Event Detection in Field Sports Video Using Audio–Visual Features and a Support Vector Machine

David A. Sadlier and Noel E. O'Connor, *Member, IEEE*

***Abstract*—In this paper, we propose a novel audio–visual feature-based 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 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 generically 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.**

***Index Terms*—Event detection, field sports video, MPEG, signal processing, support vector machine (SVM).**

I. Introduction

MODERN developments in digital video compression technologies have paved the way for extensive archiving of content. More and more video material is being digitized and archived worldwide [1]. However, the problems of limited bandwidth and/or battery life, which impede the development of hand-held wireless video applications, built upon such archives, suggests an increasingly crucial role for highlighting or summarization of such content. Sports video analysis in particular 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, and within this area, soccer video analysis almost saturates the field. In [2]–[4] the authors exploit visual characteristics such as camera motion, and perform object tracking, in automatically detecting soccer video highlights. Cabasson *et al.* [5] exploit both visual and aural features to achieve the same aim. Visual analysis methods, including player tracking, for automatic soccer video indexing are described in [6], while Xie *et al.* [7] present visual methods for soccer video structure analysis. Basketball video is also a well-researched specific genre. Nepal *et al.* [8] describe audio–visual techniques for the detection of basketball goals, while Zhang and Ellis [9] present an audio only approach. Further instances of automated basketball video indexing can be found in [10] and [11]. Audio and/or video based analyses for event detection toward summarization or indexing, can be found in the literature for a variety of sports genres; formula-1 [12]; baseball [13]–[15]; cricket [16]; tennis [17]–[19]; American football [20], and Gaelic football [21]. 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, which tend to be particular to each sports type. Their effectiveness is somewhat diluted by their inherent inapplicability to other sports genres. Relatively little prior work addresses a more generic, genre independent methodology, concerning the more challenging task of revealing the common structures of multiple events across multiple genres. In Hanjalic *et al.* [22], the author discusses a generic approach to highlights detection in sports video, while methods for the syntactical segmentation of generic sports video are presented in Jianyun *et al.* [23]. However, both these works suffer from the fact that only results concerning soccer video are presented. Wu *et al.* [24] present visual analysis techniques for event detection, which, albeit within integrated athletics broadcasts, are tested successfully across track-and-field sports genres. Zhong *et al.* [25] illustrate visually-based structure analysis methods, which are examined across tennis and baseball video. Generic aural and/or visual techniques for sports video are used in [26]–[28] for semantic annotation and highlights detection, respectively, and together these approaches are shown to operate across multiple genres. Further examples of multiple genre approaches may be found via [29] and [30]. Pan *et al.* [31] discuss the detection of replay segments for generic sports video by a logo detection method. However, while not genre specific, it may require *a priori* knowledge of a particular broadcaster. Finally, Babaguchi *et al.* [32] present a video-textual technique for event detection in generic sports video, which exploits closed-captioning. However, this has an inherent geographical restriction, e.g., closed-captioning of broadcast sports content is not mainstream in Europe.

For a given indexing or event detection task, it is unfeasible to consider that there exists a unique solution that will operate successfully across all genres of sports video. For example, a solution that functions effectively on golf video cannot conceivably be expected to work to the same degree on Sailing content. The work described herein, aims to set some meaningful boundary on a generic approach to sports video event detection. To this end, 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 genres

Manuscript received June 3, 2004; revised December 21, 2004. This work was supported in part by the Informatics Research Intitiative of Enterprise Ireland and in part by the Irish Research Council for Science, Engineering and Technology. This material is based on work supported in part by the IST program of the EU in the project IST-2000-32795 SCHEMA. This paper was recommended by Associate Editor P. van Beek.

The authors are with the Centre for Digital Video Processing, School of Electronic Engineering, Dublin City University, Dublin 9, Ireland (e-mail: sadlierd@eeng.dcu.ie; oconnorn@eeng.dcu.ie).

Digital Object Identifier 10.1109/TCSVT.2005.854237

that fall within this ambit. The reasoning behind this is that field sport broadcasts (i.e., soccer, American football, rugby, Australian rules football, field hockey, Gaelic football, hurling etc.) all share common characteristics, which may be generically exploited in the analysis. These include:

1) two opposing teams + referee(s);
2) enclosed playing area;
3) grass pitch;
4) field lines;
5) commentator voice-over;
6) spectator cheering;
7) on-screen video text (scoreboard);
8) three well-defined styles of camera shot: global (main), zoom-in, and extreme close-up;
9) objectives concerned with territorial advancement and directing an object (e.g., ball) toward a specific target.

For the purposes of these experiments, and for the following discourse, an *event* hereafter refers to that constituting a successful scoring incident from a broadcast field sport game, e.g., a *goal* in soccer or hockey, a *try* or *conversion* in rugby, a *point* in Gaelic football, etc. Each one of the above characteristics is present in all genres of field sports broadcasts. Thus, if event-defining feature detectors are designed so that they are rooted in such commonality, then it is possible that no other assumption need be made about the nature of the content for the analysis.

For the purposes of developing our approach a video corpus consisting of a variety of field sport genres including rugby, hockey, soccer and Gaelic football was created. To ensure generality, the content was captured from various broadcasts sources. Video images were captured at CIF resolution (352 pixels wide $*$ 288 pixels high) at a rate of 25 frames/s. Audio data was captured in 128 kbits/s stereo, with sampling frequency of 44 100 samples/s/channel. The entire corpus was compressed according to the MPEG-1 digital video standard. In all, 100 h of content was captured. This was split into two subcorpuses. The training corpus was to be used in developing model hypotheses, and the testing corpus purely for evaluation purposes. Both were manually annotated, such that advanced knowledge of all event locations was ascertained—and subsequently used as our *ground truth*.

II. Field Sport Event Characteristics

In field sports video there exist many circumstances in which events may be manifested. However, this approach does not attempt to model the individual scenarios of such events, but rather model what is common to all situations, irrespective of circumstance.

A subjective examination of multiple-events from multiple field sport genres within the training corpus was performed. From the evidence it was apparent that 98% of all such events were followed by an action replay. While generic methods for action replay detection have been researched [31], most are rooted in exploitation of slow-motion based frame repetition. Thus, with the onset of high-speed camera technology, such methods are vulnerable to breakdown. However, it was noted that 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, the evidence suggests that during the reaction-phase, the characteristics of the content typically include: 1) a close-up shot of the player(s) and/or relevant parties involved; 2) a camera shot showing the crowd celebrating; 3) an increase in audio activity (particularly in the voice band frequency range, corresponding to commentator vocals); 4) activity in the on-screen graphics (scoreboard); and 5) a surge in near-field motion activity (as the camera attempts to capture the intense celebratory behavior of the scorer). Fig. 1 illustrates the distribution of reaction-phase durations across all types of multigenre events contained in the training corpus. The mode reaction-time duration was 18 s, which corresponds to 16% of events. It is clear from this data that a negligible amount of reaction-phase durations are in excess of 25 s. Considering this 25-s limit as a post-event critical-seek-window (CSW), it was manually quantified (using prescribed feature extraction tools where required) that over said events, within this CSW: 1) 96% were immediately followed by a close-up image; 2) 73% were followed by a sequence of crowd images; 3) 84% of the accompanying audio tracks had peak levels in excess of corresponding broadcast mean levels; 4) 61% exhibited a temporary removal of the on-screen graphic during scoreboard update; and 5) 76% of the near-field motion activity measures had peaks in excess of corresponding broadcast mean levels. Furthermore, considering point 4) in Section I, the evidence also suggests that for a given event, it is typical for the action to be situated at the end-zone region of the playing field. In fact, it was manually confirmed that, over all events contained in the training corpus, 73% occurred with the camera focused on action in the end-zone region of the field.

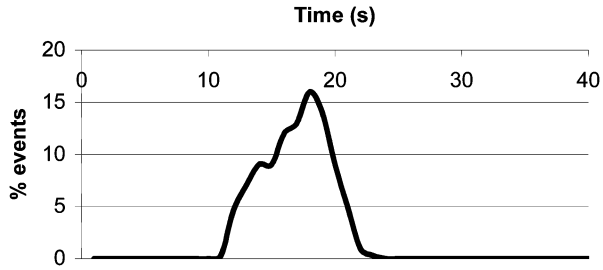

Fig. 1. Distribution of reaction-phase durations across multiple events contained in the training corpus.

III. Content Preprocessing

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

A. Shot Boundary Detection

Because of the high tempo nature of field sports, during the live action segments the broadcast director has little chance to utilize shot transition types other than hard shot-cuts. In fact it was manually quantified that 94% of all shot transitions within our multigenre training corpus were of this nature. It was found that dissolves, wipes, etc., tend only to occur when the director has time to be more creative, i.e., during a break in the play or

Fig. 2. Close-up image samples.

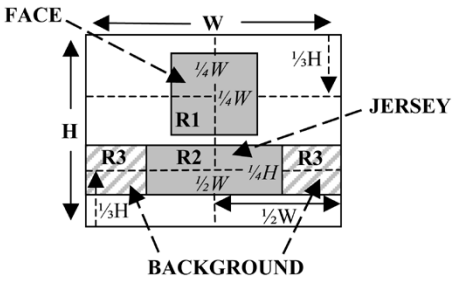

Fig. 3. Regions of expectation for close-up images.

during a break in the live action (e.g., during action replays). For these reasons we concerned ourselves only with hard shot-cut detection. Hard shot-cut boundary detection is effectively a solved problem in the area of digital video analysis, and for this purpose our own algorithm [33] was employed.

B. Probing Domain Restriction

Supplementary content (e.g., advertisements) typically accompanies the main event in a sports broadcast. Segments such as athlete profiles, highlights of recent events, etc., may contain audio–visual signal attributes with patterns similar to the events of interest. Hence, it is desirable that the temporal boundaries of the live play segments are detected within the overall broadcast, and the retrieval *probing domain* restricted accordingly. For a given sports genre, the play segments may be detected by searching through the entire audio track for extended periods of sustained volume [34]. However, this method requires advanced knowledge of specific sports genres. For our generic approach, an advertisement detection algorithm [35] is utilized, which has been shown to operate successfully on generic video. Once advertisements are located, they are disregarded from the probing domain.

C. Close-Up Image Detection and Shot Filtering

Although color-based object recognition may not be practical in many video scenarios, it is suitable for sports, where colors are purposely used to differentiate players, and clearly defined rules constrain the action [36]. As a result the colors of the playing surface and the players/referee shirts usually consist of one or two (striped) dominant colors. Two close-up images are displayed in Fig. 2. From these examples it is clear that the salient characteristics of such images are: 1) the presence of a face in the top-middle-center region (i.e., the focus) of the frame, together with 2) a jersey in the bottom-middle region of the frame, occluding an arbitrary background. Based on the evidence of numerous close-up images selected from the training corpus, estimations of the regions of expectation for said characteristics were delineated—see Fig. 3. *Region-1* (R1), the region

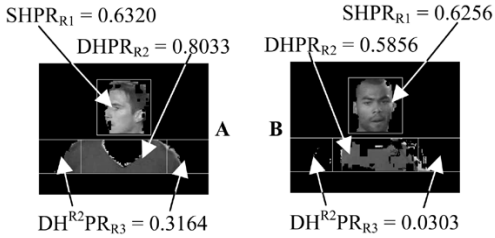

Fig. 4. Hue analysis of critical regions of interest.

of expectation for the face, was chosen to be a square of dimension 1/4 W centered on the vertical median, at a horizontal position corresponding to 1/3 H from the top of the image. *Region-2* (R2), the region of expectation for the shirt, was chosen to be a rectangle of dimensions 1/4 H $\times$ 1/2 W centered on the vertical median, at a horizontal position corresponding to 1/3 H from the bottom of the image. *Region-3* (R3), the region of expectation for the background, is simply the outstanding regions in extending the dimensions of Region-2 to the image width.

A proposed close-up detection model is based on the degree to which the image has: 1) a skin-toned entity (i.e., a face) in Region-1 and 2) a monochromatic entity (i.e., a shirt) in Region-2, occluding the background in Region-3. Consider the examples A and B in Fig. 4. It has been shown that skin-color clusters well in the hue space at $[10°–55°]$ [37], and therefore may be easily discriminated from other colors in the images. For Regions-1 the skin-hue pixel ratios $(\mathrm{SHPR_{R1}})$ are calculated. For image A this corresponds to 0.6320, likewise for image B this is 0.6256. For Regions-2 the dominant hue pixel ratios $(\mathrm{DHPR_{R2}})$ is computed. For images A and B this corresponds to 0.8033 and 0.5856, respectively.

For Regions-3 the pixel ratios of the same dominant hue of Region-2 $\left(\mathrm{DH^{R2}PR_{R3}}\right)$ is computed. For images A and B this corresponds to 0.3164 and 0.0303, respectively. For an ideal close-up image with face and shirt perfectly encapsulated in the appropriate regions, it is expected that both $\mathrm{SHPR_{R2}}$ and $\mathrm{DHPR_{R2}}$ will be relatively large values, while $\mathrm{DH^{R2}PR_{R3}}$ will be a relatively small value. Hence, a formula for image close-up confidence (CuC) exploiting these perceptions is defined as

$$\mathrm{CuC} = \mathrm{SHPR_{R1}} * (\mathrm{DHPR_{R2}} - \mathrm{DH^{R2}PR_{R3}}). \quad (1)$$

Using the values of examples A and B in this formula yields close-up confidence values of 0.3474 for image A, and 0.3078 for image B, respectively. Comparing these values with those generated for nonclose-up images shows that this model works well at discriminating such images in the context of the limited image domain of field sports content. For example, similar analyses performed on images 2 and 3 in Fig. 5 yield CuC values of 0.0390 and 0.0133, respectively.

As outlined in Section II, it was estimated that 97% of all training corpus events included a close-up image sequence during the post-event, preaction replay, 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 (as defined in Section II). This should provide for a reasonable

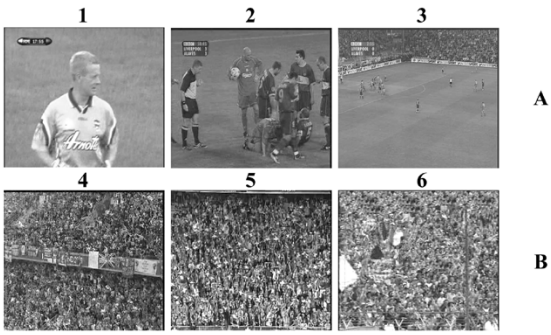

Fig. 5. Field-sports video images. Block A: Generic images from standard camera perspectives. Block B: Crowd Images.

TABLE I
VARIANCE OF EVENT SHOT-RETENTION VERSUS IMAGE CUC THRESHOLD

CuC Threshold	0.01	0.1	0.15	0.2	0.25
% Retained Events	97%	96%	88%	81%	67%

initial retention of eventful shots and rejection of uneventful shots. Experiments were performed on the training corpus. The variance of event shot-retention versus image CuC threshold is presented in Table I. As expected the number of retained event-shots decreases as the condition threshold becomes more stringent. Based on these observations, a CuC threshold of **0.1** was chosen since this provided **96%** retention of all training corpus events, which is acceptably close to the 97% maximum. When combined with the probing domain limit, this condition was found to provide overall preprocessor content rejection of **43%** across the entire training corpus.

IV. FEATURE DETECTORS: DESIGN AND IMPLEMENTATION

A. Feature Detector 1: Crowd Image Detection

As outlined in Section II, it was estimated that of all the events contained in the training corpus, 73% contained a crowd image sequence within the post-event reaction-phase. Thus, it is postulated that crowd image sequence detection is a valuable cue, which should contribute effectively to the event detection task. Crowd image detection may be performed by exploiting the inherent characteristic that, in the context of a typically noncomplex image environment, such images are relatively detailed—see Fig. 5. It is proposed that discrimination between detailed and nondetailed pixel blocks may be made by examining the number of nonzero frequency (ac) discrete cosine transform (DCT) coefficients used to represent the data in the frequency domain. It may be assumed that an (8×8) pixel block, which is represented by very few ac-DCT uniform coefficients, contains spatially consistent, nondetailed data. A block which requires a considerable amount of ac-DCT coefficients for its 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. 5, block A. Therefore, in the context of this limited environment, it is proposed that crowd images may be isolated by detecting such uniformly, high frequency images.

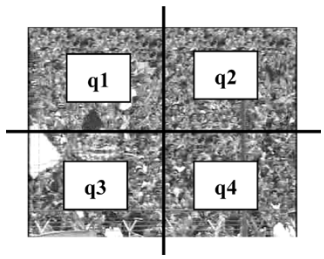

Fig. 6. Division of image into four quadrants.

Each I-frame is divided into four quadrants—see Fig. 6. For each quadrant of each image, the ac-DCT coefficients of every (8×8) luminance pixel block are analyzed. If the number of ac-DCT coefficients used to encode such blocks is greater than a preselected 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.

These values are normalized to lie between zero and one, and values for both mean number of high-frequency blocks $(\mathrm{HF}_{\mathrm{mean}})$ and standard deviation per quadrant (σ_{qx}), are calculated from the four quadrant values. It was noted that for uniform crowd images, $\mathrm{HF}_{\mathrm{mean}}$ and σ_{qx} should have high and low values, respectively. A crowd image confidence feature set, $\{Fv_1\}$, is calculated as follows:

$$\{Fv_1\} = \mathrm{HF}_{\mathrm{mean}} - \mathrm{Avg}(\sigma_{\mathrm{q1}}, \sigma_{\mathrm{q2}}, \sigma_{\mathrm{q3}}, \sigma_{\mathrm{q4}}). \quad (2)$$

B. Feature Detector 2: Speech-Band Audio Activity

Again, as mentioned in Section II, of the audio tracks accompanying the events in the training corpus, during the reaction-phases, 84% exhibited (speech-band) energy levels greater than that of corresponding broadcast mean levels. A comprehensive discourse on how speech-band energy may be estimated by examining bit-stream scalefactor [38] weights may be found via [34].

From the audio tracks, scalefactor data, from subbands 2–7, is parsed from the audio bit-stream and grouped together in 0.5-s intervals. The average of the root-mean-square scalefactor values is then calculated, to yield feature set $\{Fv_2\}$—speech-band energy levels for each temporal interval. For each particular broadcast these values are normalized to lie between zero and one

$$\{Fv_2\} = [\mathrm{Avg}(\mathrm{rms}(\mathrm{scalefactors}_{\mathrm{subbands\ 2–7}})]_{0.5\mathrm{s}}. \quad (3)$$

C. Feature Detector 3: On-Screen Graphics Tracking

As mentioned in Section II, it was estimated that of all the events contained in the training corpus, 61% were characterized by a temporary disappearance of the graphic during the scoreboard updating procedure during the post-event reaction-phase. The on-screen scoreboard graphic is a synthesized component placed over video content. See Fig. 7, images 1 and 2. The format of each graphic is particular to each broadcaster, and may even occasionally change format on an intra-broadcaster basis.

Fig. 7. Image 1: Scoreboard graphic location. Image 2: Enlarged view scoreboard graphic. Image 3: Scoreboard graphic following contrast enhancement. Image 4: Mode values for contrast-enhanced scoreboard graphic pixels.

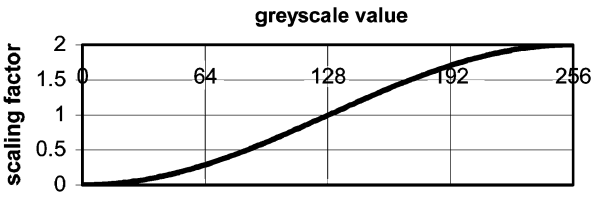

Fig. 8. Contrast enhancer: Pixel value scaling characteristic based on 180° cycle of sine function.

1) Scoreboard Graphic Locating: One of the prominent characteristics of a scoreboard graphic is the presence of text. For text to be visible, there must be a strong luminance contrast between the foreground and background. Thus, during the encoding process, this text requires a relatively large amount of ac-DCT luminance coefficients to represent it. Furthermore, for a given broadcast, the location of the graphic is static, and it is present on-screen for the main duration of the game. Hence, for a given broadcast, the pixel blocks that constitute the graphic will exhibit a large number of ac-DCT luminance coefficients consistently over the duration of the game. In contrast, nongraphic associated pixel blocks will naturally constitute many different aspects of the images captured over the course of the broadcast and, hence, will not exhibit a consistently high number of said coefficients. On this basis the pixel block location of the graphic may pinpointed.

2) Scoreboard Presence/Absence Tracking: It is not uncommon for the scoreboard graphic to have some degree of transparency in its background. This is done to limit the occlusion disturbance to the viewer. Thus, the graphic pixel values, although reasonably constant throughout a given broadcast, are subject to *transparency-noise*, which especially affects those constituting the background. To combat the effects of this on graphic presence/absence tracking, the contrast of the luminance pixel spectrum $[0\text{–}255]$ of the images is warped (enhanced), such that the effects of fleeting contrast variations in the scoreboard pixels are suppressed. Specifically, a 256-bin scaling characteristic (4), based on a 180° cycle period of the sine function, is used to perform this task. This characteristic is illustrated in Fig. 8

$$\text{Contrast Enhancer} = \frac{[2 + 2\text{Sin}\theta]}{2} \quad \frac{3\pi}{2} \leq \theta \leq \frac{5\pi}{2}. \tag{4}$$

The effect of this scaling operator in the luminance domain is to push reasonably dark pixels to very dark, reasonably bright pixels to very bright, while leaving other greylevels relatively unaffected. Pixel values outside the permitted range are clipped accordingly—see Fig. 7, images 2 and 3.

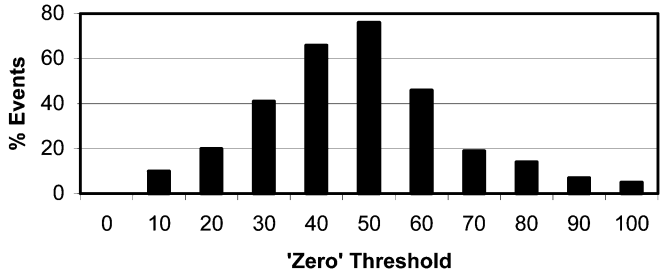

Fig. 9. Variance of number of events with reaction-phase visual activity peaks in excess of corresponding mean levels with "zero" threshold.

Once the pixels of the graphic are pinpointed, their temporal *mode* luminance value is computed, by examining the values over all I-frames within the broadcast. An example of the resulting mode is given in Fig. 7, image 4. The mode values of the graphic region, by definition, represent the most continuously recurring luminance pixel values, over the duration of a given broadcast. Therefore, for a given image, a high absence confidence value will correspond to a high sum of absolute difference (SAD) value between the current image values, and those of the mode.

Thus, for each I-frame, an image-mode SAD operation is performed on the values of the critical pixels. The outputs constitute another feature set $\{Fv_3\}$

$$\{Fv_3\} = \text{SAD}[\text{I-frame}_N, \text{Mode}]_{\text{Graphic Pixels}}. \tag{5}$$

D. Feature Detector 4: Motion Activity Measure

As previously discussed, during the post-event reaction-phase, it is typical for the camera operator to follow a celebrating player(s), generally by means of a close-up shot. Due to the dynamic nature of the content, coupled with the type of camera shot used, a significant increase in near-field visual motion activity is typically apparent.

Visual motion activity is estimated from the evidence conveyed by the motion vectors present in the MPEG video bit-stream [39]. From the video content, every P-frame is extracted and from these images, motion vectors for each macroblock are extracted directly from the encoded bit-stream. From the motion vectors of each P-frame image, a critical statistic, the nonzero vector value (NZVV) is calculated. This is calculated by counting the number of macroblocks in the frame whose motion vector length is greater than a preselected *"zero" threshold*. The "zero" threshold value should be chosen to be large enough as to ignore slow, smooth motion, while detecting, jerky, uneven motion, i.e., the type of turbulent motion expected during the celebratory moments. Fig. 9 illustrates how the percentage of training corpus events with reaction-phase visual activity peaks in excess of corresponding broadcast mean values, varies with the selection of the "zero" threshold. As mentioned in Section II, a maximum value of 76% may be achieved using a "zero"-threshold of **50**. Hence, it is this threshold value that is used for this analysis. The NZVV statistic is calculated for each P-frame, thus yielding a visual activity feature set $\{Fv_4\}$. Higher values should indicate increased visual activity. These data sets are smoothed, i.e.,

Fig. 10. Orientation of most prominent field lines during field sport events. A: rugby. B: hockey. C: soccer.

mean-filtered, using a one-dimensional (1-D) kernel of length $= 7$. The values are then normalized such that they lie between zero and one

$$\{Fv_4\} - \text{Smoothed}[\text{NZVV}_{\text{P-frame}}]. \quad (6)$$

E. Feature Detector 5: Field Line Orientation

As outlined in Section II, a typical characteristic of field sport events is that they are characterized by activity particular to the end-zone region of the playing field. In fact, it was estimated that 73% of all training corpus events adhere to this scenario. For example, events such as *goals*, *tries*, *points*, *conversions*, etc., are achieved either by: 1) directing the ball toward a target in the field end-zone, or 2) player, with ball, advancing toward the end-zone. Due to the fixed position of the camera, the resulting perspective is such that the field lines may only assume certain angles, which lie within a particular narrow interval, relative to the point of observation. Fig. 10 displays the final I-frame images of camera shots corresponding to a try-score in rugby (image A), a goal-score in hockey (image B) and a goal-score in soccer (image C). It is clear from these images that the angles of the most prominent field lines lie within a common narrow interval. Of the events in the training corpus that adhere to these circumstances, it was manually recorded that of the most prominent field lines, only a negligible number of the angles of the final I-frames in the shots mapped outside the interval (5°–25°). Thus, by detecting the field lines and continuously tracking their angle of orientation, it should be possible to detect these images that correspond to field end-zone shots. Again, it was anticipated that analysis of I-frames would be sufficient for this task.

It is assumed that the global mode hue (GMH) value, computed over the duration of the probing domain, corresponds to the pixels of the playing field. This is a reasonable assumption since the playing field is the largest reoccurring object that constitutes the images of a given broadcast. For a given image, the playing field pixels are segmented in the hue space based on this value. Segmentation of field lines from the playing field is achieved via a binarization of the luminance space of the image, using a threshold equivalent to the mode greyscale value of the playing field pixels. This provides for a segmentation of the white field lines (which are bright, i.e., luminance values generally exceed threshold) from the playing field (which is dark, i.e., luminance values generally suppressed by threshold)—see Fig. 11.

A Roberts' edge detector [40] was utilized to isolate the edges of the binarized images. For each edge-detected binary image, the field lines were extracted by means of the Hough transform [41]. Specifically, the pixels of each binary image were transformed from spatial space to Hough space. For each case the angle of the most prominent detected line was chosen as the representative value for that image. These values were calculated for each I-frame, thus yielding a field-line orientation angle feature set $\{Fv_5\}$

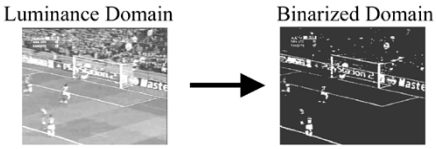

Fig. 11. Field line segmentation based on dynamic threshold.

$$\{Fv_5\} - [\text{Angle of Most Prominent Line}]_{\text{I-frame}}. \quad (7)$$

F. Other Features Investigated

So far, five critical features have been described for the pattern recognition phase of the event detection task. Clearly, to maximize the accuracy of the overall system, exploitation of other useful features would be desirable.

As outlined in Section III, in the case of 61% of training corpus events, the scoreboard update procedure occurs while the graphic temporarily disappears off-screen. In other circumstances, it would be desirable to detect the relatively minute activity of the on-screen update of individual scoreboard numerals. However, there are certain problems associated with compressed video images that prevent an optical character recognizer from easily recognizing such characters. For example, an effect of transparency-noise is that the background luminance is typically unstable, which has detrimental consequences for the segmentation of the text region into foreground/background regions. For these reasons, no further development of this feature has been investigated to date.

In the audio domain, as well as a surge in energy level during exciting moments, subjective evidence shows that it is not uncommon to observe an increase in the pitch of the commentator vocals during periods of heightened enthusiasm. Hence, an indicator of this characteristic should further contribute to the event detection task. There exist many reliable vocal pitch estimation techniques in the literature, e.g., [42], [43]. However, they assume pure speech signal input, i.e., free from background noise, which is certainly not characteristic of the content dealt with in these experiments. Thus, the task initially concerns the isolation of a clean speech signal from a noisy ensemble. If it is a case that the vocal signal is a mono signal, center panned in a stereo pair, then, by exploiting both this and the assumed stereo asymmetry of the background/spectator noise, it is possible to isolate the vocal signal from the original stereo signal. However, this approach is critically dependent upon the above assumption, and the inconsistency observed in such has discouraged any further development.

V. Feature Data Aggregation

For a given shot, it is required to aggregate its corresponding feature data, estimated exactly as described in the previous section, such that it is tagged with its own *shot feature vector* (SFV)

$$(\text{SFV}) = [\text{Vcc}_1, \text{Vcc}_2, \text{Vcc}_3, \text{Vcc}_4, \text{Vcc}_5]. \quad (8)$$

The vector should convey *event*-critical information, i.e., for a given SFv, 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 content of the shot exhibits an event. Specifically, within the CSW from the end-boundary of each retained shot, Vcc_1 is defined as the maximum I-frame crowd image confidence value; Vcc_2 is defined as the maximum speech-band energy value; Vcc_3 is defined as the maximum I-frame graphic activity value; and Vcc_4 is defined as the maximum P-frame visual motion activity confidence value

$$\begin{aligned}\mathrm{Vcc}_1 &= \mathrm{Max}\{Fv_1\}_{\mathrm{CSW}}\\ \mathrm{Vcc}_2 &= \mathrm{Max}\{Fv_2\}_{\mathrm{CSW}}\\ \mathrm{Vcc}_3 &= \mathrm{Max}\{Fv_3\}_{\mathrm{CSW}}\\ \mathrm{Vcc}_4 &= \mathrm{Max}\{Fv_4\}_{\mathrm{CSW}}.\end{aligned} \tag{9}$$

$\{Fv_5\}$ 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 activity located in the field end-zone, the field-line angles of the last six I-frames preceding the shot-end boundary (*critical I-frames*) are examined. As the corresponding orientations are found to lie in the key range (5°–25°), the confidence value increases accordingly, i.e.,

$$\begin{aligned}&\mathrm{Vcc}_5 = i * 0.166\,666\\ &\{i = \#\text{critical I-frame orientations within the key range}\}.\end{aligned} \tag{10}$$

If a given shot duration does not contain at least six I-frames, Vcc_5 is set to zero.

These five-dimensional SFvs constitute the input data for the classifier.

VI. Support Vector Machine

A. Overview

The performance of a learning machine is measured by its generalization [44]. This is the ability of a resultant decision function to correctly classify data points not in the training set. Support vector machine (SVMs) are an implementation of the latest generation of machine learning algorithms based on recent advances in statistical learning [44]. SVMs offer a solution to optimizing the generalization performance of a decision function, inferred from a given set of training data. For event detection, the desired output is a binary (positive or negative) decision. It is for this reason that an SVM (with radial basis function kernel) was chosen as the learning algorithm for the purposes of this experiment. Alternative classification schemes such as hidden Markov models are likely to be more appropriate in video indexing applications, where continuous knowledge of past and present states is desired [45]. A special feature of an SVM is the *regularization parameter*, C. This is a user-defined value that is set during the training phase. Variation of this parameter essentially provides for an effective tuning aspect to the classification. A comprehensive discourse of this topic can be found in [44].

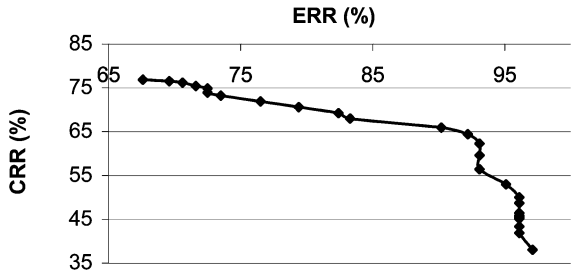

Fig. 12. Rugby content: ERR versus CRR.

VII. Experimental Evaluation

A. Training and 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. To provide an indication of the range of possible results obtainable from the model, 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 *noneventful*. Consecutive shots with identical classification were grouped together yielding eventful or noneventful *scenes*.

B. Results and 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. 12 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 that constitute 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 toward a saturation limit of 78% CRR and 68% ERR. Similar statistical analyses are performed for the other field sport genres. Fig. 13 plots CRR versus ERR for retrieval of *goals* in soccer, Fig. 14 plots CRR versus ERR for retrieval of *goals* in hockey, and Fig. 15 plots CRR versus ERR for retrieval of *goals* and *points* in Gaelic football.

For cross-genre performance comparison, Table II 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 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 nonevents as opposed to event omissions. Thus, the evaluation

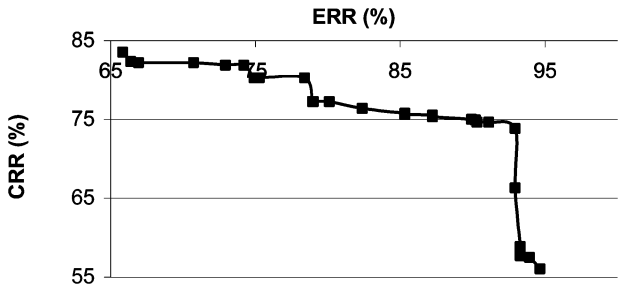

Fig. 13. Soccer content: ERR versus CRR.

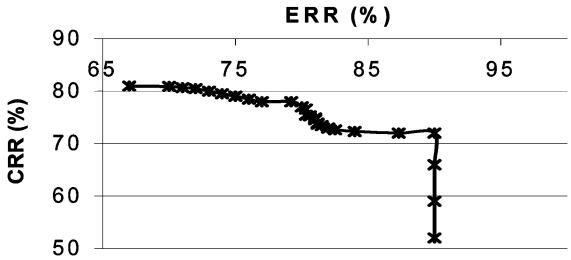

Fig. 14. Hockey content: ERR versus CRR.

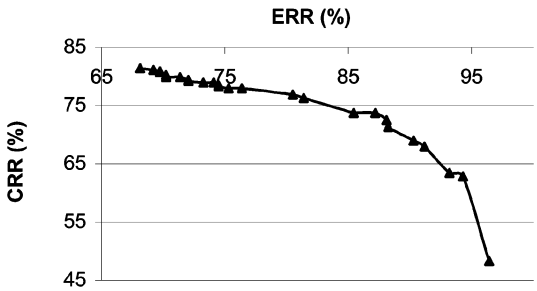

Fig. 15. Gaelic football content: ERR versus CRR.

TABLE II
ERR/CRR PREPROCESSOR LIMITS AND 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%

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. 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 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, at said point.

It is postulated that the reasons for these 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. On the other hand, 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 relatively sporadic abundance of the features critical to this analysis.

VIII. CONCLUSION AND FUTURE WORK

In this paper we have outlined an approach to 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 generically 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 the classification characteristic of the model.

Future work will focus on further investigation on certain key aspects of the scheme. First, an investigation is required that shows the individual contribution of each feature to the task. Second, an analysis of the effect of feature threshold selection on overall system performance (not just on feature performance) is desirable. A final issue is the efficiency. Where possible, the feature detectors are implemented from low-level data taken directly from the compressed domain audio/visual bit-stream. This aspect of the system will be quantified in the future, with a view to making the overall approach as computationally efficient as possible.

REFERENCES

[1] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Video abstracting," *Commun. ACM*, vol. 40, no. 12, pp. 55–62, Dec. 1997.
[2] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala, "soccer highlights detection and recognition using HMMs," in *Proc. IEEE ICME*, Lausanne, Switzerland, 2002, pp. 825–828.
[3] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Process.*, vol. 12, no. 6, pp. 796–807, Jun. 2003.
[4] D. Yow, B.-L. Yeo, M. Yeung, and B. Liu, "Analysis and presentation of soccer highlights from digital video," in *Proc. Asian Conf. Computer Vision*, Singapore, 1995, pp. 499–503.
[5] R. Cabasson and A. Divakaran, "Automatic extraction of soccer video highlights using a combination of motion and audio features," in *Proc. Symp. Electronic Imaging: Science and Technology: Storage and Retrieval for Media Databases*, vol. 5021, Jan. 2002, pp. 272–276.
[6] O. Utsumi, K. Miura, I. Ide, S. Sakai, and H. Tanaka, "An object detection method for describing soccer games from video," in *Proc. IEEE ICME*, Lausanne, Switzerland, 2002, pp. 45–48.
[7] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden Markov models," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal (ICASSP)*, 2002, pp. 4096–4099.
[8] S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic detection of goal segments in basketball videos," in *Proc. ACM Multimedia*, 2001, pp. 261–269.

[9] D. Zhang and D. Ellis, "Detecting sound events in basketball video archive," Dept. Electronic Eng., Columbia Univ., New York, 2001.
[10] W. Zhou, A. Vellaikal, and C.-C. J. Kuo, "Rule-based video classification system for basketball video indexing," in *Proc. ACM Multimedia 2000*, Los Angeles, CA, Nov. 2000, pp. 213–216.
[11] D. D. Saur, Y.-P. Tan, S. R. Kulkarni, and P. J. Ramadge, "Automated analysis and annotation of basketball video," in *Proc. Symp. Electronic Imaging: Science and Technology: Storage and Retrieval for Image and Video Databases*, vol. 3022, Jan. 1997, pp. 176–187.
[12] M. Petkovic, V. Mihajlovic, M. Jonker, and S. Djordjevic-Kajan, "Multimodal extraction of highlights from TV formula 1 programs," in *Proc. IEEE ICME*, Lausanne, Switzerland, 2002, pp. 817–820.
[13] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proc. ACM Multimedia*, Los Angeles, CA, 2000, pp. 105–115.
[14] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden Markov models," in *Proc. IEEE ICIP*, 2002, pp. 609–612.
[15] T. Kawashima, K. Tateyama, T. Iijima, and Y. Aoki, "Indexing of baseball telecast for content-based video retrieval," in *Proc. IEEE ICIP*, 1998, pp. 871–875.
[16] M. Lazarescu, S. Venkatesh, and G. West, "On the automatic indexing of cricket using camera motion parameters," in *Proc. IEEE ICME*, 2002, pp. 809–813.
[17] R. Dayhot, A. Kokaram, and N. Rea, "Joint audio–visual retrieval for tennis broadcasts," in *Proc. IEEE ICASSP*, 2003, pp. 561–564.
[18] F. Sudhir, J. C. M. Lee, and A. K. Jain, "Automatic classification of tennis video for high-level content-based retrieval," in *Proc. Int. Workshop Content-Based Access of Image and Video Databases (CAIVD'98)*, 1998, pp. 81–90.
[19] E. Kijak, L. Oisel, and P. Gros, "Temporal structure analysis of broadcast tennis video using hidden Markov models," in *Proc. Symp. Electronic Imaging: Science and Technology: Storage and Retrieval for Media Databases*, vol. 5021, Jan. 2003, pp. 277–288.
[20] Li and M. I. Sezan, "Event detection and summarization in american football broadcast video," in *Proc, Symp. Electronic Imaging: Science and Technology: Storage and Retrieval for Media Databases*, vol. 4676, Jan. 2002, pp. 202–213.
[21] D. Sadlier, N. O'Connor, S. Marlow, and N. Murphy, "A combined audio–visual contribution to event detection in field sports broadcast video. Case study: Gaelic football," in *Proc. IEEE ISSPIT*, Darmstadt, Germany, 2003, pp. 552–555.
[22] A. Hanjalic, "Generic approach to highlights extraction from a sport video," in *Proc. IEEE ICIP*, 2003, pp. 1–4.
[23] C. Jianyun, L. Yunhao, L. Songyang, and W. Lingda, "A unified framework for semantic content analysis in sports video," in *Proc. 2nd Int. Conf. Information Technology for Application (ICITA)*, 2004, pp. 149–153.
[24] C. Wu, Y.-F. Ma, H.-J. Zhang, and Y.-Z. Zhong, "Events recognition by semantic inference for sports video," in *Proc. IEEE ICME*, Lausanne, Switzerland, 2002, pp. 805–808.
[25] D. Zhong and S.-F. Chang, "Structure analysis of sports video using domain models," in *Proc. IEEE ICME*, Japan, 2001, pp. 920–923.
[26] J. Assfalg, M. Bertini, C. Colombo, and A. Del Bimbo, "Semantic annotation of sports videos," in *Proc. IEEE Multimedia*, vol. 9, 2002, pp. 52–60.
[27] Z. Xiong, R. Radhakrishnan, and A. Divakaran, "Generation of sports highlights using motion activity in combination with a common audio feature extraction framework," in *Proc. IEEE ICIP*, 2003, pp. 5–8.
[28] R. Radhakrishan, Z. Xiong, A. Divakaran, and Y. Ishikawa, "Generation of sports highlights using a combination of supervised & unsupervised learning in audio domain," in *Proc. Int Conf. Pacific Rim Conf. Multimedia*, vol. 2, Dec. 2003, pp. 935–939.
[29] K. A. Peker, R. Cabasson, and A. Divakaran, "Rapid generation of sports video high-lights using the MPEG-7 motion activity descriptor," in *Proc. Symp. Electronic Imaging: Science and Technology: Storage and Retrieval for Media Databases*, vol. 4676, Jan. 2002, pp. 318–323.
[30] B. Li and M. I. Sezan, "Event detection and summarization in sports video," in *Proc. IEEE Workshop Content-Based Access of Image and Video Libraries (CBAIVL)*, 2001, pp. 132–138.
[31] H. Pan, B. Li, and M. I. Sezan, "Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions," in *Proc. IEEE ICASSP*, 2002, pp. 3385–3388.
[32] N. Babaguchi, Y. Kawai, and T. Kitashi, "Event based indexing of broadcasted sports video by intermodal collaboration," *IEEE Trans. Multimedia*, vol. 4, no. 4, pp. 68–75, Mar. 2002.
[33] C. O'Toole, A. Smeaton, N. Murphy, and S. Marlow, "Evaluation of shot boundary detection on a large video test suite," in *Proc. Challenges in Image Retrieval*, Newcastle, U.K., Feb. 1999.
[34] D. A. Sadlier, S. Marlow, N. O'Connor, and N. Murphy, "MPEG audio bit-stream processing toward the automatic generation of sports programme summaries," in *Proc. IEEE ICME*, Lausanne, Swizerland, 2002, pp. 77–80.
[35] ——, "Automatic TV advertisement detection from MPEG bit-stream," *J. Pattern Recognit.*, vol. 35, no. 12, pp. 2719–2726, Dec. 2002.
[36] L. Wang, B. Zeng, S. Lin, G. Xu, and H.-Y. Shum, "Automatic extraction of semantic colors in sports video," in *Proc. IEEE ICASSP*, 2004, pp. 617–620.
[37] J.-C. Terrillon and S. Akamatsu, "Comparative performance of different chrominance spaces for color segmentation and detection of human faces in complex scene images," in *Proc. 12th Conf. Vision Interface*, vol. 2, May 1999, pp. 180–187.
[38] D. Pan, "A tutorial on MPEG/audio compression," in *Proc. IEEE Multimedia*, vol. 2, 1995, pp. 60–74.
[39] X. Sun, B. S. Manjunath, and A. Divakaran, "Representation of motion activity in hierarchical levels for video indexing and filtering," in *Proc. IEEE ICIP*, New York, 2002, pp. 149–152.
[40] L. Roberts, "Machine perception of 3-D solids," in *Optical and Electro-Optical Information Processing*. Cambridge, MA: MIT Press, 1965.
[41] T. Risse, "Hough transform for line recognition," *Proc. Computer Vision and Image Processing*, vol. 46, pp. 327–345, 1989.
[42] Y. Ishimoto, M. Unoki, and M. Akagi, "A fundamental frequency estimation method for noisy speech based on instantaneous amplitude and frequency," in *Proc. Eurospeech*, vol. 4, 2001, pp. 2439–2442.
[43] G. S. Ying, L. H. Jamieson, and C. D. Mitchell, "A probabilistic approach to AMDF pitch detection," in *Proc. Int. Conf. Spoken Language Processing*, Philadelphia, PA, Oct. 1996, pp. 1201–1204.
[44] C. Burgess, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
[45] M. Baillie, M. J. Joemon, and C. J. van Rijsbergen, "HMM model selection issues for soccer video," in *Proc. CIVR*, Dublin, Ireland, 2004, pp. 70–78.

David A. Sadlier received the B.Eng. degree in electronic engineering from the National University of Ireland, Dublin, Ireland, in June 2000, and the M.Eng. degree from Dublin City University (DCU), Dublin, Ireland, in March 2002, where he is currently working toward the Ph.D. degree.

In October 2000, he joined the Centre for Digital Video Processing (CDVP), DCU, as a full-time Researcher, working primarily in the field of digital signal processing applied to multimedia retrieval. Since June 2001, he has published eight peer-reviewed papers in international conferences and two journal papers. He has also filed an international patent based on his research work. His research interests include developing scene-level analysis techniques for sports-video summarization applications.

Noel E. O'Connor (M'99) is a Principal Investigator (PI) in the Centre for Digital Video Processing (CDVP)—an interdisciplinary University Designated Research Centre at Dublin City University, Dublin, Ireland.

The Centre consists of seven faculty members, seven postdoctoral researchers, and 23 postgraduate students. He is a Lecturer in the School of Electronic Engineering and Programme Chair of the B.Eng. in Digital Media Engineering undergraduate degree program. Since July 2000, he has published three journal papers, 33 peer reviewed papers in international conferences, and has edited two sets of conference proceedings. He has filed six international patents based on his research work. His current research interests include audio–visual analysis for knowledge extraction from digital content, region-based and object-based segmentation and feature extraction for indexing and retrieval, scene-level analysis, automatic summarization, and power-efficient hardware architectures for video processing.