Soccer Event Detection via Collaborative Multimodal Feature Analysis and Candidate Ranking

Alfian Abdul Halin[1], Mandava Rajeswari[2], and Mohammad Abbasnejad[3]
[1]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia
[2]School of Computer Sciences, Universiti Sains Malaysia, Malaysia
[3]College of Engineering & Computer Science, Australian National University, Australia

Abstract: *This paper presents a framework for soccer event detection through collaborative analysis of the textual, visual and aural modalities. The basic notion is to decompose a match video into smaller segments until ultimately the desired eventful segment is identified. Simple features are considered namely the minute-by-minute reports from sports websites (i.e. text), the semantic shot classes of far and closeup-views (i.e. visual), and the low-level features of pitch and log-energy (i.e. audio). The framework demonstrates that despite considering simple features, and by averting the use of labeled training examples, event detection can be achieved at very high accuracy. Experiments conducted on ~30-hours of soccer video show very promising results for the detection of goals, penalties, yellow cards and red cards.*

Keywords: *Soccer event detection, sports video analysis, semantic gap, webcasting text.*

Received August 20, 2011; accepted December 30, 2011

1. Introduction

Technological advances have caused a boom in the broadcast, capture, transfer and storage of digital video. Optimizing consumption of such huge repositories has spurred great interest in automatic indexing and retrieval techniques, where the most effective is through content-based semantics [21]. Semantic concepts strongly rely on specific domain context. Therefore, restricting the domain being addressed can help narrow down the semantic gap between low-level features and the semantics they inherently represent. Sports in particular, have attracted wide interest where interesting works have been reported in domains such as tennis [9, 26], baseball [6, 20, 27] and basketball [23, 28]. In this paper, we propose a framework for soccer event detection utilizing the textual, visual and aural modalities in a collaborative fashion. Generally, the match video is decomposed into a shortlist of candidate segments, where the actual eventful segment is finally identified via a ranking process. Evaluations are based on the measurements of precision and recall.

2. Related Works in Soccer Event Detection

Two main steps are commonly involved to detect soccer events. Firstly, the appropriate features are selected to represent the audio/visual content evolution surrounding the events. This is followed by the event detection process where features are used to build event models.

2.1. Semantic Feature Extraction

Recent trends mostly rely on semantic-level features (SLF) as opposed to their low-level counterparts. Examples of visual SLFs are semantic shot classes [15, 17], object classes such as the referee, players and the ball [20], playfield positions and player deployments [2, 10], and the camera motion parameters of pan, tilt and zoom [2]. Aural SLFs on the other hand commonly relate to the type of referee sound (e.g. long and short whistling), and commentator or crowd sounds such as excited/plain commentator speech and crowd cheering [18, 24].

Both the visual and aural SLFs are derived from low-level features. Camera zoom for instance is determined through MPEG motion vector analysis [2]. Plain and exciting commentator speeches are derived from the audio signal's mel-frequency cepstral coefficients (MFCC), Linear-Prediction Coefficients (LPC) and Zero Crossing Rates (ZCR) [24]. Feature representations at the semantic level can greatly simplify event modeling since reliance on arrays of low-level numerical values is not required.

2.2. Event Modeling

The derived SLFs are incorporated into event models, which are commonly rule-based or built upon supervised machine learning.

Spatio-temporal sequential or simultaneous SLFs instances surrounding event occurrences can be

defined using rule-sets. Templates following an IF-THEN structure were used in [10, 11]. Both detected events such as goals and fouls by isolating video segments containing semantic shot classes of specific durations. In [11], fouls were detected when replays lasted between 4 and 15 seconds. In [24], aural templates were used to detect shots/saves/passes when excited commentator speech and crowd sounds occurred simultaneously. Alternatively, graphical models such as Finite State Machines (FSM) are also used. FSMs treat event detection as a logical rule fulfillment process. If a series of SLF observations are successfully made, this means that the event has occurred and hence detected [2].

Other than rules, event detection is also popularly performed using supervised machine learning. The most popular by far is supervised classification. Basically, given labeled training examples $\{X_n, Y_n\} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, the algorithm learns a function $f : X \rightarrow Y$ that predicts future input-output mappings [4]. X and Y are the input and output spaces, respectively, which comes in the form of SLFs. The output hence is the desired event class. Provided sufficient training examples, supervised learning can be very effective. Examples of popular learning algorithms are such as Hidden Markov Models [23], Dynamic Bayesian Networks [7] and Support Vector Machines [20]. The clear advantage of supervised learning is that feature patterns are automatically discovered without the need for hand crafted rules.

3. Problem Statement

3.1. Issues Regarding SLF Identification

Having many SLFs can assist in accurately representing event models. However, identifying or deriving them can be cumbersome. In [10, 20], various visual SLFs were derived via multi-level image processing and threshold-based tasks, such as edge detection and texture analysis. In [15], the authors derived 17 visual concepts such as replays, team players and camera views using a hierarchical classification tree. Most of the concepts necessitated color-based threshold comparisons at multiple levels. The work in [16] identified six playfield positions via edge detection, projection profile analysis, and region of interest color, shape and texture analyses. Each position was then classified using hierarchical SVMs, assisted by heuristic-rules. The case is similar for aural SLFs where in [3, 24], 3-levels of SVMs were trained to derive excited and plain commentator/crowd sound. As can be seen, complex processes are required to derive the SLFs. In some cases (e.g. [16]), erroneous feature sets were generated due to the hierarchical decision making structure. Similarly, the work in [3] reported error rates between 20%-26% for aural SLF classification. Having erroneous features can be detrimental to the accuracy and reliability of event models.

3.2. Issues Regarding Event Modeling

Rule-based approaches have the advantage of being simple in terms of rules insertion, deletion and modification. However, conditions pertaining to feature duration and threshold values might not be exclusive to only one event [14]. Consequently, considering that detection algorithms have to wade through entire video durations, many false positives can be detected, resulting in low detection precision. Moreover, defining different rule-sets for different events is cumbersome, especially when many events are considered. Therefore, rule sets alone, especially in unconstrained search spaces, can be less effective.

In soccer, non-events such as dribbling, throw-ins and goal-kicks hugely outnumber interesting events [17, 19]. This content asymmetry results in the extreme lack of positive training examples for robust classifier construction. Works in [2, 19] stress that significant amounts of positive labeled data is necessary for effective training. However, since events are scarce, such examples are hard to come by. Even if positive examples can be obtained, extensive labeling is required, which is a laborious and error prone task. The learning process also requires lengthy training time and tweaking of various parameters [8]. All of this is inconvenient especially when many events need to be considered.

4. The Proposed Framework

We present a framework that circumvents the above mentioned issues. The aims are to limit feature usage to only those that can be easily and reliably derived, and to avoid reliance on labeled training examples. Consequently, event detection is performed in a rule-based manner. Note that the framework is multimodal since it collaboratively analyzes the triplex of textual, visual and aural modalities. We demonstrate that despite being rule-based, this strategy can prevent false alarms while detecting a varied set of events using simple rule-sets.

The features considered are all easily derived. Two semantic shot classes are used as visual features namely the far and closeup camera views. These classes were chosen since they commonly appear in pairs during event occurrences. Audio features on the other hand are considered at the low-level, where the pitch and log-energy are calculated. These features were chosen since they are good indicators of potentially exciting/eventful segments [8]. This work also utilizes the textual modality, which has largely been ignored previous works. In particular, cues from Minute-By-Minute (MBM) reports obtained from sports websites are considered. Its usage in sports

event detection is quite recent, and has been an effective element in works such as [14, 23].

The proposed framework is termed as the Multimodal Collaborative Framework (MMCF). The basic flow is explained as follows:

- The time-stamp cue of a specific event is obtained from a match's MBM report via keyword matching. This cue helps localize event search to the specific minute of an event's occurrence.
- Candidate segments are then extracted from this localized search space by observing transition patterns of semantic shot classes.
- Depending on the event, the eventful segment is identified via a ranking procedure by referring to event specific feature signatures.

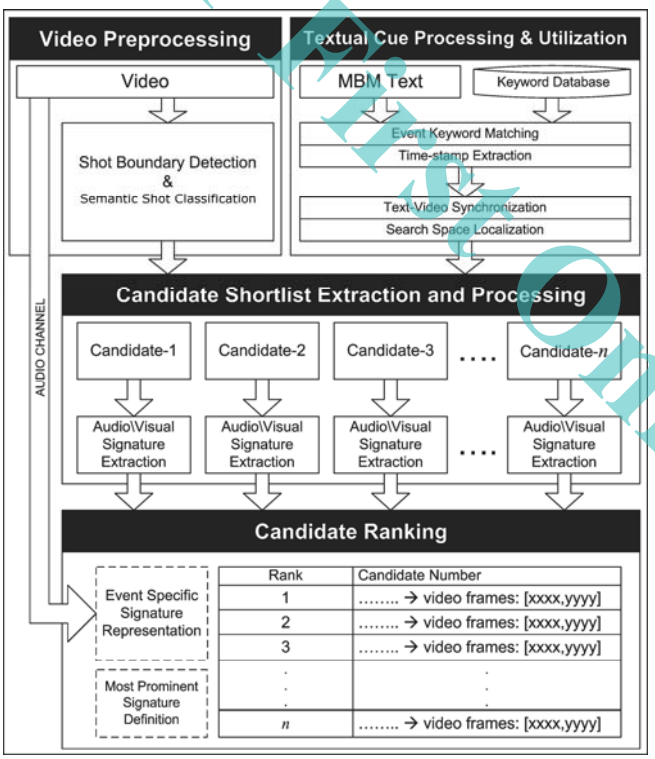

Figure 1. Block diagram of the framework.

Figure 1 indicates that four components are involved: 1) Video Preprocessing; 2) Textual Cue Processing and Utilization; 3) Candidate Shortlist Extraction and Processing; and 4) Candidate Ranking. Note that the events being considered are goals, penalties, yellow cards and red cards.

4.1. Video Preprocessing

Two steps are performed here namely Shot Boundary Detection (SBD) and Semantic Shot Classification (SSC). SBD divides the video into *m*-number of shots, where a shot can be referred to as a sequence of video frames taken by a continuous camera action. We used the algorithm in [1] for SBD. Next, each shot is classified into either a far or closeup view through the semantic shot classification algorithm in [13], whose accuracy is between 90%-94%. An example output of both these processes is shown in Table 1. Columns 2 and 3 show the beginning and ending frames of each shot after SBD, whereas column 4 indicates the assigned semantic shot label after SSC.

Table 1. An example of shots being assigned semantic labels.

Shot Number	Start Frame	End Frame	Shot Class
1	1	29	Far
2	30	40	Closeup
3	41	62	Closeup
...	...	...	...
495	41530	41539	Far

4.2. Textual Cues Processing and Utilization

MBMs of specific matches are obtained freely from broadcasters' websites such as ESPN[1] and BBC[2], as well as sports information providers such as Sportinglife[3] and UEFA[4]. An MBM is annotated during the course a match's progression by experts. It provides a log of all the goings-on in the granularity of minutes, which are finally published on specific web pages. We use two crucial cues namely the specific event name and its minute time-stamp. An excerpt of an MBM is shown in Figure 2 from ESPN. These MBM cues are invaluable since they help eliminate the guesswork in determining which event occurred and when. Moreover, they also allow all events to be accounted for.

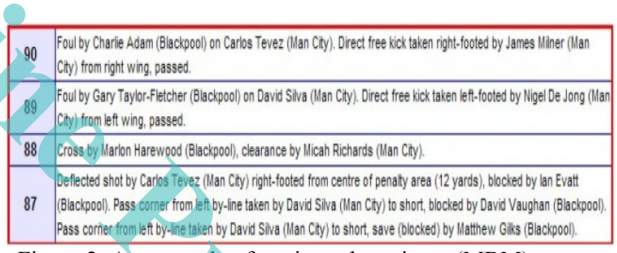

Figure 2. An example of a minute-by-minute (MBM) report.

4.2.1. Keyword Matching and Time-stamp Extraction

Figure 2 shows that the first column contains the minute time-stamps whereas the corresponding explanations are in the second column. We observed that specific events are annotated using dedicated keywords. After looking at various MBMs, we came up with a definition of specific keyword combinations for each event, which are shown in Table 2. Some consist of single keywords, whereas others contain regular expressions such as 'converts *[a-z][a-z][a-z]* penalty' for penalties. An event occurrence is detected if the specific event keyword is found within the respective match's MBM.

[1] http://soccernet.espn.go.com/
[2] http://newsimg.bbc.co.uk/
[3] http://www.sportinglife.com/football/live_match/200111.html
[4] http://www.uefa.com/

Table 2. Keywords and keyword combinations for each event.

Event	Keywords
Goal	goal!, goal by, scored, scores, own goal, convert
Penalty	penalty spot, power penalty, placed penalty, penalty kick, penalty-kick, converts [a-z][a-z][a-z] penalty
Yellow Card	yellow, yellow card, booking, booked, bookable, caution, cautioned
Red Card	dismissed, sent-off, sent off, sending off, red card, red-card, sees red, second booking, second bookable

Events are represented as a set $E \in \{g, p, y, r\}$, where the set elements are abbreviations for *goal*, penalty, yellow card and red card, respectively. Given the task of detecting i-occurrences of an event $e \subset E$, if matching keywords are found within the MBM, the time-stamp of each of the i-th occurrence of e are noted. These can be written as $T_e = \{t_i^e\}$, where $i > 0$ if there is at least one occurrence of e. Then, for each of the i-th occurrence, the event search is initiated within the one-minute segment of each t_i^e.

4.2.2. Text-Video Synchronization and Event Search Localization

The time-stamp t_i^e indicates the minute within which the event has occurred. Directly inferring the corresponding video frame however, can be erroneous due to the misalignment with the actual game time. Therefore, synchronization between t_i^e and its corresponding video frame is performed manually where a reference frame is determined by matching a frame number to the actual elapsed game time of the match. This process is illustrated in Figure 3, where the elapsed game time of 0.25-minutes (15-seconds) corresponds with the 245-th frame of the match video. These values are each denoted as t_{ref} and f_{ref}, respectively. We are however aware that this process can be automated via the method in [23]. However, since some matches fail to display superimposed game clocks; this technique is not always reliable. Therefore, we sacrifice a bit of automation for the sake of reliability.

Figure 3. Reference frame and elapsed game-time synchronization.

The values t_{ref} and f_{ref} are then used localize event search to the one-minute eventful segment. With t_i^e being the minute time-stamp of the event e, the beginning and ending frames for the eventful minute can be determined using equations 1 and 2, respectively:

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

$$f_{i,end}^e = \left\lfloor f_{i,begin}^e + 60\,fr \right\rfloor \quad (2)$$

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

$$\chi_i^e = [f_{i,begin}^e, f_{i,end}^e] \quad (3)$$

4.3. Candidate Shortlist Extraction and Processing

The localized space χ_i^e is still coarse since events only take up a very short time span. Therefore, potential event segments within χ_i^e need to be identified. It was discovered that during events, certain visual properties can be observed. These properties are exploited to decompose the one-minute segment into a shortlist of candidate segments:

- The camera transitions from a far-view to a closeup-view. The former is meant to capture the buildup whereas the latter focuses on player/crowd/coach reactions.
- Closeup-views normally last at least 6-seconds.
- It takes approximately 12-seconds to fully observe an event's progression from conception to finish.

From these properties, the relevant 12-second segments are shortlisted as candidates. Note that the transition points between shot views serve as the mid-point, where each preceding and superseding video segments have equal lengths of 6-seconds. Resultantly, n-candidates are shortlisted from χ_i^e, which can be represented as:

$$D_i^e = \{d_{ik}^e\} \quad (4)$$

where $D_i^e \subset \{\chi_i^e\}$, is the set containing the shortlisted candidates; and d_{ik}^e is the k-th 12-second candidate

within D_i^e (for $k = 1, \ldots, n$). An illustration is given in Figure 4 for two shortlisted goal event candidate segments.

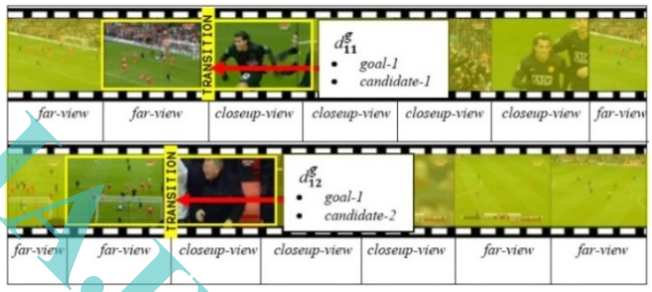

Figure 4. Two far and closeup view pairs resulting in two shortlisted candidate segments.

4.3.1. Candidate Feature Representation

Audio/visual properties during event occurrences were carefully scrutinized from many hours of soccer video. Table 3 provides a summary of the most prominent observations:

Table 3. The prominent modalities during each event.

Event	Aural Properties	
	Crowd Cheering	Excited Speech
Goal	√	√
Penalty	√	√
Yellow Card	√	-
Red Card	√	√

As shown, all the events exhibit audio modality prominence. Visual modalities such as shot class durations were also present, but were deemed insignificant. To allow computerized analysis, these observations were translated into feature representations. Based on the literature, we selected two features:

- *Pitch determination for excited speech*: Works by [22, 8] demonstrated that pitch (*f*0) is reliable for excited human speech detection. Normally, as a direct result of speech excitement, *f*0-values will increase. In this paper, the sub-harmonic to harmonic ratio analysis technique was chosen due to its insensitivity towards noise and prominent unvoiced segments [25];
- *The 1st. MFCC coefficient for crowd cheering*: From [5], crowd reactions such as cheers, jeers, chants and applause can be represented using Mel-frequency Cepstral Coefficients (MFCC). Normally, 12-coefficients are calculated. However, selecting the most appropriate coefficient is vital since each (or a combination) can be used to represent different audio properties. For our work, the 1-st coefficient was chosen due to it representing the log-energy [12]. Log-energy is reliable for measuring loud crowd reactions because of its insensitivity towards speech variations and noise. The formula for MFCC can be represented as:

$$C_n = \sum_{j=1}^{J} [\log S(j)] \cos[n \frac{(j - \frac{1}{2})\pi}{J}] \quad (5)$$

where J is the number of the sub-bands and L the length of the cepstrum. S_j, for $0 \leq j < J$, is the filter bank energy after the j-th triangular band-pass filtering, which is a spectrum envelope scaled according to the mel-frequency. Note that $n = 1, \ldots, L$.

4.3.2. Feature Representation for Goal/Penalty/Red Card Candidates

These events are always accompanied by excited commentary and crowd cheering. Therefore, pitch (*f*0) and log-energy (*le*) are calculated for each of the d_{ik}^g, d_{ik}^p or d_{ik}^r candidates.

Within each candidate, *f*0 measurements are sampled at every 20-millisecond audio frame, with a time-step of 10-milliseconds. Once all samples are taken, a reference clip index $t_{f0_{\max}}$ corresponding to the maximum *f*0 value within the candidate is identified (highest commentator excitement level). A range $[t_{f0_{\max}} - 2, t_{f0_{\max}} + 6]$ at 0.5-seconds per clip is defined around $t_{f0_{\max}}$ to calculate the pitch evolutions. Frame-level *f*0s within this range are initially averaged followed by a final averaging of all frames across clips. This produces, $\overline{f0_{ik}^g} / \overline{f0_{ik}^p} / \overline{f0_{ik}^r}$, which represents the mean pitch for the respective candidate.

For crowd cheering, the *le* is calculated at every 20-millisecond audio frame with no overlap. The maximum of each frame-level measurement is firstly obtained across frames followed by determining the maximum values at each 0.5-second clip. The maximum log-energy $\max le_{ik}^g / \max le_{ik}^p / \max le_{ik}^r$ across all clips is taken to represent the strongest crowd reaction within the respective candidate. Ultimately, the two features form the audio signature for each candidate and can be written as equation 6, where $e \in \{g, p, r\}$.

$$d_{ik}^e \equiv (\overline{f0_{ik}^e}, \max le_{ik}^e) \quad (6)$$

4.3.3. Feature Representation for Yellow Card Candidates

Yellow cards mostly trigger only pronounced crowd reactions. Therefore, *le* measurements are calculated to represent each d_{ik}^y. Two features are computed, namely the $\max le$ and *mean-of-the-*$\max le$ across clips. The former is calculated in a similar fashion to the $\max le$ for the other three events, hence represented as $\max le_{ik}^y$. The latter is obtained by firstly identifying a reference clip index $t_{\max le_{\max}}$ corresponding to the maximum $\max le$

measurement within the candidate. Then, a range $[t_{\max le_{\max}} - 2, t_{\max le_{\max}} + 4]$ is defined around $t_{\max le_{\max}}$ to calculate the le evolutions. Frame-level $\max le$ values within this range are initially averaged followed by a final averaging of all frame $\max le$ values across clips. This produces $\overline{\max le_{ik}^{y}}$, which represents the mean $\max le$ for the candidate. Ultimately, each candidate's audio signature is represented by the two features, and written as equation 7:

$$d_{ik}^{e} \equiv (\overline{\max le_{ik}^{y}}, \max le_{ik}^{y}) \quad (7)$$

4.4. Candidate Ranking

Determining the actual eventful segment is not as simple as determining which candidate has the maximum feature measurements. We discovered that in some cases, during goal events, both pitch and log energy were not both at the highest. The case was similar for red cards, penalties and yellow cards. The most common scenario is for one of the features to be highest, whereas the other is relatively high across candidates. Due to this, a different approach is proposed.

4.4.1. Overall Approximated Non-Event Audio Feature Representation

Since the events can be described by each of their own feature signatures, close scrutiny of how the signatures evolve through time provides a clue of the overall audio representation for each match video. By taking many consecutive audio samples consistent with each event's feature combinations, it was discovered that most of the measurements were biased towards non-event type audio. This is consistent with the inherent nature of soccer broadcasts; where non-events such as dribbling, goal kicks, short/long passes etc. dominate game proceedings [5, 19]. These non-events do not trigger intense audio measurements and are prevalent throughout most of the match's proceedings. The aural measurements are also more consecutive and less sparse, resulting in stable temporal feature evolution.

Due to this, when significant amounts of samples are consecutively taken from a match's audio signal, it is highly likely that the overall distribution will be biased towards less intense, non-eventful measurements. We represent in Fig. 5, the 800-consecutive samples of the $\max le$ and $\overline{\max le}$ combination taken from one match[5]. It clearly shows dominance of relatively lower intensity measurements. Another observation is that the samples are approximately Gaussian (Normal). Therefore, the data points can be described by the mean(s) μ and variance σ^2 (or covariance matrix $\sum$ for the multivariate case), and can make use of the Gaussian probability density function (pdf) as representation.

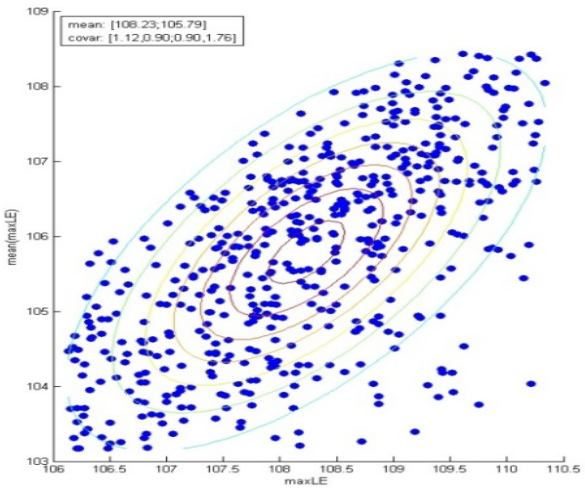

Figure 5. Yellow card event feature samples for a particular match.

Overall, we conclude that if sufficient consecutive samples are taken from a match video, the resultant data point distribution will be an approximate representation of the un-eventfulness of the video itself. This can then serve as a reference against each candidate to determine a likelihood score of whether the candidate belongs to the generated distribution or not. In other words, if the score is higher, this indicates higher likelihood that the candidate does not contain an event. Lower scores on the other hand indicate that the candidate is more likely to be eventful.

4.4.2. Goals, Penalties and Red Cards

To generate the approximated representation consistent with these events, the similar feature types of $f0$ and $\max le$ are sampled. 800 consecutive audio samples are taken from each match video. For pitch, samples are obtained from consecutive, non-overlapping 0.5-second windows. Log energy samples are obtained from consecutive, non-overlapping 3.5-second windows. The audio features $f0$ and $\max le$ are uncorrelated, where the former measurement is unrelated to the latter. Therefore, two separate Gaussians are generated for each of the features. These can be represented in equation 8 and 9, respectively:

$$\overline{f0} \sim N(\mu_{\overline{f0}}, \sigma^2_{\overline{f0}}) \quad (8)$$

$$\max le \sim N(\mu_{\max le}, \sigma^2_{\max le}) \quad (9)$$

Each sample's probability density is calculated using Eq. 10, where $p(x_f)$ is the probability density for the feature sample x_f, μ_f being the sample mean, and σ_f (σ_f^2) the standard deviation (variance) of the sampled data points, for $f \in \{\overline{f0}, \max le\}$.

[5] Recall that $\max le$ and $\overline{\max le}$ are the feature combination (signature) for yellow card events.

$$p(x_f) = 1/(\sigma_f\sqrt{2\pi}) \exp(-(x-\mu_f)^2/2\sigma_f^2) \quad (10)$$

4.4.3. Yellow Cards

800 consecutive samples are taken from each match video under consideration. Each sample are of the $\max le$ and $\overline{\max le}$ features. The sampling window is set to be at every 3.5-seconds, with no overlap. Both these features are correlated since $\overline{\max le}$ is calculated based on $\max le$. Therefore, a single 2-dimensional Gaussian is used, as shown in equation 11:

$$(\max le, \overline{\max le}) \sim N_2(\mu_{\max le}, \mu_{\overline{\max le}}, \Sigma) \quad (11)$$

Each sample's probability density can be calculated using Eq. 12, where $p(x)$ is the probability density for the signature sample x, for $x = (\max le, \overline{\max le})$. μ is a vector of the means of $(\mu_{\max le}, \mu_{\overline{\max le}})$ and $\sum$ the covariance matrix of the sampled data. The number of features of the distribution is represented by m. Note that $m = 2$, since two features are considered namely $\max le$ and $\overline{\max le}$.

$$p(x) = \frac{1}{2\pi^{\frac{m}{2}}|\Sigma|^{\frac{1}{2}}} \exp(-\tfrac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \quad (12)$$

4.4.4. Final Candidate Ranking via Likelihood Calculation

The main task is to calculate the probability density (i.e. likelihood score) for each candidate, with regards to its respective approximated non-event representation.

- *Minimum likelihood score for Goals/Penalties/Red Cards*: The score for a goal, penalty or red card candidate is calculated by substituting its feature representation $d_{ik}^g / d_{ik}^p / d_{ik}^r$ into Eq. 10. The final score is calculated as the product between $p(f0_{ik}^e)$ and $p(\max le_{ik}^e)$, for $e \in \{g, p, r\}$. The candidate with the minimum likelihood score indicates that it is least likely to be a non-event, and hence ranked top most in its shortlist. This can be written as equation 13:

$$min(p(f0_{ik}^e) \times p(\max le_{ik}^e)) \quad (13)$$

- *Minimum likelihood score for Yellow Cards*: The score for the yellow card candidate is calculated by substituting its feature representation d_{ik}^y into Eq. 12, which consists of $(\max le_{ik}^y, \overline{\max le_{ik}^y})$. The candidate producing the minimum score indicates that it is least likely to be a non-event, and hence ranked top most in its respective shortlist. This can be written as equation 14:

$$min(p(\max le_{ik}^y, \overline{\max le_{ik}^y})) \quad (14)$$

5. Experimental Results

Experiments were conducted to evaluate the three components of Event Keyword Matching, Candidate Shortlist Extraction and Candidate Ranking. The MMCF was implemented using MATLAB-R2007a and tested on 21 matches spanning ~30-hours. The videos were in AVI format with frame rates set at 15-fps. Audio was encoded in MP3-mono, 22.5-kHz at 32kbps. Each of the match halves were processed separately where non-game footage such as commercial breaks and half-time commentaries were omitted. The soccer matches were from the European Champions League (3-matches), Barclay's English Premier League (15-matches), Italian Serie-A (2-matches) and Spanish La Liga (1-match).

5.1. Event Detection via Keyword Matching

Currently, the MBMs for each match are manually imported into Microsoft Excel worksheets. Keyword matches are searched for based on the listing in Table 2. The precision equation 15 and recall equation 16 for event keyword detection is shown in Table 4.

$$Precision = Detected / Detected + False \quad (15)$$

$$Recall = Detected / Detected + Missed \quad (16)$$

The results suggest that all events were successfully detected, and the *100% precision* indicates irrelevant keywords were avoided. Avoiding false alarms is critical as the candidate shortlist extraction and ranking processes might fail when working with irrelevant footage.

Table 4. Event detection via keyword matching.

Event	Ground Truth	Detected	False	Missed	Precision	Recall
Goal	62	62	0	0	100%	100%
Penalty	10	10	0	0	100%	100%
Yellow Card	70	70	0	0	100%	100%
Red Card	3	3	0	0	100%	100%

5.2. Candidate Shortlist Extraction

Recall is more critical in this stage as all actual event occurrences should be accounted for. *Precision* should be preferably low to avoid extraneous candidates. Table 5 shows the results of the shortlist extraction process. The abbreviations in Table 5 are explained as: GT-ground truth, RC-relevant number of candidates, TNC-total number of candidates extracted, and ACPS-average number of candidates per shortlist. equations 17 and 18 calculate the precision and recall, respectively:

$$Precision = RC / TNC \quad (17)$$

$$Recall = RC / RC + Missed \quad (18)$$

Table 5. Candidate segments extraction results for all matches.

Event	GT	RC	TNC	ACPS	Missed	Precision	Recall
Goal	62	62	144	2	0	100%	100%
Penalty	10	10	26	3	0	100%	100%
Yellow Card	70	70	151	2	0	100%	100%
Red Card	3	3	8	3	0	100%	100%

The results demonstrate that all relevant candidates were successfully extracted. The ACPS was also low, i.e., between 2 to 3 per shortlist. This is crucial for candidate ranking where, in case the highest ranked candidate does not contain the actual event, it will not be difficult to access the actual eventful segment as the total number of extracted candidates is low. This is very useful in a retrieval setting.

5.3. Candidate Ranking

Table 6 shows the results of the ranking process, where the columns Rank-1, Rank-2 and Rank-3 shows the number of actual event segments assigned to that respective rank. The following provides a discussion regarding the final ranking results:

Table 6. Candidate ranking results for all events.

Event	Ground Truth	Candidate Rank		
		Rank-1	Rank-2	Rank-3
Goal	62	58 (93.55%)	4 (6.45%)	0 (0.00%)
Penalty	10	9 (90.00%)	1 (10.00%)	0 (0.00%)
Red Card	70	55 (78.57%)	12 (17.14%)	3 (4.29%)
Yellow Card	3	3 (100.00%)	0 (0.00%)	0 (0.00%)

- *Goals*: For the two wrongly ranked *goal* segments, the commentators were less exited and were not vocally aroused. Closer observation revealed that the goals were not detrimental to the match's outcome, resulting in less vocal arousal from the commentator. One goal was wrongly ranked since the commentators became excited after the 12-second window, causing the goal segment to have high crowd cheer only, but very low commentary excitement. The final ranking error was due to interference in the audio signal where 'clanking' sounds were present in non-eventful segments. This caused pitch measurements to significantly rise causing these segments to be assigned lower likelihood scores;
- *Penalties*: The sole 2nd ranked penalty was due to it taking place at the end of the game. Since it had little significance to the match's outcome, low commentator and crowd reactions were recorded. On the contrary, the foul causing the penalty had higher pitch and log-energy measurements;
- *Yellow cards*: All the 2nd and 3rd-ranked segments exhibited lower measurements compared to the non-event candidates. In one scenario, a bad foul was committed but was not awarded a yellow card. In the other cases, the yellow card happened very quickly, resulting in less intense crowd reactions.

Table 7. Precision and Recall for overall event detection.

Event	Ground Truth	Relevant	Irrelevant	Missed	Precision	Recall
Goal	62	58	4	0	93.55%	100.00%
Penalty	10	10	1	0	90.91%	100.00%
Yellow Card	70	70	15	0	82.35%	100.00%
Red Card	3	3	0	0	100.00%	100.00%

Overall, a high percentage of the actual event segments were ranked 1st within each shortlist. Table 7 shows the precision equation 19 and recall equation 20 for each event as an overall performance indicator. The value for Irrelevant in column three is equivalent to the total number of actual event segments ranked other than first. From a retrieval perspective, the results are encouraging since high precision entail that users need not wade through many irrelevant results to retrieve the correct segment. The perfect recall scores are good for both retrieval and summarization applications since all the desired eventful segments are returned.

$$Precision = Relevant / Relevant + Irrelevant \quad (19)$$

$$Recall = Relevant / Relevant + Missed \quad (20)$$

6. Conclusions

We have presented a soccer event detection framework for goals, penalties, yellow cards and red cards. The overall accuracy is impressive, demonstrating the effectiveness of the approach. The textual cues are crucial to indicate event occurrences as well as to localize and significantly reduce the search space. This allows simpler observations to be made for identifying potential eventful segments. Finally, the actual event segments are identified via a ranking process based on the most prominent feature combinations of pitch and/or log-energy. The experimental results are very encouraging with very high precision and recall. Some limitations have been identified. Firstly, MBMs are sometimes unavailable after a certain period, without which, the MMCF would fail. Therefore, it is necessary to capture the MBM data directly after a match is annotated. Next, although it is difficult to obtain labeled training examples for soccer video, labeling small sets is still feasible. Therefore, semi-supervised learning methods such as [29] might be promising where the learning process can benefit from partially labeled training instances. Finally, we aspire to detect more events with the hope of constructing a comprehensive index for soccer videos.

References

[1] Abd-Almageed W., "Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing", *15th IEEE*

International Conference on Image Processing, pp.3200-3203, 2008.

[2] Bertini M., Del Bimbo A., and Nunziati W., "Highlights modeling and detection in sports videos", *Pattern Analysis & Applications*, vol.7, no.4, pp. 411-421, 2004.

[3] Byungho M., Jinhyuck K., Chongyoun C., Hyeonsang E., and McKay R., "A compound framework for sports results prediction: A football case study", *Knowledge-Based Systems*, vol.21, no.7, pp.551-562, 2008.

[4] Camastra F., and Vinciarelli A., *Machine Learning for Audio, Image and Video Analysis: Theory and Applications (Advanced Information and Knowledge Processing)*, Springer, 1st. edition, 2007.

[5] Chih-Chieh C., and Chiou-Ting H., "Fusion of audio and motion information on HMM-based highlight extraction for baseball games", *IEEE Transactions on Multimedia*, vol.8, no.3, pp.585-599, 2006.

[6] Chih-Hao L., Wei-Ta C., Jin-Hau K., Ja-Ling W., and Wen-Huang C., "Baseball event detection using game-specific feature sets and rules", *IEEE International Symposium on Circuits and Systems*, vol.4, pp.3829-3832, 2005.

[7] Chung-Yuan C., Huang-Chia S., and Chung-Lin H., "Semantics-based highlight extraction of soccer program using DBN", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol.2, pp.1057-1060, 2005.

[8] Coldefy F., and Bouthemy P., "Unsupervised soccer video abstraction based on pitch, dominant color and camera motion analysis", *12th Annual ACM International Conference on Multimedia*, pp. 268-271, 2004.

[9] Dahyot R., Kokaram A., Rea N., and Denman H., "Joint audio visual retrieval for tennis broadcasts", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol.3, pp.561-564, 2003.

[10] Ekin A., Tekalp A., and Mehrotra R., "Automatic soccer video analysis and summarization", *IEEE Transactions on Image Processing*, vol.12, no.7, pp. 796-807, 2003.

[11] Eldib M., Zaid B., Zawbaa H., El-Zahar M., and El-Saban M., "Soccer video summarization using enhanced logo detection", *Proc. 16th IEEE Internatinal Conference on Image Processing*, pp.4345-4348, 2009.

[12] Feng L., Nielsen A., and Hansen L., "Vocal segment classification in popular music", *9th International Conference on Music Information Retrieval*, pp.121-126, 2008.

[13] Halin A., Rajeswari M., and Ramachandram D., "Shot view classification for playfield-based sports video", *IEEE International Conference on Signal and Image Processing Applications*, pp.410-414, 2009.

[14] Jinjun W., "Content-Based Sports Video Analysis and Composition", *Ph.D. thesis*, Nanyang Technological University, School of Computer Engineering, 2006.

[15] Kolekar M., Palaniappan K., Sengupta S., and Seetharaman G., "Semantic concept mining based on hierarchical event detection for soccer video indexing", *Journal of Multimedia*, vol.4, no.5, pp.298-312, 2009.

[16] Lin K., Joo-Hwee L., Qi T., Kankanhalli M., and Xu C., "Visual keywords labeling in soccer video", *17th International Conference on Pattern Recognition*, vol.3, pp.850-853, 2004.

[17] Min C., Shu-Ching C., and Mei-Ling S., "Hierarchical temporal association mining for video event detection in video databases", *23rd IEEE International Conference on Data Engineering Workshop*, pp.137-145, 2007.

[18] Min X., Xu C., Lingyu D., Jesse S., and Suhuai L., "Audio keywords generation for sports video analysis", *ACM Transactions on Multimedia Computing, Communications and Applications*, vol.4, no.2, pp.1-23, 2008.

[19] Ren R., "Audio-visual football video analysis, from structure detection to attention analysis", *Ph.D. thesis*, University of Glasgow, 2008.

[20] Sadlier D., and O'Connor N., "Event detection in field sports video using audiovisual features and a support vector machine", *IEEE Transactions on Circuits and Systems for Video Technology*, vol.15, no.10, pp.1225-1233, 2005.

[21] Snoek C., "The authoring metaphor to machine understanding of multimedia", *Ph.D. thesis*, University of Amsterdam, 2005.

[22] Tjondronegoro D., "Content-based video indexing for sports applications using multi-modal approach", *Ph.D. thesis*, Deakin University, 2005.

[23] Xu C., Wang L., Lu L., and Zhang Y., "A novel framework for semantic annotation and personalized retrieval of sports video", *IEEE Transactions on Multimedia*, vol.10, no.3, pp.421-436, 2008.

[24] Xu M., Maddage N., Xu C., Kankanhalli M., and Qi T., "Creating audio keywords for event detection in soccer video", *International Conference on Multimedia and Expo*, vol.2, pp.281-284, 2003.

[25] Xuejing S., "Pitch determination and voice quality analysis using subharmonic-toharmonic ratio", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol.1, pp.333-336, 2002.

[26] Yo-Ping H., Ching-Lin C., and Sandnes F., "An intelligent strategy for the automatic detection of highlights in tennis video recordings", *Expert*

Systems with Applications, vol.36, no.6, pp.9907-9918, 2009.

[27] Zakaria E., Abdellatif R., and Mohamed B., "Using Wordnet for text categorization", International Arab Journal of Information Technology, vol.5, no.1, pp. 16-24, 2008.

[28] Zhang Y., Xu C., Rui Y., Jinqiao W., and Hanqing L., "Semantic event extraction from basketball games using multimodal analysis", *IEEE International Conference on Multimedia and Expo*, pp.2190-2193, 2007.

[29] Zhang T., Xu C., Zhu G., Liu S., and Lu H. "A generic framework for event detection in various video domains", *Proc. of the ACM International Conference on Multimedia*, pp. 103–112, 2010.

Alfian Abdul Halin is a lecturer at the Faculty of Computer Science & Information Technology, University Putra Malaysia. He obtained his PhD in Computer Science from University Sains Malaysia in 2011. His research interests include image/video processing and multimedia content understanding.

Mandava Rajeswari received her Ph.D. from the University of Wales in 1995. She heads the Computer Vision Lab at the School of Computer Sciences, Universiti Sains Malaysia. Her research interests are in computer vision and medical image analysis.

Mohammad Ehsan Abbasnejad completed his Masters in Computer Science from Universiti Sains Malaysia in 2010. He is currently a PhD student at the Australian National University, Australia. His interests are in machine learning and its wide range of real-world applications.