A COMBINED AUDIO-VISUAL CONTRIBUTION TO EVENT DETECTION IN FIELD SPORTS BROADCAST VIDEO. CASE STUDY: GAELIC FOOTBALL

David A. Sadlier, Noel O'Connor, Sean Marlow, Noel Murphy

Centre for Digital Video Processing, Dublin City University, Ireland
sadlierd@eeng.dcu.ie

ABSTRACT

In this paper we propose novel, audio-visual analysis techniques for event detection in broadcast TV sports video content. The scope of the design is constrained to the specialized domain of 'field sport,' and specifically, Gaelic Football is presented as an experimental case study.

We will show that a combination of speech-band energy tracking in the audio domain, coupled with colour dominance pattern recognition in the video domain, provides a useful contribution to event detection for broadcast Gaelic Football matches. It is projected that, any conclusions made therein may be extended such that they function on sports content of a similar nature such as American Football, Australian Rules, Rugby Union etc.

1. INTRODUCTION

Recent developments in video compression technologies have paved the way for substantial allowances in extensive archiving of content. However the limited bandwidth availability for an online or wireless video streaming application, rooted in such archives, suggests an increasingly crucial role for highlighting or summarization of such content.

In [1], Tovinkere *et al.* present object tracking techniques for the detection of semantic highlights in Soccer matches and Cabasson *et al.* [2] tackle the same subject utilizing audio and visual cues. Methods for the automatic retrieval of semantic events from Tennis video are detailed in both Sudhir *et al.* [3] and Dayhot *et al.* [4]. Meanwhile, Zhang *et al.* [5] describe segmentation and analysis of superimposed text captions towards the automatic annotation of Baseball video. Zhou *et al.* [6] discuss methods for automatic basketball video indexing.

These works provide fairly comprehensive solutions to the tasks outlined, however the challenge of developing a solution or scheme that can reveal common structures of multiple events across multiple domains remains under-investigated. In practice though, such a scheme could not exist without some limit of domain constraint, i.e. the design of common feature extraction metrics applied to two vastly different sports types such as golf and darts would be nonsensical. On the other hand it is important to avoid becoming too context specific.

With this trade-off in mind, our research is aimed towards designing techniques such that they can be globally applied to all sports types which come under the umbrella of 'ball-and-field sport.' This is feasible since such sports (Soccer, Rugby, Gaelic Football, American Football Australian Rules etc.) all share common characteristics, some of which are listed below;

(a) Two opposing teams + referee(s).
(b) Grass pitch.
(c) Enclosed play area.
(d) Commentator voice-over.
(e) Spectator cheering.
(f) Player close-ups following significant events
(g) Objectives consistent with territorial advancement and/or directing a ball towards a specific target.

These traits, since common to all field sports, represent a good basis upon which to build low-level feature extraction metrics, which may then be exploited to provide some semantic knowledge of the content.

In this paper we describe a case study detailing the contribution made by a single particular combination of feature extractors, based on traits (b), (d), (f) and (g), towards event detection task in Gaelic Football. Since these traits are common across the scope of the domain, any deductions made should transfer effectively.

2. BACKGROUND

The Centre for Digital Video Processing (CDVP) at Dublin City University is concerned with developing technologies fundamental to the realization of efficient video content management. The current stage of research is demonstrated in the web-based digital video system, *Físchlár*, which has been used to index several video libraries. One of these libraries, *Físchlár-TV* [7], captures TV broadcast programmes and encodes them according to the MPEG-1 video standard. It then provides for efficient analyzing, browsing, and viewing of the recorded content. Currently, a user can preset the recording of programmes selected from an online TV broadcast schedule, and then navigate through the recorded material via a number of browser interfaces. Upon development, more personalized features such as news story segmentation, remote (mobile) interaction, text-based retrieval, etc. are plugged in and utilized.

3. CASE STUDY: GAELIC FOOTBALL

Ireland's national sport, Gaelic Football, is presented as a case study for the analysis discussed herein. It is a classic example of a ball-field sport and can be described as a mixture of Soccer and Rugby. A game of two 35-minute halves, a team may score in two ways (i) scoring a conventional 'goal' (as in Soccer), or (ii) scoring a 'point' by playing the ball over the crossbar and between the upright posts (as in American Football, Rugby etc.).

The objective of this analysis is to be able to discriminate these two significant events from the large amount of less consequential content, which constitutes an entire match. Our methods are based on exploiting traits (b), (d), (f) and (g) described above, and are hence split into two distinct sections, audio and visual analysis.

4. AUDIO ANALYSIS

As mentioned above, field sports such as Gaelic Football matches are typically audio-busy broadcasts, characterized by spectator cheering and enthused announcer commentary.

4.1. Assumptions and Hypothesis

The analysis makes two assumptions about the characteristics of Gaelic Football broadcast audio tracks. The first is that the variance of the overall amplitude primarily reflects the noise level exhibited by the commentator, and to a lesser extent, the attending spectators; i.e. the effect of noise sources other than commentator/spectator noise on the audio amplitude is assumed to be minimal. The second assumption is that there is a direct correlation between the variation of commentator/spectator enthusiasm and the momentary significance of the content.

These two assumptions suggest that peaks in a speech-band energy characteristic, of the audio track, may flag the camera shots associated with significant events such as goals, points or other exciting segments.

It was proposed that these peaks would be detected by thresholding a speech-band energy profile, which spanned the entire duration of the audio track. Furthermore, it was anticipated that the energy profile itself, could be efficiently established by exploiting data stripped directly from the encoded audio bitstream (MPEG-1 Layer-II).

4.2. MPEG-1 Layer-II Audio & Speech Band Energy Profile

The MPEG-1 Layer-II (mp2) compression algorithm (which *Físchlár-TV* uses) encodes audio signals as follows: the frequency spectrum of the audio signal, bandlimited to 20kHz, is uniformly divided into 32 subbands. Layer-II audio frames consist of 1152 samples; 3 groups of 12 samples from each of 32 subbands. A group of 12 samples gets a bit-allocation and, if this is non-zero, a scalefactor. Scalefactors are weights that scale groups of 12 samples such that they fully use the range of the quantiser. The scalefactor for such a group is determined by the next largest value (in a look-up table) to the maximum of the absolute values of the samples. Therefore, scalefactors, which are fundamental components of MPEG-1 Layer-II audio bitstreams, provide an indication of the maximum power exhibited by any sample within a group of 12.

Since the frequency band (0-20kHz) is divided uniformly into 32 subbands, subbands 2-7 represent the frequency range 0.625kHz – 4.375kHz. Hence, the position of these 6 subbands approximates that of speech, which is known to typically reside within the 0.1kHz - 4kHz band. Therefore, manipulation of scalefactor weights from subbands 2 through 7 should be sufficient for the establishment of an approximate speech-band energy profile.

Because it operates on compressed bitstream data, this novel approach to speech band energy tracking exhibits far superior efficiency to that of any previously known method.

5. VISUAL ANALYSIS

As with 'drop-goals' in Rugby Union, 'goals' in Australian Rules etc., scoring 'points' in Gaelic Football involves directing the ball between two target posts located in an 'endzone'.

5.1. Assumption and Hypothesis

The detection of such events assumes that the television camera follows the trajectory of the ball as it travels from the player towards the goalposts and not otherwise. Upon viewing numerous field sports, it was clear that this was an ever present trait of the production style, and hence a reasonable assumption.

Armed with this assumption, it was proposed that classification of camera shots which exhibit this type of scoring may be achieved by exploiting the natural green colour of the grass playing field. Specifically, during a point score, the camera's attention is initially concentrated on the player in the playing field, which presents a heavy monochromatic green bias, but then it's focus moves, leaving the player, and tracking the ball as it ascends towards the goalposts, whose surroundings are much less green orientated. It is proposed that recognition of this continuously decreasing fall-off pattern (from an initial high) of the grass coloured pixel ratio characteristic in the video signal, may flag such events.

If we could further classify the camera shots of the video into two additional categories (i) Distant Shot (e.g. main overhead camera) or (ii) Non-Distant Shot (e.g. close-up), it would be useful for two reasons. The first benefit would be to corroborate the apparent detection of an event; i.e. following a goal/point score, it is typical for the broadcast director to cut directly to a close-up of the player concerned. Therefore, if a non-distant shot does not immediately follow a flagged camera shot, this suggests that a false positive has been detected and it can thus be eliminated.

Secondly, presenting the user with a rapid succession of just the significant events of a given match would seem quite visually disturbing. A more agreeable, user-friendly, way of presenting this data would be to include all subsequent close-ups (non-distant shots) following an event. In this way the viewer has more time to digest each individual event and also has a chance to view this content, which typically consists of the reactions of the relevant parties.

Contrary to a non-distant shot, a distant shot would tend to capture a large area of the playing field, and hence, a significant amount of its pixel data would reside within the grass coloured hue interval. Hence, it was proposed that a rudimentary classification could be made via a simple thresholding of values representing the average grass colour content of a shot.

5.2. Grass Coloured Pixels

Grass coloured pixel samples from many different field sports were used in an experiment to find where 'grass colour' lies in the hue space. The weather conditions of the matches varied from bright and sunny to wet and muddy. The tests showed that no samples mapped outside the 60° - 105° hue interval. Intuitively, this seems correct for natural turf, since this interval maps to the brown side of pure green (=120°).

5.3. Video Sampling Rate (MPEG-1 Video Frame Encoding)

At the encoder each video frame of an MPEG-1 sequence is either independently (*intra-*) coded (I-frames), or they exhibit some form of predictive (*interframe-*) coding, i.e. a derivation based on some previously known information (P/B-frames).

It was anticipated that a mere examination of I-frames would be sufficient for our analysis. Generally, every 12th image in a video sequence is intra-coded, and these I-frames are easily and efficiently decoded. One in twelve images, at a frame rate of 25fps, corresponds to a video sampling rate of one image every half second, approximately. It was estimated that this sampling rate was sufficiently frequent, such that the transience of the grass coloured pixel ratio characteristic could be captured with acceptable precision, according to the application requirements.

6. EXPERIMENTATION

6.1. Test Subjects & Shot Boundary Detection

Físchlár-TV captured eight Gaelic Football matches, broadcast from different sports grounds under varying weather conditions, and encoded them according to the MPEG-1 standard. These matches, distinct from those in Section-5.2, comprised our experimental corpus for the said analysis.

As a preamble step, it was required to locate the boundaries of the individual camera shots within the content. For this purpose we employed our own shot boundary detection algorithm [8] and applied it to all eight test subjects.

6.2. Audio Analysis

6.2.1. Speech Band Energy Profile

For each of the eight subjects, the scalefactors associated with subbands 2-7 were stripped from the audio bitstream. A speech band energy profile with 'per video-frame' resolution was generated by a superposition of scalefactors over a window length corresponding to that of one video frame (=1/25s at 25fps).

6.2.2. Profile Thresholding & Event Detection

For each case, a threshold, $\mathbf{P_{th}}$, was initialised to the value corresponding to the largest peak found in the profile. This threshold was then gradually reduced and a peak defined if the profile exceeded $\mathbf{P_{th}}$. Via this technique, peaks were progressively detected such that temporally aligned camera shots became flagged based on significant speech band audio energy. It was decided to cap this operation once **16** individual segments had been isolated.

We had assumed that the energy exhibited by the commentator during an important event exceeds that for other times. Hence, through this analysis we expected to detect goal scores, point scores and other, non-specific, significant events.

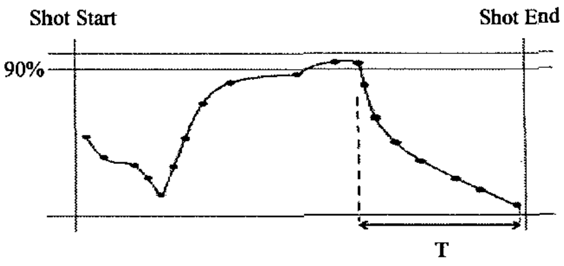

Figure-1 *GCPRC: Interpolation of I-frame data within a shot boundary. [If 1.5s<T<10s, this pattern will be recognised.]*

6.3. Visual Analysis

6.3.1. Grass Coloured Pixel Ratio Characteristic (GCPRC)

For each test subject, the region of interest of each I-frame was examined and a figure representing the I-frame grass coloured bias was calculated. This information was then interpolated over I-frames within shot boundaries, generating a temporal grass coloured pixel ratio characteristic (GCPRC) for each shot. This is illustrated in Figure-1.

6.3.2. Distant/Non-Distant Classification

A mean value for each shot's GCPRC was derived. This value was thresholded such that each shot was pigeonholed into one of two categories (i) distant shot (ii) non-distant shot, accordingly.

6.3.3. Pattern Recognition

The GCPRC 'point score pattern' was defined as a breach of a preliminary high threshold of **90%** (initial high: corresponds to strong grass colour bias e.g. player in playing field), followed by a continuous drop-off which is time constrained towards the end of the shot. See Figure-1.

The GCPRCs of each (distant) shot were examined and searched for the above pattern. If this pattern was found in a given shot and the following shot was non-distant, only then was the shot flagged.

7. RESULTS AND CONCLUSIONS

7.1. Audio/Visual Results

Table-1 presents the results of the audio and video analyses performed.

Consider the match labeled subject-A. 3 goals and 25 points were scored in total. The visual analysis retrieved 22 clips, locating 20 points (missing 5), with 2 false positives. Thus for subject-A, the visual-based retrieval has a recall value of 100*[(25-5)/25] = 80%, and a precision value of 100*[(25-5)/(25-5+2)] = 91%. For practicality, the audio-based retrieval was capped once 16 individual clips were isolated. Since precision/recall figures are nonsensical in this scenario, they are not calculated. However, we can approximately quantify the performance by calculating the ratio between tallies of content

		AUDIO RESULTS				**VISUAL RESULTS**		
	goals-points (game totals)	**# clips returned**	**goals-points Detected**	**other significant**	**insignificant detections**	**# clips returned**	**detected points (*# these also detected by audio*)**	**false positives**
A	3-25	16	3-5	6	2	22	20 (*3*)	2
B	3-27	16	2-7	4	3	26	20 (*4*)	6
C	2-24	16	2-9	2	3	28	22 (*9*)	6
D	4-24	16	2-7	3	4	22	21 (*5*)	1
E	5-19	16	3-8	3	2	22	19 (*8*)	3
F	1-17	16	1-8	3	4	17	15 (*7*)	2
G	1-22	16	1-14	1	0	21	18 (*13*)	3
H	2-25	16	2-9	3	2	24	20 (*5*)	4

Table-1 *Results of the audio/visual analyses.*

deemed significant to that deemed insignificant. In subject-A, the audio analysis returned 16 clips corresponding to 3 goals, 5 points, a further 6 events deemed significant, and 2 uneventful segments. Hence the ratio of significant content to insignificant content was 3+5+6:2 (=14:2), which is approximately 87:13 percent.

By computing these statistics for all eight subjects, and then averaging, we calculated that the visual-based retrieval had overall recall-precision figures of **85%** and **85%** respectively. For the audio analysis the overall mean ratio of significant/insignificant retrievals was **84:16** percent.

7.2. Conclusions

Of the most important events, 5-out-of-21 goal scores and 15-out-of-183 point scores were missed. A post-analysis investigation suggested reasons including the following.

Several shots containing goals were not flagged because a shot change quickly followed the goal score, hence the peaks in the speech-band energy were aligned with subsequent shots and not the actual shots of interest. Some form of time shift compensation could be used to remedy this. Occasionally a stationary camera approach was undertaken for the photography of point scores arising from dead-ball situations e.g. free kicks. This presented an unanticipated problem for the point detector system. When the camera had problems keeping up with the pace of the game, the jerky footage resulted in high numbers of false positives from the shot boundary detection algorithm. This also had a consequential effect on the detection of some events.

These faults aside, our results suggest that the audio/visual metrics designed, make a useful contribution to the task of event detection in Gaelic football, and as explained, because they are based on exploiting traits common to many other field sports, they should transfer efficiently for detection of similar events in such sports.

This work, while by no means a final solution, shows how a particular combination of selected audio-visual characteristics were exploited to contribute to event detection in field sport video. Our future work involves applying this analysis to related sports types to gauge its generalization performance, and investigating how other combinations, exploiting further common characteristics, can provide the desired increase in recall and precision.

8. REFERENCES

[1] Tovinkere, V., Qian, R.J., "Detecting Semantic Events in Soccer Games: Towards a Complete Solution," Proc. ICME 2001, Tokyo, Japan, August 2001.

[2] Cabasson, R., Divakaran, A., "Automatic Extraction of Soccer Video Highlights Using a Combination of Motion and Audio Features," Proc. Electronic Imaging (EI) 2003, Santa Clara, California, USA, January 2003.

[3] Sudhir, G., Lee, J.C.M., Jain A.K., "Automatic Classification of Tennis Video for High-Level Content-Based Retrieval," Proc. IEEE CAIVD 1998, Bombay, India, January 1998.

[4] Dayhot, R., Kokaram A., Rea, N., Denman, H., "Joint Audio-Visual Retrieval for Tennis Broadcasts," Proc. ICASSP 2003.

[5] Zhang, D., Chang, S.F., "Event Detection in Baseball Video using Superimposed Caption Recognition," Proc. ACM Multimedia 2002, Juan Les Pins, France, December 2002.

[6] Zhou, W., Vellaikal, A., Kuo, C-C.J., "Rule-based Video Classification System for Basketball Video Indexing," Proc. ACM Multimedia 2000, Los Angeles, USA, November 2000.

[7] O'Connor, N., *et al., Físchlár: An On-line System for Indexing and Browsing of Broadcast Television Content*. Proc. ICASSP 2001, Salt Lake City, UT (USA), 7-11 May 2001.

[8] O'Toole, C. *et al., Evaluation of Shot Boundary Detection on a Large Video Test Suite*, Proc. Challenges in Image Retrieval, Newcastle (UK), February 1999.

Acknowledgements: The support of both the Informatics Research Initiative of Enterprise Ireland and the Irish Research Council for Science, Engineering and Technology is gratefully acknowledged.
This material is based on work supported by the IST programme of the EU in the project *IST-2000-32795 SCHEMA*.