for Application (ICITA 2004)*.
- [11] J. Assfalg, M. Bertini, C. Colombo and A. Del Bimbo, "Semantic annotation of sports videos," Proc. *IEEE Multimedia*, Vol. 9, No 2, pp 52-60, 2002.
- [12] Z. Xiong, R. Radhakrishnan, A. Divakaran, "Generation of sports highlights using motion activity in combination with a common audio feature extraction framework," Proc. *IEEE ICIP 2003*.
- [13] C. O'Toole, A. Smeaton, N. Murphy, S. Marlow, "Evaluation of Shot Boundary Detection on a Large Video Test Suite," Proc. *Challenges in Image Retrieval*, Newcastle (UK), February 1999.
- [14] D.A. Sadlier, S. Marlow, N. O'Connor and N. Murphy, "MPEG audio bitstream processing towards the automatic generation of sports programme summaries," Proc. *IEEE ICME 2002*, Lausanne, Switzerland.
- [15] X. Sun, B.S. Manjunath and A. Divakaran, "Representation of motion activity in hierarchical levels for video indexing and filtering," Proc. *IEEE ICIP*, New York, USA, 2002.
- [16] L. Roberts, "Machine Perception of 3-D Solids," *Optical and Electro-optical Information Processing*, MIT Press, 1965.
- [17] T. Risse, "Hough Transform for Line Recognition," Proc. *Computer Vision and Image Processing*, 1989, 46, 327-345, 1989.