Pertanika J. Sci. & Technol. 21 (2): 423 - 442 (2013)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Goal Event Detection in Soccer Videos via Collaborative Multimodal Analysis

Alfian Abdul Halin[1]* and Mandava Rajeswari[2]

[1]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[2]School of Computer Sciences, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia

ABSTRACT

Detecting semantic events in sports video is crucial for video indexing and retrieval. Most existing works have exclusively relied on video content features, namely, directly available and extractable data from the visual and/or aural channels. Sole reliance on such data however, can be problematic due to the high-level semantic nature of video and the difficulty to properly align detected events with their exact time of occurrences. This paper proposes a framework for soccer goal event detection through collaborative analysis of multimodal features. Unlike previous approaches, the visual and aural contents are not directly scrutinized. Instead, an external textual source (i.e., minute-by-minute reports from sports websites) is used to initially localize the event search space. This step is vital as the event search space can significantly be reduced. This also makes further visual and aural analysis more efficient since excessive and unnecessary non-eventful segments are discarded, culminating in the accurate identification of the actual goal event segment. Experiments conducted on thirteen soccer matches are very promising with high accuracy rates being reported.

Keywords: Soccer event detection, shot view classification, sports video summarization, semantic video indexing, webcasting-text

INTRODUCTION

Technological advances have greatly enhanced broadcast, capture, transfer and storage of digital video (Tjondronegoro *et al*., 2008). Effective consumption of such huge repositories has spurred interest in automatic indexing and retrieval techniques, especially those that cater for content-based semantics (Snoek, 2005). Semantic concepts strongly rely on specific domain context. Therefore, restricting the domain being addressed is to some extent, imperative in order to bridge the semantic gap between low-level features and the inherent semantics they represent. Sports,

Article history:
Received: 20 June 2011
Accepted: 13 December 2011

E-mail addresses:
alfian@ieee.org (Alfian Abdul Halin),
mandava@cs.usm.my (Mandava Rajeswari)
*Corresponding Author

ISSN: 0128-7680 © 2013 Universiti Putra Malaysia Press.

in particular, has attracted wide interest in the past decades. Domain knowledge of specific sports has been used to extract important semantic concepts such as tennis that serves and rallies (Huang *et al*., 2009), baseball pitch trajectories (Chen *et al*., 2008), as well as basketball dunks and layouts (Changsheng *et al*., 2008b).

Soccer is one particular domain that has also received great attention, where much work focuses on event detection. Since events are meaningful and easily recalled, they are highly suitable as semantic indices (Tjondronegoro, 2005). The identified events can in turn be used to facilitate production tasks including specialized programme production, automatic video summary generation, video abstracts creation, and archival tasks such as production and posterity logging (Assfalg *et al*., 2003). However, soccer event detection is very challenging due to the unscripted recordings and the loose dynamic structure of soccer broadcasts. Lengthy running time and sparseness of event occurrences further complicate matters, where traditional VHS "rewind and fast-forward" cannot cope for timely footage access. For this reason, content-based analysis assisted by domain knowledge is required to automate and facilitate event detection.

RELATED WORKS

A great body of literature has been dedicated to events or highlights detection in soccer, as well as other sports. Basically, audio and/or visual features are firstly extracted directly from the video contents, followed by a decision making algorithm to determine whether or not a video segment contains events. Sadlier and O'Connor (2005) analyzed soccer and Gaelic Football videos. A 5-dimensional feature vector containing visual concepts such as crowd presence and motion activity and aural speech band energy to train Support Vector Machines (SVM) and to classify whether shot are eventful or uneventful. Snoek and Worring (2003) fused camera views, object classes, aural energy and textual features into a Maximum Entropy interval based algorithm to detect soccer goals, yellow cards and substitutions. The work by Chung-Lin *et al*. (2006) detects penalties, cards, goals and corner kicks using colour, motion, texture, object and audio energy from each video frame. A Dynamic Bayesian Network architecture was constructed to calculate the probability of event occurrences in the video. In their work, Jinjun *et al*. (2004) applied a hierarchical architecture where audio and visual keywords are firstly inferred from low-level features using an SVM. Left-to-right four state Hidden Markov models were then trained to detect soccer goals, goal-kicks, corner kicks and shot on goals. Meanwhile, Leonardi *et al*. (2004) used Controlled Markov Chains to detect candidate soccer goal events based on camera motion features. Ranking of each candidate was then performed based on audio loudness, where the top results contained the goal events.

All the mentioned works share three common fundamental characteristics. Firstly, event detection solely relied on features directly extracted from the video. Secondly, the event search space spanned the whole duration of the video. Thirdly, the ultimate task of event detection is carried out using supervised learning algorithms, which discovers the audiovisual patterns of specific events based on a learned model using labelled training examples. Each of the characteristics above will be further elaborated and their underlying issues will also be stated.

It is important to note that sole reliance on the directly extracted audiovisual features can lead to: (1) missed detections, and (2) confused detections. Missed detections occur when event patterns are not detected due to feature patterns being less prominent during event occurrences, such as the missed occurrences in Snoek and Worring (2003) and Jinjun *et al*. (2004) despite the use of sophisticated event models. Meanwhile, confused detections are events being misclassified as other events. For instance, Chung-Lin *et al*. (2006) reported off-sides being labelled as goals and long-passes as corner-kicks. These occurred due to the different events sharing somewhat similar audiovisual patterns. Furthermore, event boundaries are blurry and difficult to be effectively localized using audiovisual features alone.

A complete soccer game lasts for at least 90 minutes (~135,000-frames at 25-frames per second). Within this duration, the video signal is very asymmetric and noisy, where non-events[1] outnumber interesting events by an enormous margin (Min *et al*., 2007). Hence, event detection algorithms face the challenge of recognizing event patterns from the majority of non-event patterns, which is not an easy task.

As for supervised learning algorithms, their robustness may be questionable since for some soccer events, the number of positive training examples is very limited. In the work of Ren (2008), it was asserted that even with 90-hour of footage and 190 instances of soccer goals, it was still insufficient to train a classifier for robust goal event detection. They further emphasized that the sample was too small to robustly estimate any form of sequential event pattern.

Due to these reasons, it is imperative that missed and confused detections be eliminated. Furthermore, event detection could benefit from a constrained or localized search space so that algorithms could work in a more symmetric and less noisy environment. Last but not least, due to the difficulty in obtaining sufficient positive training examples for some events, unsupervised approaches should be explored. Moreover, audiovisual information in a particular video should be fully utilized instead of relying on offline training from multiple sources.

CONTRIBUTIONS

In this paper, we proposed a soccer event detection framework that circumvents the issues mentioned in the previous section. Instead of solely relying on audiovisual features that are directly extracted from the video, an external textual resource was utilized to initiate event detection. Then, a rule-based approach was used to identify potential event candidates by firstly short-listing video segments exhibiting specific visual characteristics based on semantic shot view class transitions. Finally, event segments were determined from the shortlist by a ranking procedure based on the aural feature of mean pitch. Since the framework is unsupervised, all the audiovisual considerations and assumptions were based on prior knowledge by observing ~23-hours of soccer video from various broadcasters. Furthermore, the audiovisual considerations were solely made within each video itself, and without relying on any pre-trained model.

This work offers two insights. Firstly, the issues of missed and confused detections, as well as the issue of the huge and asymmetric search space, are solved by utilizing the minute-by-minute (MBM) reports from sports websites to initiate event detection. Since MBMs contain detailed and reliable annotations of a match's progression, two crucial cues (namely, the

[1] Uninteresting portions of the video such as throw-ins, normal passing of the ball, ball out of bounds, etc.

event name and its corresponding minute time-stamp) were used. Here, missed and confused detections can be avoided since events are explicitly identified by the event name, and its time of occurrence is indicated by the time-stamp. This furthermore allows localization of the event occurrence to the one-minute video segment indicated by the time-stamp. Therefore, the event search space is significantly reduced to only the one-minute segment instead of the entire video duration. The use of MBMs is quite recent and has been shown to be effective in works such as by Changsheng *et al*. (2008a), and Changsheng *et al*. (2008b).

Secondly, since obtaining sufficient training examples for certain events is difficult (Ren, 2008), a rule-based method is proposed. It is crucial to highlight that the approach used in this study analyzed the visual and aural information only from the particular video under consideration. Therefore, no reliance is put on specific event models based on offline training. All the audiovisual considerations and assumptions are uncomplicated and therefore able to effectively identify goal event occurrences. More specifically, the authors relied on observed prior knowledge of the visual and/or aural characteristics during goal events. Initially, each shot was classified as belonging to either one of two labels, i.e., far or close-up view. These labels are consistent with camera views used during soccer broadcasts. Based on these labels, a shortlist of candidate segments exhibiting event-like characteristics is generated from within the localized one-minute video segment. Ultimately, one candidate is chosen as containing the actual goal event through a ranking procedure.

FRAMEWORK FOR GOAL EVENT DETECTION

The framework consists of four main components, which are: (1) video pre-processing, (2) textual cues utilization, (3) candidate shortlist generation, and (4) candidate ranking. An overview of the proposed framework is shown in Fig.1.

Video Pre-processing

For tractability, each video is firstly segmented into shots and assigned semantic labels. Both these steps are performed by the shot boundary detection and shot view classification steps, respectively.

Shot Boundary Detection

For videos to be processed and semantically analyzed, they firstly have to be segmented into tractable units. This process is commonly referred to as shot boundary detection (SBD), where video frame sequences taken by one continuous camera action are grouped together forming individual units called shots (Yuan *et al*., 2007). Since this work does not concern SBD, the existing algorithm by Abd-Almageed (2008) was used. This algorithm was chosen due to its robustness against variations to illumination and intensity. Basically, SBD partitions a video V into m-shots, represented as $V = \{S_i, S_{i+1}, \dots S_m\}$.

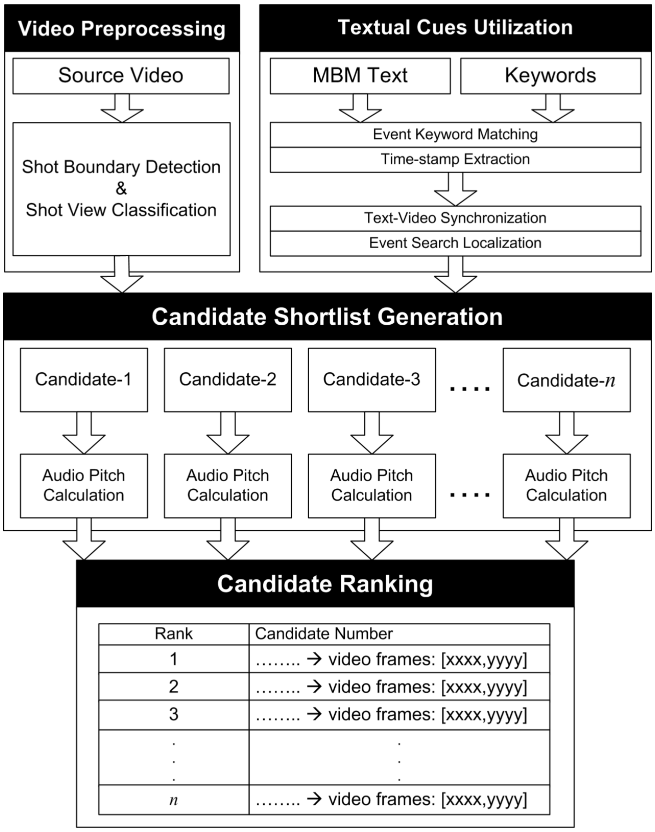

Video Preprocessing
Source Video
Shot Boundary Detection & Shot View Classification
Textual Cues Utilization
MBM Text
Keywords
Event Keyword Matching
Time-stamp Extraction
Text-Video Synchronization
Event Search Localization
Candidate Shortlist Generation
Candidate-1
Candidate-2
Candidate-3
Candidate-n
Audio Pitch Calculation
Candidate Ranking
Rank
Candidate Number
1 …….. → video frames: [xxxx,yyyy]
2 …….. → video frames: [xxxx,yyyy]
3 …….. → video frames: [xxxx,yyyy]
n …….. → video frames: [xxxx,yyyy]

Fig.1: The proposed framework

Shot View Classification

Shots alone convey no semantic meaning. Soccer videos are broadcasted with different camera shooting styles. A shot showing most of the fields normally indicates nothing interesting happening. A close-up, on the other hand, indicates something interesting that is worth paying attention to. For this reason, shots are classified into two labels, namely, *far-views* or *close-up views* (see Fig.2 and Fig.3). As will be explained in the sub-sections, these labels are used to shortlist candidate goal segments. Our shot view classification (SVC) algorithm is briefly described in the Fig.2 and 3.

1. *Dominant Hue Detection*: Firstly, all the frames are converted to the HSV (Hue, Saturation and Value) colour space and 64-bin histograms are generated for each colour component. The dominant colour is determined by the peak index idx_{peak} of the hue component. By assuming a green playfield, a dominant hue between 0.155 and 0.350 indicates the existence of predominantly green grass pixels. Note that when the dominant hues outside this range are detected, an immediate *close-up view* label can be assigned since it highly likely indicates footage outside of the playfield, or an actual close-up with little or no grass.

2. *Playfield Region Segmentation*: A range is defined to determine other pixels perceived to be similar to the dominant hue. Empirically, the range $[idx_{peak} - \alpha, idx_{peak} - \alpha]$ was determined with the optimal value for α being 0.1. All the pixels within this range are also resultantly classified as playfield pixels, if the *saturation* component is between 0.0 to 0.1, with the *value* component exceeding 0.13. For further refinement of the playfield region, morphological image processing and connected components analysis are applied (Halin *et al*., 2009).

3. *Object-size Determination and Shot Classification*: Large objects overlapping the playfield indicate *close-up views*, whereas smaller objects indicate *far views*. Some SVC techniques (e.g., Shu-Ching *et al*., 2004; Min *et al*., 2006; Xu *et al*., 2001) solely rely on the playfield ratio alone, which as suggested in Halin *et al*. (2009), can yield many incorrect labels. By calculating the largest object size, the accuracy of shot classification can improve by 8%-13% (Halin *et al*., 2009). From our experiments, a playfield with an overlapping object pixel count of more than 13,500 can be classified as a *close-up view*. A shot is then labelled as either a *close-up* or *far view* based on the majority voting of all the frame labels within the respective shot.

Textual Cues Utilization

The MBMs are obtained online from broadcasters such as ESPN (Espn, 2010) and BBC (Bbc, 2009), as well as sports information providers such as Sportinglife (2009) and UEFA (2010). An MBM of a match is annotated during a match's progression by experts. It basically logs details of the match in the granularity of minutes. Two crucial cues are considered from this

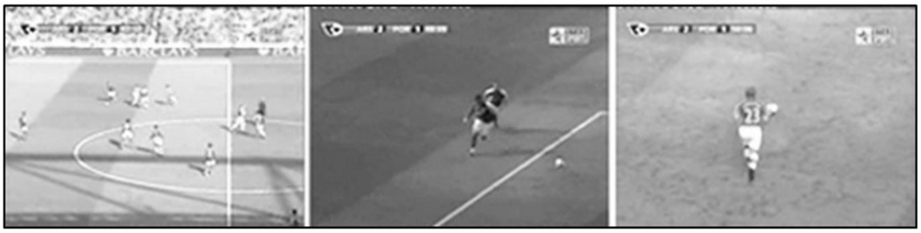


Fig.2: Far-views

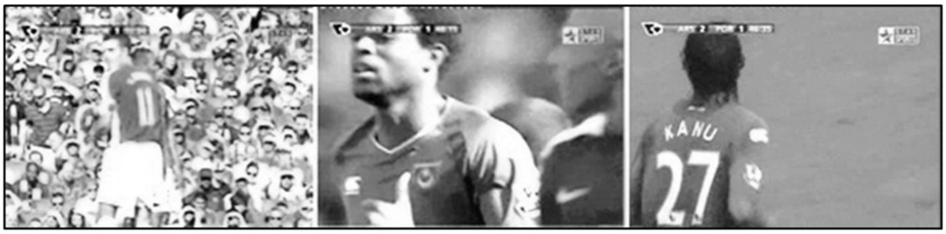


Fig.3: Close-up views

source, namely, the *event name* and its minute *time-stamp*. A screen capture of an MBM from ESPN is shown in Fig.4. These two cues from MBMs are invaluable since they eliminate guesswork in determining which event has occurred, as well as in allowing the event search to be localized to the approximate minute within the match video.

Goal Event Keyword Matching and Time Stamp Extraction

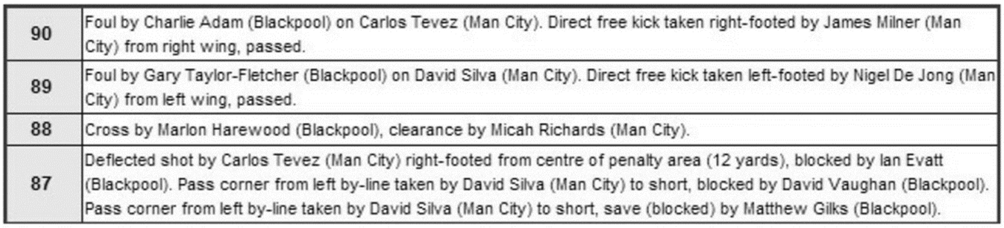
Fig.4 shows that the first column represents the minute time-stamps, whereas the match happenings are explained in the second column. Close inspection reveals that goals are annotated using dedicated keywords. After scrutinizing various sources of MBMs, a set of goal-related keywords was defined:

$$G = \{goal!, goal\ by, scored, scores, own\ goal, convert\} \tag{1}$$

Given the task of detecting a goal event *g*, that has *i*-number of occurrences: If matching keywords are found within the MBM, the *time-stamp* of each of the *i*-th occurrences are noted. These can be written as a set $T^g = t_i^g$, where $i > 0$ if there is at least one goal event in the match. Then, for each *i*, the goal event search is initiated within the one-minute segment of each t_i^g .

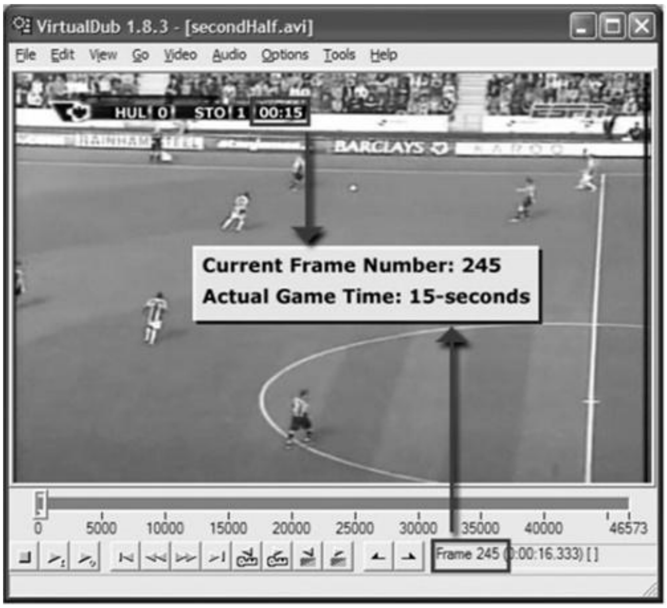
Text-Video Synchronization and Event Search Localization

The time-stamp t_i^g indicates the minute within which the event has occurred. However, directly mapping to the corresponding video frames can be erroneous due to the misalignment with the actual game time. Due to this, synchronization between the t_i^g and its corresponding video frame is necessary. A manual step is introduced here, where a reference frame is determined by matching a frame number to the actual elapsed game time of a match. This process is illustrated in Fig.5, where an elapsed game time of 0.25-minutes (15-seconds) corresponds with the 245-th frame of the match video. Each of these values is denoted as t^{ref} and f^{ref} , respectively. Although the authors are aware that the process can be automated using the technique proposed by Chansheng (2008a), some matches fail to display superimposed game clocks, therefore this particular technique cannot be used. We argue that bit of automation needs to be sacrificed for the sake of reliability.



90	Foul by Charlie Adam (Blackpool) on Carlos Tevez (Man City). Direct free kick taken right-footed by James Milner (Man City) from right wing, passed.
89	Foul by Gary Taylor-Fletcher (Blackpool) on David Silva (Man City). Direct free kick taken left-footed by Nigel De Jong (Man City) from left wing, passed.
88	Cross by Marlon Harewood (Blackpool), clearance by Micah Richards (Man City).
87	Deflected shot by Carlos Tevez (Man City) right-footed from centre of penalty area (12 yards), blocked by Ian Evatt (Blackpool). Pass corner from left by-line taken by David Silva (Man City) to short, blocked by David Vaughan (Blackpool). Pass corner from left by-line taken by David Silva (Man City) to short, save (blocked) by Matthew Gilks (Blackpool).

Fig.4: An expert of a minute-by-minute report from ESPN


VirtualDub 1.8.3 - [secondHalf.avi]
File Edit View Go Video Audio Options Tools Help
HUL 0 STO 1 00:15
BARCLAYS
0 5000 10000 15000 20000 25000 30000 35000 40000 46573
Frame 245 (0:00:16.333) [ ]

Current Frame Number: 245
Actual Game Time: 15-seconds

Fig.5: Manual determination of the reference frame and time

The values t^{ref} and f^{ref} can then be used to localize the event search to the one-minute eventful segment. With t_i^g being the minute time-stamp of the goal event, the beginning ($f_{i,begin}^g$) and ending ($f_{i,end}^g$) frames for the event minute can be determined via:

$$f_{i,begin}^g = \left\lfloor \left| f^{ref} - \left(fr\left(t^{ref} + 60\left(t_i^g - 1 \right) \right) \right) \right| \right\rfloor \tag{2}$$

$$f_{i,end}^g = \left\lfloor f_{i,begin}^g + 60fr \right\rfloor \tag{3}$$

where fr is the video frame rate and $\lfloor \cdot \rfloor$ rounds-down the calculations to the nearest integer. Note that for $f_{i,begin}^g$, the time-stamp t_i^g (after being converted to seconds) is subtracted by 60-seconds since the actual event occurs between the minute range of $\left[t_i^g - 1, t_i^g \right]$. Note also that for $f_{i,end}^g$ end, fr has been multiplied by 60 (seconds) in order to position the end boundary at one-minute after $f_{i,begin}^g$. Consequently, the localized one-minute event search space is:

$$\Upsilon_i^g = \left[f_{i,begin}^g, f_{i,end}^g \right] \tag{3}$$

Candidate Shortlist Generation

Returning the whole one minute segment would be too coarse since a goal's conception to finish is normally shown within a very short and condensed time period. Therefore, potential goal segments within Υ_i^g need to be identified. During goal events, certain visual properties can be observed. By observing more than ~23-hours of soccer video footage from various broadcasters (including the footage used in this paper), three generic visual-related properties were identified, and when present, they would highly likely indicate a potential goal event. These properties were exploited to decompose the one-minute segment into a shortlist of candidate segments, which could be explained as:

1. The camera transitions from a *far-view* to a *close-up view*. The former is meant to capture the build-up, whereas the latter focuses on player/crowd/coach reactions;
2. *Close-up views* during goals normally last at least 6-seconds;
3. It takes approximately 12-seconds to fully observe an event's progression from conception to finish.

Note that these properties can be quite generic. In other words, they can also indicate other event occurrences such as yellow cards or red cards. However, since the search space has already been localized to the one-minute eventful segment, detecting other events is very unlikely. Moreover, it is already known from the MBM that this particular minute contains a goal event.

From these properties, the relevant 12-second segments are extracted as shortlisted candidates. Note that the transition point between the shot views serves as the mid-point, where preceding and superseding segments have equal lengths of 6-seconds. Consequently, the *n*-number of candidate segments is generated from Υ_i^g:

$$C_i^g = \left\{c_{ik}^g\right\} \quad for\ k = 1, \ldots, n \tag{4}$$

where $C_i^g \subset \Upsilon_i^g$ is the set containing the shortlisted candidates, and c_{ik}^g s the *k*-th 12-second candidate segment within C_i^g. An illustration is shown in Fig.6, where two candidate segments have been shortlisted for a goal event.

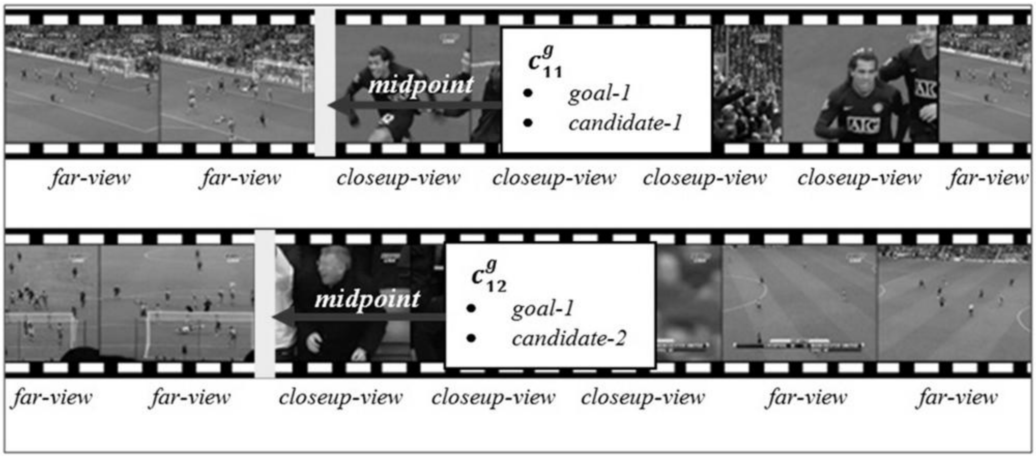


Fig.6: Example of two shortlisted candidates

Candidate Ranking

At this stage, we have obtained the candidate segments c_{ik}^{g}, where one of them is the actual goal event. Since the shot view transitions are similar across the candidates, another feature is needed from each c_{ik}^{g}. It was observed that whenever a goal is scored, there would be an increase in the commentator's excitement and intonation. Since the commentator can be considered a neutral party during any soccer match, his/her excited speech is less biased to any one team. The works by Tjondronegoro (2005), and Coldefy and Bouthemy (2004) demonstrated that pitch or the fundamental frequency of an audio signal $f0$ is reliable to detect excited human speech. Normally, as a direct result of speech excitement, $f0-$ measurements will increase. For the purpose of this work, the sub-harmonic to harmonic ratio analysis technique was chosen due to its insensitivity towards noise and prominent unvoiced segments. The algorithm used is called *shrp.m*, which is a MATLAB implementation of Xuejing (2002), and is available from Sun (2008).

For audio temporal segmentation, 20-milisecond (ms) frames and 500-ms clips were used. The value of 20-ms was chosen since audio signals could be assumed to be (pseudo) stationary within the 10 to 40-ms range (Yao, 2000), and that 20-ms seemed to be appropriate to capture a snapshot of the evolving audio signal for the data in this study. The clip size of 500-ms, on the other hand, was chosen as it managed to accurately capture the average measurement within a longer time window.

Basically, $f0$ is initially calculated at the frame level. The frame level results are then combined to obtain an averaged value at the clip level. Finally, all the clips' average $f0$ values within the respective 12-second segments were averaged to obtain the mean pitch $\overline{f0}_{ik}^{g}$ for each candidate.

The rule being applied here is that the candidate with the goal event will cause commentator speech to be very pronounced. This will result in high *f0* values across audio frames, leading to high $\overline{f0}_{ik}^{g}$ within the segment. It can then be argued that segment c^*, with the maximum $\overline{f0}_{ik}^{g}$, contains the actual goal event.

EXPERIMENTAL RESULTS AND DISCUSSION

The experiments were conducted to evaluate the effectiveness of the SVC algorithm, shortlist generation and goal segment identification processes. A total of thirteen soccer matches, consisting of European Champions League, Barclay's Premier League and Italian Serie-A matches, were used as test data. All the videos were in AVI format at 15-frames per second. The audio channel was MP3-mono, sampled at 22.5-KHz with a bit-rate of 32-kbps. Each of the match halves was recorded in a separate file, with non-game footage such as halftime commentaries, commercial breaks, match highlights, etc. excluded.

It is worth noting that, to the best of our knowledge, there is no publicly available dataset(s) for soccer video. Therefore, we have resorted to recorded matches from television broadcasters. Details for each of the recordings are available in Table 1. Note that some matches managed to be fully recorded (i.e. 90-minutes), whereas other were partially recorded (i.e. 45-minutes). The partial recordings were due to equipment setup failures during the recording process.

TABLE 1
Description of the dataset. ECL – European Champions League, BPL – Barclay's Premiere League and ISA – Italian Serie – A

Match number	Teams	League	Duration
1	Chelsea vs. Man. United	ECL	90-mins
2	Man. United vs. Blackburn	BPL	90-mins
3	Man. United vs. Liverpool	BPL	90-mins
4	Newcastle vs. Fulham	BPL	45-mins
5	Stoke vs. Hull	BPL	90-mins
6	West Brom vs. Liverpool	BPL	90-mins
7	West Ham vs. Liverpool	BPL	45-mins
8	Barcelona vs. Man. United	ECL	45-mins
9	Arsenal vs. Celtic	ECL	90-mins
10	Arsenal vs. Portsmouth	BPL	90-mins
11	Inter Milan vs. Bari	ISA	90-mins
12	Bolton vs. Liverpool	BPL	90-mins
13	Everton vs. Wigan	BPL	90-mins
		TOTAL	1035-mins

Note that for the following sub-sections of *Shot View Classification* and *Candidate Shortlist Generation*, the evaluation criteria used are *precision* and *recall*. Although sharing the same terminologies, the contexts within which both sections are evaluated are different. Sub-section *Shot View Classification* is basically a classification task, whereas Sub-section *Candidate Shortlist Generation* is a retrieval task. Therefore, two sets of formulas are presented to cater for each of these contexts, which will be further explained in detail within each of the respective sections.

Shot View Classification

In this work, SVC was tested on different subsets of our data, totalling 599-shots. Different subsets from different matches were used to demonstrate the robustness of the algorithm across different broadcasters. *Precision* and *recall* are calculated using Equations 5 and 6, respectively. SVC is treated as a two-class classification task. For clarity, *true positives*, *false positives* and *false negatives* are explained, supposing that the positive class being predicted is a *far-view*:

- *True positive*: Assigning a shot as *far-view*, when the actual class is indeed a *far-view*;
- *False positive*: Assigning a shot as *far-view*, when the actual class is a *close-up view*;
- *False negative*: Assigning a shot as *close-up view*, when the actual class is a *far-view*.

The Table 2 shows the results obtained by the SVC algorithm.

$$Precision = \frac{\#\ true\ positives}{\#\ true\ positives + \#\ false\ positives} \tag{5}$$

$$Recall = \frac{\#\ true\ positives}{\#\ true\ positives + \#\ false\ negatives} \tag{6}$$

The results are encouraging where very high *recalls* are reported. This is very important since further visual processing requires that each shot be properly labeled.

TABLE 2
Average Precision and Recall for Shot View Classification

Shot Type	# of shots	Data	Precision	Recall
Close-up view	415	Match-1	94.29%	98.51%
		Match-2	100.00%	100.00%
		Match-3	98.46%	98.46%
		Match-4	100.00%	92.00%
		Match-5	100.00%	93.24%
		Match-6	96.88%	95.38%
		Average	**98.27%**	**96.27%**
Far-view	184	Match-1	96.56%	87.50%
		Match-2	100.00%	100.00%
		Match-3	97.14%	97.14%
		Match-4	80.65%	100.00%
		Match-5	83.87%	100.00%
		Match-6	91.67%	94.29%
		Average	**91.65%**	**96.49%**

Candidate Shortlist Generation

The total number of goals from the thirteen matches was 36 altogether. In all, 84-candidates c_{ik}^{g} were generated for each time-stamp. Similar to the previous section, *precision* and *recall* are calculated. However, the calculations in this section are from the context of retrieval. Therefore, the formulas differ from that of the previous section, and are calculated using Equations 7 and 8, respectively. *Relevant* refers to the number of candidate segments generated that actually contain goal events. *Retrieved* refers to the total number of candidates generated based on the assumption of the changing camera views.

$$Precision = \frac{\#\ relevant \cap \#\ retrieved}{\#\ retrieved} \tag{7}$$

$$Recall = \frac{\#\ relevant \cap \#\ retrieved}{\#\ retrieved} \tag{8}$$

Table 3 shows the overall performance of the candidate segment generation process. It can be observed that the *Average Number of Candidates per Shortlist* is quite low (i.e. 2). This value indicates the average number of candidates being identified for each event occurrence.

Preferably, this value should be as low as possible so that in case ranking errors occur (i.e. Sub-section *Candidate Ranking*), the actual segment can still be retrieved without the need for extensive browsing.

Overall, the most important measurement is *recall*. It is crucial that this be 100% for all cases since it is mandatory that an actual goal event segment be present within each of the candidate segments for a particular minute. If for instance, a candidate set does not contain an actual goal event segment, the final step of candidate ranking (i.e. Sub-section *Candidate Ranking*) would be in vain as the top-ranked candidate would never be a goal event (and hence, be missed).

TABLE 3
Segment Generation Performance

Ground Truth	36
Relevant	36
Retrieved	84
Missed	0
Average Number of Candidates per Shortlist	2
Precision	42.86%
Recall	100.00%

Candidate Ranking

Table 4 shows the results of the candidate ranking process, where ideally, the candidate ranked top most will contain the actual goal event. The first column shows the respective matches whereas the second records the time-stamps of each goal event. The variable *n* represents the number of candidate segments c_{ik}^{g} identified for each of the *i-th* goal instance. The mean pitch $\overline{f0}_{ik}^{g}$ for each *k-th*. segment (for $k = 1, ..., n$) is recorded in the sub-columns of column 5, where the numeric boldface values indicate the maximum $\overline{f0}_{ik}^{g}$ of that time-stamp, which is the top-ranked candidate within that particular shortlist. The final column (Decision) indicates whether the top-ranked candidate indeed contains the goal event. In all, 35 goal segments were ranked top most. This high percentage is very promising and shows that the modality collaboration technique is reliable at identifying goal events. The final row indicates one actual goal event segment being wrongly ranked second (i.e. Decision value **X**). Close observation showed that it was a goal resulting from a penalty kick. For this particular occurrence, the commentator's reaction was louder during the foul that led to the penalty, than when the actual goal was scored.

TABLE 4
Results of Candidate Ranking. The ‘Match number’ in column 1 corresponds to the same matches shown in TABLE 1

Match number	Num. of goals	T^g	N	Mean pitch of candidate segments ($\overline{f0}^g_{ik},...\overline{f0}^g_{in}$)				Decision
1	3	$t^g_1 = 46$	3	**285.55**	254.08	255.52		√
		$t^g_2 = 56$	2	246.87	**290.08**			√
		$t^g_3 = 86$	3	238.23	247.62	**271.64**		√
2	1	$t^g_1 = 90$	2	**301.55**	273.28			√
3	5	$t^g_1 = 23$	3	262.38	245.55	**264.59**		√
		$t^g_2 = 28$	1	**275.99**				√
		$t^g_3 = 44$	2	257.88	**273.13**			√
		$t^g_4 = 77$	3	239.41	245.73	**267.05**		√
		$t^g_5 = 91$	3	230.30	**293.08**	260.71		√
4	1	$t^g_1 = 41$	2	**305.79**	285.74			√
5	2	$t^g_1 = 73$	2	248.68	**284.33**			√
		$t^g_2 = 95$	3	234.65	250.76	**281.00**		√
6	2	$t^g_1 = 28$	2	268.31	**299.02**			√
		$t^g_2 = 63$	2	276.63	**277.35**			√
7	2	$t^g_1 = 02$	3	230.00	**290.87**	237.45		√
		$t^g_2 = 38$	2	251.61	**286.89**			√
8	1	$t^g_1 = 10$	2	290.96	**295.47**			√
9	4	$t^g_1 = 28$	3	**279.46**,	268.85	263.39		√
		$t^g_2 = 53$	1	**283.63**				√
		$t^g_3 = 74$	2	272.63	**274.09**			√
		$t^g_4 = 92$	2	248.21	**278.12**			√
10	5	$t^g_1 = 18$	2	**298.86**	261.12			√
		$t^g_2 = 21$	2	228.40	**301.02**			√
		$t^g_3 = 37$	4	217.25	215.02	**279.22**	211.02	√
		$t^g_4 = 51$	2	224.77	**300.38**			√
		$t^g_5 = 68$	2	212.52	**298.66**			√
11	2	$t^g_1 = 56$	3	280.48	**281.59**	279.54		√
		$t^g_2 = 74$	*2*	228.14	**295.39**			√

TABLE 4 (continue)

12	5	$t_1^g = 33$	3	243.40	239.67	**281.84**		√
		$t_2^g = 41$	3	264.61	**272.70**	265.61		√
		$t_3^g = 47$	3	238.52	**273.82**	242.51		√
		$t_4^g = 56$	3	284.34	**306.16**			√
		$t_5^g = 83$	2	213.75	**280.70**			√
13	3	$t_1^g = 57$	3	252.49	**300.00**			√
		$t_2^g = 62$	2	266.33	**270.98**			√
		$t_3^g = 93$	2	286.75	**293.71**			**X**

COMPARISON

The proposed framework was compared against three works to detect 12-goals from three randomly selected soccer matches taken from the same dataset. These works were selected due to their similarities in feature usage and event modelling (i.e., rule-based and/or unsupervised). Explanations regarding these works are as follows:

- Ekin *et al*. (2003) - (CT): Goals occur during game breaks lasting between 30 and 120 seconds. Throughout this duration, there must be at least one close-up and one replay shot, and the replay must directly follow the close-up shot;

- Eldib *et al*. (2009) - (CT2): Goals are detected based on a replay analysis. They further impose the rule that goals occur when replays last between 16 and 40-seconds;

- Yina *et al*. (2008) - (FCM-HR): Goals exhibit relatively lengthier close-up and replay durations. Data points representing the combination of the two durations are clustered using the fuzzy C-means (FCM) algorithm, where goals are expected to exhibit significant inter-cluster separability . The final goal event clusters are determined via specific conditions based on the known number of goals in a particular match.

Note that the approach proposed in this paper only considers two shot labels; the *far* and *close-up* views. The three other approaches, however, require the additional label of *replay* shots. Therefore, each shot corresponding to a replay had to be relabelled accordingly for the three matches, where the entire process was done manually.

From the results in Table 5, CT, CT2 and the proposed framework were able to obtain perfect recall scores, indicating that all actual goal segments were accounted for. FCM-HR, however, missed two goals for Match-2 (i.e. 60% recall), and this was due to the failure of the assumption that the goals and non-goals would exhibit clear inter-cluster separability. Precision-wise, the proposed work was able to achieve perfect scores for all the matches. As for the other works, precision was imperfect since certain segments sharing similar feature properties for goal events were also retained. This occurred mainly because the search space was unconstrained (i.e., the entire video duration), resulting in the detection of non-goal segments as well. As can be seen, this problem can be alleviated (and in this case, eliminated) by considering the time-stamp textual cue.

CT, CT2 and FCM-HR returned relevant frame ranges containing the goal events. However, event boundaries were not specified and this resulted in lengthy clip durations, including the replay segments. The proposed framework benefits from the specification of the 12-second time window, where although manually determined, is able to show relatively sufficient footage from the goal's conception to finish. Eventful frame boundaries, however, are very subjective and depend on viewers' preferences (Changsheng *et al*., 2008a). Therefore, although the proposed framework does return succinct event segments, further user study is still necessary to determine ideal frame boundaries.

TABLE 5
Comparisons of the goal event detection. The 'match number' in column 1 corresponds to the same matches shown in TABLE 1

Match number	Work	Ground Truth	Relevant	Irrelevant	Missed	Precision	Recall
5	CT	2	2	4	0	33.33%	100.00%
	CT2		2	3	0	40.00%	100.00%
	FCM-HR		2	4	0	33.33%	100.00%
	Proposed framework		2	0	0	100.00%	100.00%
10	CT	5	5	2	0	71.43%	100.00%
	CT2		5	2	0	71.43%	100.00%
	FCM-HR		3	7	2	30.00%	60.00%
	Proposed framework		5	0	0	100.00%	100.00%
12	CT	5	5	8	0	38.46%	100.00%
	CT2		5	3	0	62.50%	100.00%
	FCM-HR		4	28	0	12.50%	100.00%
	Proposed framework		5	0	0	100.00%	100.00%

CONCLUSION

In this work, a multimodal collaborative framework was proposed for goal event detection. The framework uses event names and time-stamps extracted from minute-by-minute reports to distinctly identify event occurrences and to localize the video search space to only relevant and eventful portions, respectively. Shot labels analysis, which is used as the basis for candidate eventful segments shortlisting, was also carried out. A ranking step was finally performed based on the event-specific signatures, where the top-ranked candidate is most likely to contain the actual event occurrence. In all, the use of MBMs greatly facilitates event detection by decreasing the guesswork of locating specific events. Besides preventing misses and confusions, it also enables processing of only relevant portions of the video. This further allows simple assumptions to be made about events' visual and aural patterns, which could fit into an effective rule-based framework. Moreover, since all the relevant segments have been retained during shortlist generation, all the desired events were accounted for. Ultimately, although one ranking error has been reported, the framework ensures that eventful footage can still be accessed in a timely fashion since on average only 2 candidates are generated for each shortlist. In the future, this work will be improved by considering more robust techniques to extract textual resources from more than one provider. In addition, we would also like to detect other important events such as free-kicks on goal, shots on goal, saves, penalties and own-goals.

REFERENCES

Abd-Almageed, W. (2008). *Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing*. Paper presented at the 15th IEEE International Conference on Image Processing, pp. 3200-3203.

Assfalg, J., Bertini, M., Colombo, C., Del Bimbo, A., & Nunziati, W. (2003). Semantic annotation of soccer videos: automatic highlights identification. *Computer Vision and Image Understanding, 92,* 285-305.

BBC. (2009). *Sport news images*. BBC. Retrieved from http://newsimg.bbc.co.uk/sport1/hi/football/teams/a/arsenal/livedtext/default.stm?refresh.

Changsheng, X., Wang, J., Lu, L., & Zhang, Y. (2008a). A Novel Framework for Semantic Annotation and Personalized Retrieval of Sports Video. *IEEE Transactions on Multimedia, 10,* 421-436.

Changsheng, X., Yi-Fan, Z., Guangyu, Z., Yong, R., Hanqing, L., & Qingming, H. (2008b) Using webcast text for semantic event detection in broadcast sports video. *IEEE Transactions on Multimedia, 10,* 1342-1355.

Chen, H. T., Chen, H. S., Hsiao, M. H., Tsai, W. J., & Lee, S. Y. (2008). A trajectory-based ball tracking framework with visual enrichment for broadcast baseball videos. *Journal of Information Science and Engineering*, *24,* 143-157.

Chung-Lin, H., Huang-Chia, S., & Chung-Yuan, C. (2006). Semantic analysis of soccer video using dynamic Bayesian network. *IEEE Transactions on Multimedia, 8,* 749-760.

Coldefy, F., & Bouthemy, P. (2004). *Unsupervised soccer video abstraction based on pitch, dominant color and camera motion analysis*. In the 12th Annual ACM International Conference on Multimedia*,* pp. 268-271.

Ekin, A., Tekalp, A. M., & Mehrotra, R. (2003). Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing, 12,* 796-807.

Eldib, M. Y., Zaid, B., Zawbaa, H. M., El-Zahar, M., & El-Saban, M. (2009). *Soccer video summarization using enhanced logo detection*. In the 16th IEEE International Conference on Image Processing, pp. 4345-4348.

ESPN Soccernet. (2010). Retrieved from http://soccernet.espn.go.com/commentary?id=244509\&cc=4716\&league=eng.1.

Halin, A. A., Rajeswari, M., & Ramachandram, D. (2009). *Shot view classification for playfield-based sports video.* In the IEEE International Conference on Signal and Image Processing Applications ICSIPA09, pp. 410-414.

Huang, Y.-P., Chiou, C.-L., & Sandnes, F. E. (2009). An intelligent strategy for the automatic detection of highlights in tennis video recordings. *Expert Systems with Applications, 36,* 9907-9918.

Jinjun, W., Changsheng, X., Engsiong, C., & Qi, T. (2004). *Sports highlight detection from keyword sequences using HMM*. Paper presented at the IEEE International Conference on Multimedia and Expo, pp. 599-602.

Leonardi, R., Migliorati, P., & Prandini, M. (2004). Semantic indexing of soccer audio-visual sequences: A multimodal approach based on controlled Markov chains. *IEEE Transactions on Circuits and Systems for Video Technology, 14,* 634-643.

Min, C., Shu-Ching, C., & Mei-Ling, S. (2007). *Hierarchical temporal association mining for video event detection in video databases.* Paper presented at the 23rd IEEE International Conference on Data Engineering Workshop, pp.137-145.

Min, C., Shu-Ching, C., Mei-Ling, S., & Wickramaratna, K. (2006). Semantic event detection via multimodal data mining. *IEEE Signal Processing Magazine, 23,* 38-46.

Ren, R. (2008). *Audio-visual football video analysis, from structure detection to attention analysis.* University of Glasgow.

Sadlier, D. A., & O'Connor, N. E. (2005). Event detection in field sports video using audio-visual features and a Support Vector Machine. *IEEE Transactions on Circuits and Systems for Video Technology, 15,* 1225-1233.

Shu-Ching, C., Mei-Ling, S., Min, C., & Chengcui, Z. (2004). *A decision tree-based multimodal data mining framework for soccer goal detection.* Paper presented at the IEEE International Conference on Multimedia and Expo, pp. 265-268.

Snoek, C. G. M. (2005). *The authoring metaphor to machine understanding of multimedia*. University of Amsterdam.

Snoek, C. G. M., & Worring, M. (2003). *Time interval maximum entropy based event indexing in soccer video.* Paper presented at the International Conference on Multimedia and Expo, pp.481-484.

Sportinglife. (2009). *Live Match*. Retrieved from ttp://www.sportinglife.com/football/live_match/200111.html.

Sun, X. (2008). *Matlab central - file detail - pitch determination algorithm*. Retrieved on 22 October 2010 from http://www.mathworks.com/matlabcentral/fileexchange/1230-pitch-determination-algorithm.

Tjondronegoro, D. W. (2005). *Content-based video indexing for sports applications using multi-modal approach*. Deakin University.

Tjondronegoro, D. W., Yi-Ping Phoebe, C., & Adrien, J. (2008). A scalable and extensible segment-event-object-based sports video retrieval system. *ACM Transactions on Multimedia Computing, Communications, and Applications, 4,* 1-40.

UEFA. (2010). *Uefa champions League, Match Season 2011.* Retrieved from http://www.uefa.com/uefachampionsleague/matchess/season=2011/live/index.html?matchday=3\&day=2\&match=2002855.

Xu, P., Lexing, X., Shih-Fu, C., Divakaran, A., Vetro, A., & Huifang, S. (2001). *Algorithms and system for segmentation and structure analysis in soccer video.* Paper presented at the IEEE International Conference on Multimedia and Expo, pp.721-724.

Xuejing, S. (2002). *Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio.* Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.333-336.

Yao, W., Zhu, L., & Jin-Cheng, H. (2000). Multimedia content analysis-using both audio and visual clues. *IEEE Signal Processing Magazine*, *IEEE*, *17*, 12-36.

Yina, H., Guizhong, L., & Chollet, G. (2008). *Goal event detection in broadcast soccer videos by combining heuristic rules with unsupervised fuzzy c-means algorithm.* Paper presented at the 10th International Conference on Control, Automation, Robotics and Vision, pp.888-891.

Yuan, J., Wang, H., Xiao, L., Zheng, W., Li, J., Lin, F., & Zhang, B. (2007) A formal study of shot boundary detection. *IEEE Transactions on Circuits and Systems for Video Technology, 17,* 168-186.