A Framework for Event Detection in Field-Sports Video Broadcasts based on SVM generated Audio-Visual Feature Model. Case-Study: Soccer Video.

David A. Sadlier, Noel O'Connor, Noel Murphy, Sean Marlow
Centre for Digital Video Processing, Dublin City University, Ireland

Abstract - **In this paper we propose a novel audio-visual feature-based framework, for event detection in field sports broadcast video. The system is evaluated via a case-study involving MPEG encoded soccer video. Specifically, the evidence gathered by various feature detectors is combined by means of a learning algorithm (a support vector machine), which infers the occurrence of an event, based on a model generated during a training phase, utilizing a corpus of 25 hours of content. The system is evaluated using 25 hours of separate test content. Following an evaluation of results obtained, it is shown for this case, that both high precision and recall statistics are achievable.**

Keywords: Event Detection, Field sports video, MPEG, Signal Processing, Support vector machine

I. INTRODUCTION

Due to the dramatic variances in broadcast styles for different sports genres, much of the prior art in sports video analysis is specific to particular genres. For example, methods for the automatic retrieval of semantic events from Tennis video are presented in Dayhot *et al.* [1], while Petkovic *et al.* [2] describe techniques for similar tasks in Formula-1 racing. Meanwhile, Zhou *et al.* [3] discuss methods for automatic basketball video indexing, and Rui *et al.* [4] describe methods for event detection in a baseball video context. In [5], [6], [7] and [8], the authors present techniques for the detection of semantic events, particularly goals, from soccer video.

Central to all these works are complex algorithms, performing standalone modeling of specific events, based on intrinsically critical characteristic features, particular to each sports type. Few have successfully investigated the challenge of developing a solution or scheme that reveals the common structures of multiple events across multiple genres [9,10].

The work described in this paper is termed *event detection*, and it aims to bridge this gap between the specific and wholly generic approaches to the problem. To achieve this, we limit our scope to some extent, while at the same time, we avoid becoming too content specific. Our chosen domain is *field-sport* broadcasts, encompassing all sports types that fall within this ambit. The reasoning behind this is that field-sports broadcasts (i.e. Soccer, American Football, Rugby, Australian Rules Football, Field Hockey, Gaelic Football etc.) all share common characteristics, which may be exploited in the analysis. These include (i) two opposing teams + referee, (ii) grass pitch, (iii) enclosed playing area, (iv) commentator voice-over, (v) field lines, (vi) on-screen scoreboard, (vii) spectator cheering, (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.

Ultimately we hypothesize a field-sport feature ontology, exploiting the characteristics of multi-genre commonality described above, where numerous atomic (low-level) audio-visual signal trackers rigorously mine the video content. Any number of these atomic units are then integrated in such a way that their combination constitutes a particular (high-level) feature detector. The high level features are then themselves combined and mapped to various semantic events.

Eventually, subsequent to a comprehensive ontology implementation, we aim to develop a multi-event classification system, albeit still limited to field-sport content, but which will automatically differentiate between many types of semantic events, within allowable genres, via an extensive network of low-level to high-level feature mappings.

II. EVENT MODEL

In field-sports video, it is evident that there exists many circumstances in which an *event* (i.e. *goal* in Soccer, *try* in Rugby etc.) may be manifest. However, our approach attempts not to model the individual scenarios of *events*, but rather model what is common to all situations, irrespective of circumstance, i.e. the general characteristic patterns of the audio-visual signals of the content, following their occurrence.

Empirical evidence suggests that immediately following an *event*, the characteristics of the content typically includes; (i) a

close-up shot of the player(s) and/or relevant parties involved, (ii) a surge in motion activity as the camera attempts to capture the intense celebratory behaviour of the scorer, (iii) a camera shot showing the crowd celebrating, (iv) an increase in audio activity (particularly in the voice band frequency range, corresponding to commentator vocals). Certainly, these features may occur sporadically throughout the duration of the content. However, it is proposed that when they are found occurring within closely bound localities we can deduce an increase in the probability that an *event* has occurred.

III. PRE-PROCESSING: SHOT FILTERING

It is desirable to pre-filter the shots, such that only those that can possibly contain an *event* are retained, and hence further analysed for the pattern recognition stage.

Shot boundary detection is effectively a solved problem in the area of digital video analysis, and for this purpose our own tried and tested algorithm [11] was employed.

Field-sports *events* are inherently dynamic in nature. Therefore, camera capture generally necessitates a panoramic (*global*) camera perspective. As a consequence of their distant perspective, these shots (images) tend to capture a large amount of the grassy playing field within their video frames. Furthermore, such *events* are characterized by activity particular to the *end-zone* region of the playing field. For example, goals, tries, points, touchdowns etc. are achieved either by (i) directing the ball towards a target in the field *end-zone*, or (ii) player, with ball, advancing to the *end-zone*.

It is assumed that for a given global camera perspective image, the *mode* hue value corresponds to the grass of the playing field. Coupling this with the knowledge that grass colour clusters reasonably well in the hue space (60°-100°) [12], allows the accurate segmentation of grass regions from the images. On this basis, grass-coloured pixel ratio (GCPR) values are computed for all I-frames. Averaging between shot boundaries then yields a mean GCPR value for each shot. These values form the basis for the detection of *global* perspective shots.

Due to the fixed position of the camera in *global* shots, the resulting perspective is such that, during an *event*, the principal field lines, relative to the point of observation, may only assume certain angles, which lie within a particular narrow interval (determined by experiment). Thus, once the global shots are isolated, the field lines within the detected grass regions are extracted (via the Hough Transform [13]), and their angle of orientation is continuously tracked. Subsequently, only *global* shots, which meet the required field-line angle criterion, were retained for further analysis

IV. FEATURE DETECTOR DESIGN

Where possible, the feature detectors are designed such that they operate on data extracted directly from the compressed domain audio-video bitstream.

A. Feature Detector 1: Close-up Image Detection

It is proposed that close-up image detection may be performed by detecting (i) the presence of skin-coloured pixels (i.e. player's face), and/or (ii) the occlusion of a grass-coloured pixel background by a single, homogeneous, monochromatic region (i.e. player's torso)

Our field-sport close-up image model is defined by the weighted prevalence of two attributes. The first is a concentrated presence of skin-coloured pixels (i.e. player's face) within the top-middle-centre region (i.e. the focus) of the frame. The second is the presence of a single, homogeneous, monochromatic region occluding the (assumed) densely populated grass-coloured pixels of the bottom-middle region of the background (i.e. player's torso) - see Fig. 1, block-A, image-1. It has been shown that both grass-colour and skin-colour cluster well in the hue space ([60°-100°] and [10°-55°] respectively), and therefore may be easily discriminated from other colours in the images [12], [14].

The focus of the analysis for each I-frame is on two regions of interest, corresponding to the expected position of the face and torso - see Fig. 1, block-A, image-1. The first (R1) corresponds to the top-middle-centre region (the focus) of the image. It is within this region where we attempt to detect skin-toned pixels, corresponding to a face. For this region, the (normalised) skin coloured pixel ratio (SCPR) is calculated. This value represents the number of skin-coloured pixels, per total number of pixels, for the region. The second region of interest (R2) corresponds to the bottom-middle section of the image. This region itself is again divided into two sub-regions corresponding to positional expectation of torso (R2A) and background (R2B). Within these regions the (normalised) grass-coloured pixel ratios (GCPRs) are calculated. These values represent the number of grass-toned pixels, per total number of pixels per sub-region. A close-up confidence feature set, $\{Fv_1\}$, is then calculated as:-

$$\{Fv_1\} = SCPR_{R1} * (GCPR_{R2B} - GCPR_{R2A})$$

B. Feature Detector 2: Motion Activity Measure

It is proposed that visual motion activity may be estimated from the evidence conveyed by the motion vectors present in

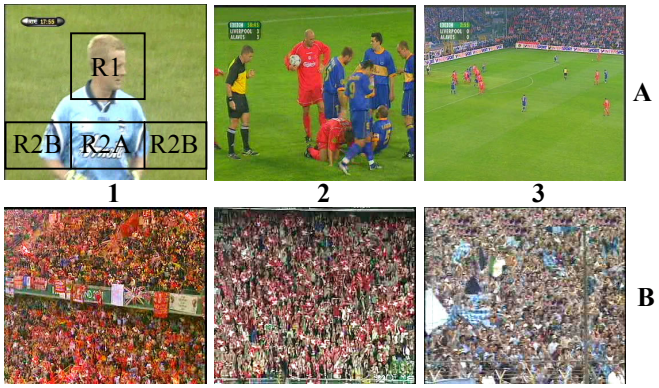

Fig. 1. Field-sports video images. Block-A: Generic images from standard camera perspectives. Block-B: Crowd Images

the MPEG video bitstream. From the video content every P-frame is extracted and from these images, motion vectors for each macroblock are extracted directly from the encoded bitstream. From the motion vectors of each P-frame image, two different statistics are calculated: The (normalised) non-zero vector value (NZVV) is calculated by counting up the number of macroblocks in the frame whose motion vector length is greater than a pre-selected threshold.

Secondly, the (normalised) mean overall length value (MOLV) is calculated by an averaged superposition of all the motion vectors in the frame.

The two statistics are calculated for each P-frame, and a visual activity feature set, $\{Fv_2\}$, is calculated by averaging the associated NZVV and MOLV values.

$$\{Fv_2\} = \text{Avg (NZVV, MOLV)}$$

C. Feature Detector 3: Crowd Image Detection

It is proposed that crowd image detection may be performed by exploiting the inherent characteristic that, in the context of a typically non-complex image environment, such images are relatively detailed - see Fig. 1.

It is proposed that discrimination between detailed and non-detailed pixel blocks may be made by examining the number of non-zero frequency (AC) Discrete Cosine Transform (DCT) coefficients used to represent the data in the frequency domain. It may be assumed that an (8x8) pixel block, which is represented by very few AC-DCT uniform coefficients, contains spatially consistent, non-detailed data. Whereas, a block which requires a considerable amount of AC-DCT coefficients for it's representation, may be assumed to consist of relatively more detailed information.

In field-sports video content, the majority of images capture relatively sizeable monochromatic, homogeneous regions e.g. grassy pitch, player's shirt - see Fig. 1, block-A. Therefore, in the context of this limited environment, it is proposed that crowd images may be isolated by simply detecting such uniformly, very high frequency images.

Each I-frame is divided into four quadrants. For each quadrant of each image, the AC-DCT coefficients of every (8x8) luminance pixel block are analysed. If the number of coefficients used to encode such blocks is greater than a pre-selected threshold, it can be deduced that the block represents reasonably complex data, and is counted - obtaining an overall value representing the number of high frequency blocks, per total number of blocks, for each quadrant.

Values for both mean number of high-frequency blocks (HF_{mean}) and standard deviation per quadrant (σ_{qx}), are calculated from the four quadrant values. It was noted that for uniform crowd images, HF_{mean} and σ_{qx} should have high and low values respectively. A crowd image confidence feature set, $\{Fv_3\}$, is calculated as follows:

$$\{Fv_3\} = HF_{mean} - \text{Avg}\ (\sigma_{q1}, \sigma_{q2}, \sigma_{q3}, \sigma_{q4})$$

D. Feature Detector 4: Speech-Band Audio Activity

It is proposed that speech band energy may be estimated by examining the scalefactors of the encoded audio bitstream.

A fundamental component of MPEG audio bitstreams is the scalefactor. Scalefactors are variables that normalize groups of audio samples such that they use the full range of the quantiser. The scalefactor for such a group is determined by the next largest value to the maximum of the absolute values of the samples. Hence, they provide an indication of the maximum power (volume) of a group of samples.

The scalefactors may be individually extracted from any of 32 equally spaced frequency subbands, which uniformly divide up the 0-20kHz audio bandwidth. Therefore, they provide an efficient audio filtering, envelope tracking technique.

Subbands 2-7 correspond to the frequency range 0.6kHz – 4.4kHz, which approximates to that of the speech band. Therefore, manipulation of scalefactors from these subbands provides for the establishment of a speech-band energy profile of the audio data.

For the audio tracks of the video corpus, scalefactor data, from subbands 2-7, is stripped from the audio bitstream and grouped together in 0.5s intervals. The average of the root-mean-square scalefactor values is then calculated, to yield the feature set $\{Fv4\}$, i.e. (normalised) speech-band energy levels for each temporal interval.

$$\{Fv_4\} = [\ \text{Avg}\ (\ \text{Rms}\ (\text{scalefactors}_{\ \text{subbands 2-7}})\ )\ ]_{\ 0.5s}$$

V. FEATURE DATA AGGREGATION

For a each shot, retained from the pre-processor, it is required to process the shot's corresponding feature data, such that it is tagged with its own *feature vector (V)*. The vector should convey *event*-critical information, i.e. for a given shot vector, it is required that the individual vector component coefficient (Vcc) values, computed from the feature data sets, reflect the features' contribution to the overall probability that the shot contains an event.

As mentioned previously, the proposed approach to the event detection task is to analyze the content characteristics immediately following their incidence, i.e. for a given shot, retained from the pre-processing phase, the classification decision is to be made on the basis of the audio-visual content characteristics, corresponding to a short critical period subsequent to the shot-end.

To realize this, for each retained shot a 20-second critical seek window (CSW), beginning 5s prior to the shot-end boundary, is used. Feature data from within this window is processed as follows, which then yields the Vcc values for the associated shot vector.

$$V = [Vcc_1, Vcc_2, Vcc_3, Vcc_4]_{\ SHOT}$$
$$Vcc_1 = \text{Max}\ \{Fv_1\}_{CSW}\ ,\ Vcc_2 = \text{Avg}\ \{Fv_2\}_{CSW}$$
$$Vcc_3 = \text{Max}\ \{Fv_3\}_{CSW}\ ,\ Vcc_4 = \text{Avg}\ \{Fv_4\}_{CSW}$$

Following this process, a 4-dimensional feature vector is associated with each shot retained from the pre-processor. These shot vectors constitute the input data of the

training/testing phases, for the machine learning algorithm.

VI. EXPERIMENTAL CASE-STUDY: SOCCER VIDEO

The following is an account of an experimental case-study utilizing a 50-hour corpus of soccer video, obtained from a variety of broadcast sources. In the context of soccer video, the *event* to be detected corresponds to a *goal score*. All content pre-processing, feature extraction, and data-to-vector aggregation, was executed on the corpus exactly as described in previous sections.

A. Machine Learning Algorithm: Support Vector Machine

A special feature of the support vector machine (SVM) is the *cost-factor*. This is a user-defined parameter that determines by how much training errors in the positive examples outweigh those in the negative examples, i.e. it allows adjustment of the cost of false positives Vs cost of false negatives. Effectively, variation of this value during the training phase allows the user to tune the classifier such that it is biased towards perfect classification of either negative or positive examples.

B. SVM: Training & Testing Phases

The 50-hour experimental corpus was divided into two equal 25-hour sections, one section each for both the training and testing phases. The 25-hour training corpus was manually annotated, and this data was used to train the SVM, such that it yielded an *event* model, inferred from the key feature patterns.

During the training phase the SVM was trained for wide-varying values of cost-factor. Each resulting trained classifier (generated in the training phase, via cost-factor adjustment), was run on the 25-hour test corpus. The shots classified as those containing *events* were compared to that of a ground truth baseline, generated via a manual annotation of the content. Precision and Recall statistics were calculated.

C. Evaluation

A plot of precision against recall for varying values of *cost-factor* is presented in Fig. 2. Manually selecting a particular *cost-factor* value allows us to bias the classifier at any chosen point on the characteristic. Clearly it is desirable to maintain both 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 exciting, albeit non-*event*, moments as opposed to significant omissions. The classifier defined by a cost-factor of 3.2 (point-X), provides simultaneous precision of 54% and recall of 94%, which represents a favourable retrieval-rejection trade-off.

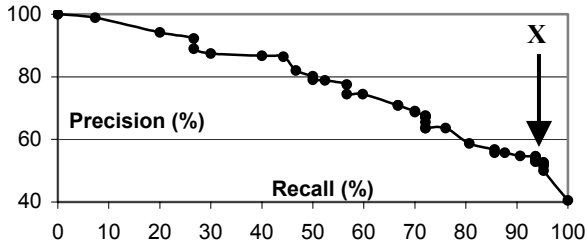

Fig. 2. Precision Vs Recall for varying cost-factor

VII. CONCLUSIONS

In this paper we have outlined a proposed framework for event detection in field-sports broadcast video, in which an *event* model is inferred by an SVM based on the evidence of four significant feature detectors, which are chosen such that they are recyclable across multiple sports genres within the field-sport domain. For efficiency, where possible, the feature detectors are implemented from low-level data taken directly from the compressed domain audio/visual bitstream.

As a preliminary experimental case-study, the techniques have been applied and tested on soccer video. A large experimental corpus, which was obtained from multiple broadcast sources, was utilized for this analysis. Compared to a manually annotated baseline, it has been shown that both high precision and recall statistics are achievable. Furthermore, it has been described how the SVM can be tuned such that the classification may be biased to any point on the precision recall characteristic of the model.

REFERENCES

[1] R. Dayhot, A. Kokaram, and N. Rea, "Joint audio-visual retrieval for tennis broadcasts," in *Proc. ICASSP 2003*.

[2] 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.

[3] W. Zhou, A. Vellaikal, and C-C.J. Kuo, "Rule-based video classification system for basketball video indexing," *Proc. ACM Multimedia 2000*, Los Angeles, USA, November 2000.

[4] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," *Proc. ACM Multimedia 2000*, Los Angeles, USA.

[5] 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.

[6] O. Utsumi, K. Miura, I. Ide, S. Sakai, and H. Tanaka, "An object detection method for describing soccer games from video," *Proc. IEEE ICME 2002*, Lausanne, Switzerland.

[7] H. Kim and K. S. Hong, "Soccer video mosaicing using self-calibration and line tracking," *Proc. IEEE Int. Conf. on Pattern Recognition (ICPR) 2000*, Barcelona, Spain.

[8] D. Yow, B-L. Yeo, M. Yeung and B. Liu, "Analysis and presentation of soccer highlights from digital video," *Proc. Asian Conference on Computer Vision 1995*, Singapore.

[9] C. Wu, Y-F. Ma, H-J. Zhang, and Y-Z. Zhong, "Events recognition by semantic inference for sports video," *Proc. IEEE ICME 2002*, Lausanne, Switzerland.

[10] D. Zhong and S-F. Chang, "Structure analysis of sports video using domain models," *Proc. IEEE ICME 2001*, Japan.

[11] 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.

[12] D. Sadlier, N. O'Connor, S. Marlow, N. Murphy, "A Combined Audio-Visual Contribution to Event Detection in Field Sports Broadcast Video. Case Study: Gaelic Football," *Proc. IEEE ISSPIT 2003*, Darmstadt, Germany.

[13] T. Risse, "Hough Transform for Line Recognition," *Computer Vision and Image Processing*, 1989, 46, 327-345, 1989.

[14] J-C. Terrillon and S. Akamatsu "Comparative performance of different chrominance spaces for colour segmentation and detection of human faces in complex scene images," *Proc. 12th Conf. on Vision Interface*, vol.2, pp.180-187, May 1999.