International Journal of Computer Graphics
Vol. 6, No.1 (2015), pp.13-36
http://dx.doi.org/10.14257/ijcg.2015.6.1.02

Fast Highlight Detection and Scoring for Broadcast Soccer Video Summarization using On-Demand Feature Extraction and Fuzzy Inference

Mohamad-Hoseyn Sigari[1], Hamid Soltanian-Zadeh[1,2] and Hamid-Reza Pourreza[3]

[1]Control and Intelligent Processing Center of Excellence (CIPCE), School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran 14399, Iran
[2]Image Analysis Laboratory, Department of Radiology, Henry Ford Health System, Detroit, MI 48202, USA
hoseyn@sigari.ir, hszadeh@ut.ac.ir, hpourreza@um.ac.ir

Abstract

In this paper, a fast highlight detection and scoring method is proposed using an on-demand feature extraction and a fuzzy inference system. The proposed method partitions video to highlights and analyzes their content using an on-demand feature extraction approach. Then, a score is assigned to each highlight using a Fuzzy Inference System (FIS) according to the analyzed content. The assigned score determines importance of the events occurred in the highlight. This method is useful for flexible video summarization. The proposed method for on-demand feature extraction is a heuristic model of attention control that reduces computational complexity of the algorithm greatly. Additionally, FIS offers a simple and robust solution for content analysis. Experimental results illustrate that the proposed method is fast and processes about 130 frames per second on a personal computer. In addition, objective and subjective evaluations show that the proposed method generates high quality results for highlight detection, scoring and video summarization.

Keywords: *Broadcast Soccer Video; Fuzzy Inference System; Highlight Detection; Highlight Scoring; On-Demand Feature Extraction; Stimulus Driven Attention; Video Summarization*

1. Introduction

Nowadays, a variety of digital videos such as movies and sport videos are available on the web, hard disks, and non-volatiles memories. Thus, researchers are interested to work on content video analysis systems to facilitate management and information extraction from video data sets. Content video analysis systems includes a wide varieties of video analysis systems such content-based video compression [1], video indexing and retrieval [2], video classification [3], action and activity analysis [4], team tactic analysis in sport videos [5] and video summarization [6].

Digital videos are growing exponentially. On the other hand, many users do not have enough time to watch whole of an interesting video. Also, some users want to receive video highlights on-line, on a narrow band network, *e.g.*, mobile networks. Highlight detection and video summarization systems are necessary to extract informative segments of video. A summarization system may convey the informative contents of video by two output types: (1) *key frames* and (2) *video segments*. In some researches, video summarization by selection of some informative video segments is called dynamic video skimming [7].

ISSN: 2093-9663 IJCG
Copyright ⓒ 2015 SERSC

Type of summary is selected according to the application of video summarization or user preference. For instance, in [8], the summarized output is presented as a set of key frames which is known as storyboard and in [9], the output is a sequence of important video shots. In [10], both of key frames and video segments can be generated by a general framework. The summarization method proposed in [10] is based on human attention.

Video summarization is a high-level processing task. Therefore, informative objects, concepts and events in video have to be detected before summarization. Current methods are usually domain specific. In [9, 11, 12], hierarchical approaches were proposed for broadcast soccer video summarization. In these systems, video summarization is performed after some low-level and mid-level processing such as logo detection, shot view recognition, goal mouth detection, and scoreboard detection. These systems may be developed as real-time systems using high performance computers. In some researches like [13], summarization is based on only replay detection and usually does not has complex computation; thus, such methods can be implemented in real-time.

In [14], summarization is essentially based on low-level and mid-level features, but goal events are detected and used as a high-level feature beside other feature to offer more summarization options for user. This system can summarize the video in three types: (1) all slow-motion segments, (2) all goal events and (3) slow-motion segments classified based on appearance of a specific object. Except type 3, video summarization can be performed on off-the-shelf hardware in real-time. In order to perform algorithms in real-time and avoid miss detection, the precision of goal event detection is low (about 49%).

The best approach for video summarization is based on event detection. This approach was used in [15]. In this system, at first, events are detected as high-level features based on some low and mid-level features. Then, detected events are classified to three classes: goals, attacks and others. Finally, the summarized video is generated based on detected events and user interests.

Soccer event detection systems such as [16, 17, 18, 27] can be used for video summarization too. In some event detection methods, features are extracted in a hierarchical semantic structure from low-level to high-level. Then, events are detected based on extracted features. Therefore, semantic gap is effectively covered and video summarization is usually robust and efficient. The main disadvantage of this approach is that usually event detection algorithms are computationally complex. Therefore, such systems are performed on off-the-shelf hardware off-line.

Above mentioned methods are domain specific and thus, domain knowledge and heuristic rules are used to develop the system. For example, although the methods presented in [16, 17] built a model for event detection based on machine learning approaches, these methods used heuristic rules and domain knowledge too. In [16], mid-level feature extractions, such as player detection, are based on the domain knowledge while event detection is based on the machine learning approaches. In [17], mid-level feature extractions are based on the machine learning methods while event detection is based on the heuristic rules. However learning-based methods present more general models using training samples, but they usually have lower precision and recall rates with respect to heuristic-based methods. Additionally, learning-based methods suffer from limited number of training samples. It is an important aspect of almost all semantic video analysis systems that a given event/concept usually occurs in various forms, but we usually can capture a limited number of samples (forms) for training of that event/concept. It should be noted that almost all interesting events/concepts are rare and this causes the problem to be more difficult.

General-purpose video summarization methods are applicable for different types of video. These systems usually summarize the video based on low-level features without minimum usage of domain knowledge. In [8, 19, 20], general-purpose summarization methods were proposed that are based on extraction of some low-level features including dominant color, color histogram, edge directions and motions. These systems serve as a pre-processing tool for high-level video analysis such as concept/event detection [20] or content identification [21]. Such systems are usually based on key frame extraction, therefore, they are disabled to identify and rank high-level concepts and events. Many of these systems can be performed on off-the-shelf hardware in real-time, because they usually based on low-level and simple features.

In this paper, authors propose a fast highlight detection and scoring method for broadcast soccer video with the aim of flexible summarization. This system detects and scores the highlighted segments of soccer video using a Fuzzy Inference System (FIS). Authors address the high computational complexity of the algorithms using a heuristic on-demand feature extraction method. The proposed method extracts required features in low-level, mid-level and high-level based on an on-demand approach. The proposed heuristic model is a stimulus driven attention control mechanism.

Attention control or selective attention is a multi-disciplinary concept which is considered in computer science, neurobiology, and psychology. Frintrop *et al.*, [22] defined selective attention as "the mechanism in the brain that determines which part of the multitude of sensory data is currently of most interest". In other words, attention control is a process to concentrate on some parts of input data selectively and ignore others.

Our proposed system may be implemented as a real-time pre-processing stage for event detection. Also, it can be used for soccer fans as a mobile value added service. For example, the system can submit highlighted soccer video segments to cell-phone of users according to their preferences.

Rest of the paper is organized as follows. In Section 2, details of the proposed method for highlight detection are presented. Experimental results are described in Section 3. Finally, conclusions and future works are explained in Section 4.

2. The Proposed Method

The proposed method for highlight detection/scoring can be presented from two different viewpoints: *data flow* and *control flow*. At first, authors introduce the data flow diagram of the proposed method.

The proposed method includes three processing levels: low-level, mid-level, and high-level. Data flow diagram of our method is shown in Figure 1.

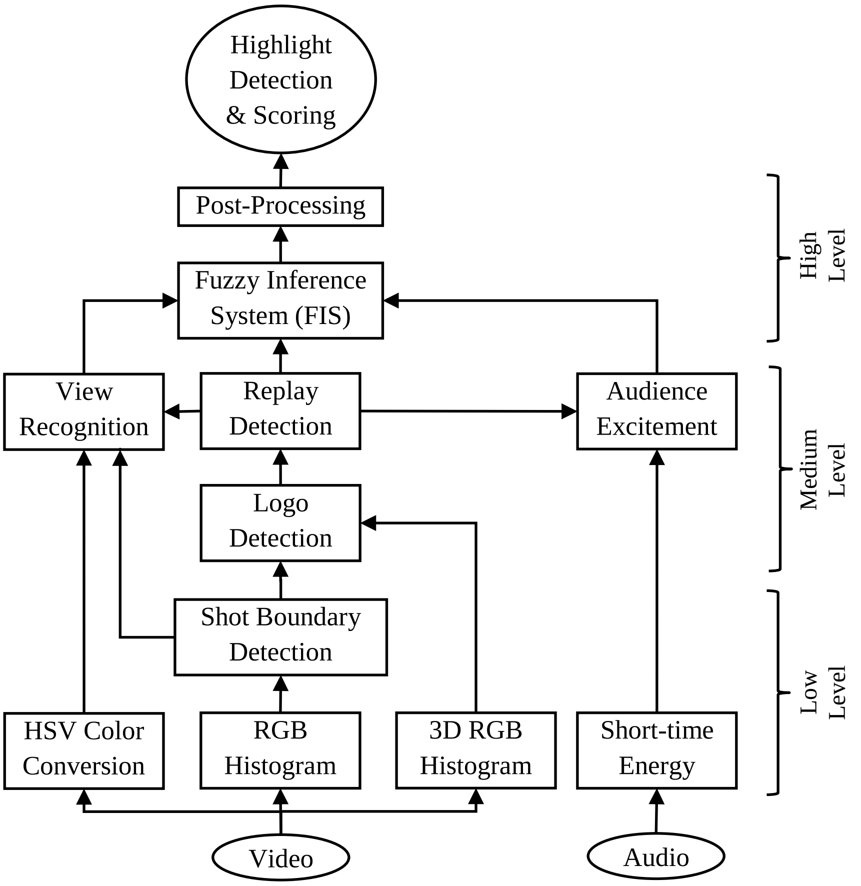

Figure 1. Data Flow Diagram of the Proposed Method for Highlight Detection/Scoring

As there are three processing levels in the proposed method, the method may look like a level-by-level processing method but it extracts features by an on-demand approach. In other words, contrary to many video analysis systems those extract features level by level; our approach is on-demand and thus runs very fast.

At first, histograms of the R, G, and B components are calculated for each frame. Then, shot transitions are detected based on the differences of the histograms of the consecutive frames. Shot boundary detection partitions the video at the low-level where each low-level partition (shot) contains similar contents.

If a shot transition is detected, 3D RGB histograms of frames during shot transitions are computed. Note that computation of a 3D histogram is more computational than an ordinary histogram of a color channel but a 3D RGB histogram provides comprehensive information about distribution of different colors in the frame and is effective for analysis of contents of the video. Thus, 3D RGB histogram is only computed for a few frames those appear during shot transition.

Copyright ⓒ 2015 SERSC

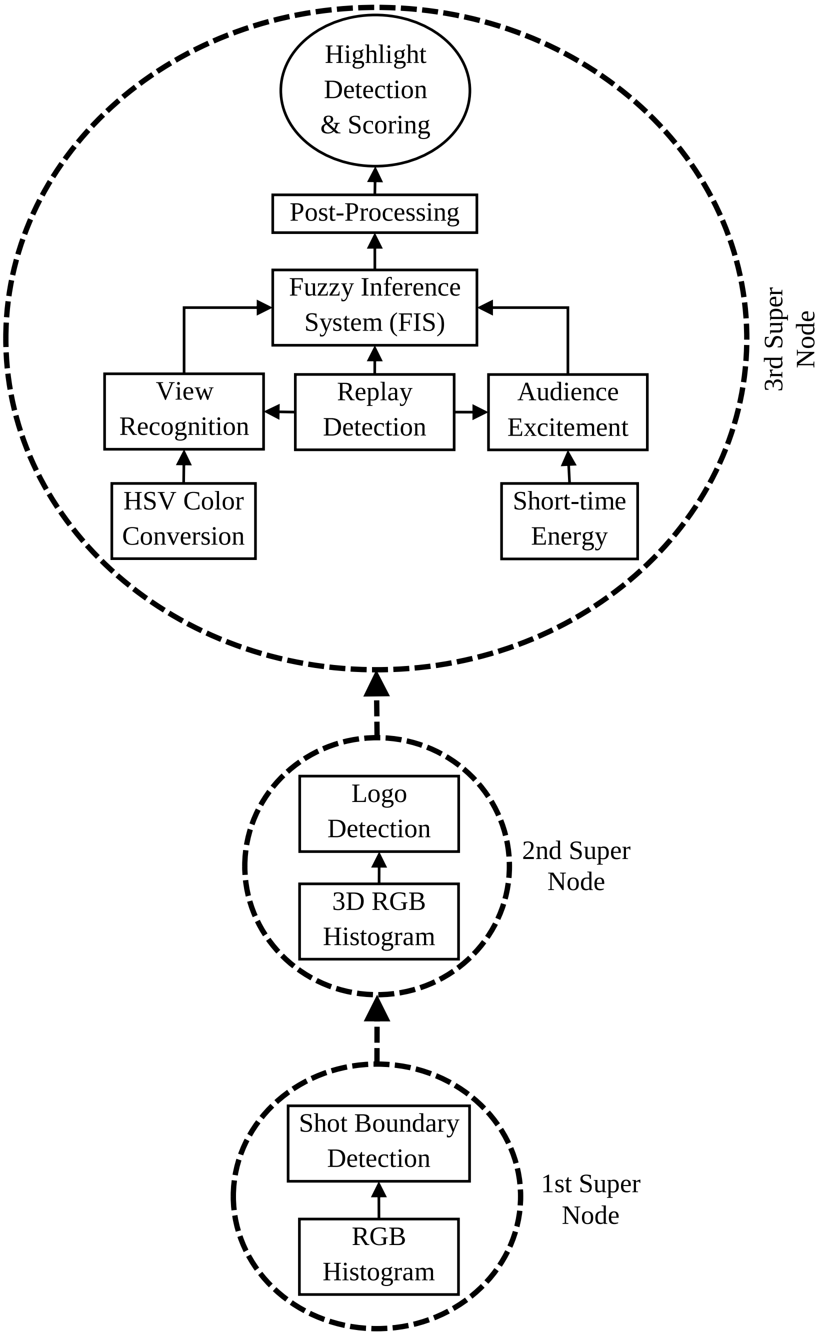

Figure 2. Control Flow Diagram of the Proposed Method for Highlight Detection/Scoring

After computing the 3D RGB histogram, logo detection is performed. Logo usually appears like a gradual shot transition in soccer video; therefore, it can be detected by shot boundary detection. To discriminate the logo from other shot transitions, a decision tree classifies the 3D RGB histogram of the logo. When the decision tree detects a logo, the appearance time of the logo is logged. Usually, all replays of a soccer video are sandwiched by two logo transitions. Thus, the replay parts can be detected based on the consecutive appearance of the logo. By replay detection, the video is partitioned at the

Copyright ⓒ 2015 SERSC

high-level conceptual paradigm where each high-level partition (the time interval between two consecutive replays) contains similar concepts.

After replay detection, a simple classifier investigates the view type of each shot during the replay part. This classifier categorizes the view type of each shot into two categories using the proposed method in the HSV color space.

In addition, the average short time energy of the audio before replay is computed as a low-level audio feature. This low-level feature is processed and excitement of the audience is calculated as a mid-level feature.

The FIS processes three features: (1) duration of replay part; (2) portion of close-up and out-field view of shots during replay; and (3) audience excitement before replay part. Finally, FIS computes importance of each replay part using fuzzy heuristic rules. The FIS output determines the importance of each replay part. In other words, FIS computes possibility of occurrence of an important event in each replay and the time before it.

A finite state machine like the one depicted in Figure 2 with three sequential super nodes can be used to show the control flow of the proposed method. The first super node analyzes all video frames. This super node triggers the second super node when it detects a shot transition. Thus, the second super node works occasionally. When the second super node detects a logo during a shot transition, it triggers the third super node. The third super node is the most complex super node but it works in rare cases. The proposed method is a stimulus driven (bottom-up) model of attention control. In a stimulus driven attention control mechanism, attention factors are derived from the raw visual data [22]. In contrary, in goal driven (top-down) model, attention factors are derived from a given goal, knowledge or expectations.

According to above explanations, in the proposed system, RGB Histogram and Shot Boundary Detection are performed on all frames but the other blocks of the system are performed occasionally.

2.1. Shot Boundary Detection

Shot boundary detection is a low-level processing that partitions the video into low-level parts. A fast and efficient method for shot boundary detection is based on the difference of the histograms of the consecutive frames. In this method, gradual shot transitions may be discarded, therefore, in the proposed method, the threshold applied for shot detection is determined adaptively to detect both of the abrupt and gradual shots.

To detect shot transitions by a fast and reliable method, a 16-bin histogram of each frame is computed for each of the RGB channels. Then, the histogram coefficients are normalized based on the size of the video frame. Therefore, there are 48 histogram coefficients for each video frame. The sum of absolute difference of the histogram coefficients for the ith frame and its previous frame is denoted by $hist\_diff_i$. The shot boundaries are detected by applying an adaptive threshold on $hist\_diff$. The adaptive threshold is defined according to the following equation for each video.

$$th_{shot\_detection} = average(hist\_diff_i \mid hist\_diff_i > mean\_hist\_diff) \quad (1)$$

Here, $mean\_hist\_diff$ is the average value of the vector $hist\_diff$ for a video. Therefore, the threshold for shot detection is equal to the average value of some elements in $hist\_diff$ that are greater than $mean\_hist\_diff$. By applying this threshold, cut and gradual shot transitions can be detected with a high detection rate. However, this method may detect some video segments with huge motion as shot transition. Since we need to minimize the overall shot transitions detection error, we may tradeoff some false shot transitions with detection of almost all cut and gradual shot transitions.

Copyright ⓒ 2015 SERSC

2.2. 3D RGB Histogram

If a shot transition is detected, we compute another visual feature named 3D RGB histogram. Despite the ordinary RGB histogram, constructed by the concatenation of histograms of the R, G, and B channels, a 3D RGB histogram is constructed by quantizing the 3D RGB color space and counting the number of pixels in each voxel of the resulting volumetric RGB space [23]. Here, we quantize R, G, and B to 4 levels generating 4×4×4=64 colors and thus, the length of the 3D RGB histograms are 64. Computational complexity of 3D RGB histogram is more than the ordinary RGB histogram. Therefore, the 3D RGB histogram is calculated for some selected frames that have to be investigated precisely.

It is worth noting that histograms of individual channels of other color spaces like HSV do not define the color information completely either. For example, histogram of H in HSV describes pure colors perfectly but does not provide information about the gray-scales (black, white, and so on). Thus, complete description of colors in a color space requires computation of 3D histograms. Since conversion of the RGB color space to other color spaces is time consuming, we use 3D histograms of the RGB color space.

2.3. Logo Detection

Usually, logo is used to separate the replay clips from other parts of the video in soccer videos. The idea of identifying replays through detection of logos dates back to 2002 [24]. Logo usually appears as a gradual transition. For logo detection, the frames during all gradual shot transitions have to be investigated. In some soccer videos, logo appears gradually and stays on screen for a short time (usually less than 1-2 seconds), then disappears gradually. Therefore, all gradual transitions and short shots have to be examined for logo detection.

In the proposed method, a Classification And Regression Tree (CART) is used to detect the logo based on 3D RGB histogram of each frame. CART is a common decision tree for classification and regression applications. This type of a decision tree is very efficient numerically.

The proposed method analyzes the video frame by frame. In each frame, the 3D RGB histogram is calculated as a feature vector. A matrix is generated by calculating this feature vector for all logo and non-logo training samples. Each column of this matrix represents a feature and each row represents a sample. This matrix is used as the training data for CART. Finally, the trained CART classifies each frame to logo and non-logo classes.

2.4. Replay Detection

Once CART has detected the logo in a sequence of frames (not in a single frame), the frames between the two consecutive logo appearances are considered as the replay part. To detect the start and stop points of a replay, a sliding window is used. If at least 50% of the frames of the current sliding window are detected as a logo frame, then the center of the sliding window is considered as the start/stop point of the replay. This method is based on a simple majority voting technique to reduce effects of logo detection errors. The length of the sliding window can be determined as a portion of the total length of the logo transition.

Replay detection is the main processing step of our proposed method for highlight detection that partitions the video into some high-level partitions. Although, detection of the start/stop point of a replay is very important for detection of other features, but the length of the replay (L_{Replay}) is also an important feature for highlight detection.

2.5. View Type Recognition

View type of soccer video shots is usually divided into four types: (1) far view (or long-shot), (2) medium view (or medium-shot), (3) close-up, and (4) out-field. Usually, a replay containing many close-up or out-field shots is presented after occurrence of a very important event in soccer video. In this case, close-up shots usually show the person who causes the event and the out-filed shots show the cheering of spectators. On the other hand, for events with low or medium importance, replay usually contains long-shots or medium-shots. In the proposed method, we classify the view types into two categories: (1) in-field views (except close-up) and (2) out-field and close-up views. View type classification is performed by detection of grass and skin in the HSV color space.

For this purpose, the grass color is detected using the approach presented in [25] at first. Then, the ratio of the grass pixels to the total pixels of the frame is calculated as R_{Grass}. In out-filed views, R_{Grass} is zero or near zero. Therefore, if $R_{Grass} < 0.05$, the view type is detected as out-filed.

In close-up views, pixels in the center of the frame are occupied by the face or body of a player. Thus, the frame is divided into 9 (3×3) sections where the width and height of the central part is 40% of the total width and total height of the frame (Figure 3).

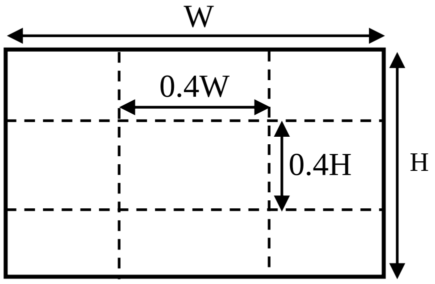

Figure 3. Dividing a Frame to 3×3 Parts for Close-Up View Recognition

The ratio of grass pixels to total pixels of the central part is named R_{Grass}^{Center}. Additionally, the number of skin pixels in the central part of frame is calculated. One of the fast approaches for skin detection is based on applying multiple thresholds on the HSV color space [26]. Thus, the skin is detected in the HSV color space by applying the below thresholds:

$$\begin{cases} 0^{\circ} \leq H \leq 50^{\circ} \\ 0.1 \leq S \leq 0.8 \\ \quad 0.25 \leq V \end{cases} \tag{2}$$

The ratio of skin pixels to the total pixels of the central part is named R_{Skin}^{Center}. If $R_{Grass}^{Center} \leq 0.5$ and $R_{Skin}^{Center} \geq 0.1$, the view type is detected as close-up. Other views except out-filed and close-up views are classified as in-field view.

Processing of all frames in a shot is not necessary for recognition of the view type of that shot, then only 5 frames are selected from 10%, 30%, 50%, 70%, and 90% of shots duration. After classification of the view type of the selected frames, the majority vote method is used to estimate the shot view type. When the view type of shots is classified, the total length of the shots that contains the out-filed and close-up views is calculated during the replay. This value is normalized with respect to the length of the replay (L_{Replay}) and used as a feature named $P_{view\ type}$.

2.6. Audience Excitement

In the proposed method, we introduce audience excitement (AE) as a mid-level audio feature. Audience excitement is used as a feature to measure cheering of the broadcaster and spectators from the audio data. This feature is computed based on a low-level audio feature named short-time audio energy. Short-time audio energy (E_{ST}) is the energy of audio samples during a window with limited size. This energy is computed for each non-overlapping window where the window size is equal to 40 ms (25 windows per second):

$$E_{ST} = \sum_i S_i^2 \tag{3}$$

When a replay is detected, the audience excitement for a limited part of the video before the replay is computed. The length of this part of the video which is used for computation of the audience excitement is noted by L_{AE} (in seconds) and calculated by:

$$L_{AE} = \min(L_{Replay}, 15) \tag{4}$$

Where L_{Replay} is the length of the corresponding replay (in seconds). The selected part of the video used for the computation of the audience excitement is noted by $[t_s, t_e]$ where $t_e - t_s = L_{AE}$. t_s and t_e are the start and the end point of the selected part respectively. t_e is always equal to the start point of the corresponding replay and $t_s = t_e - L_{AE}$. Thus, the boundary of the selected part is determined.

Portions of $[t_s, t_e]$ where short-time energy is more than 80% of the maximum short-time energy in the interval $[t_s, t_e]$ indicate cheering. These portions are selected and AE is defined as the ratio of the length of the selected portions to L_{AE}. The threshold for indication of cheering (80%) was determined by trial and error.

2.7. Fuzzy Inference System for Highlight Detection

The proposed FIS gets three inputs and estimates the importance (score) of each replay by specific fuzzy rules. Inputs are some features of the highlight: (1) the length of the replay (L_{Replay}), (2) the ratio of the total length of the close-up or out-filed shots during the replay to the length of the replay ($P_{view\ type}$) and (3) audience excitement before the replay (AE). The output of FIS is a real-value in range [0,1] which determines score of the input highlight.

At first, each crisp input is fuzzified to a fuzzy variable using a membership function. The membership functions of the inputs are depicted in Figures 4-6. The output of FIS is the importance of the replay denoted by Importance. The membership function of the output is depicted in Figure 7.

After fuzzifying the inputs, an inference engine processes the fuzzy inputs based on a set of fuzzy rules. The rules for the proposed method are listed in Table 1.

The proposed inference engine uses the Mamdani method for implication, which is based on minimum operation. Defuzzification in the proposed FIS is based on the center of gravity method.

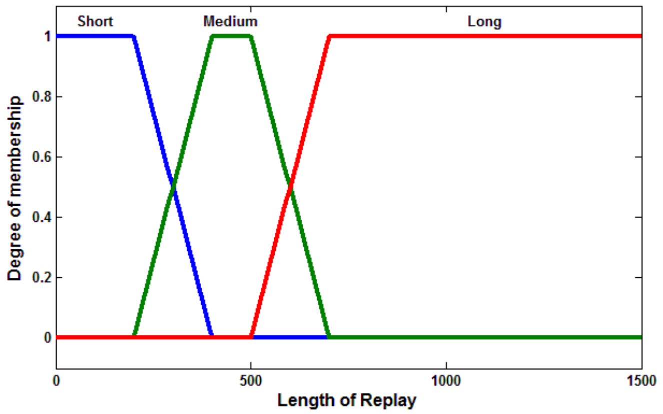

Figure 4. Membership Function of First Input: Length of Replay (L_{Replay})

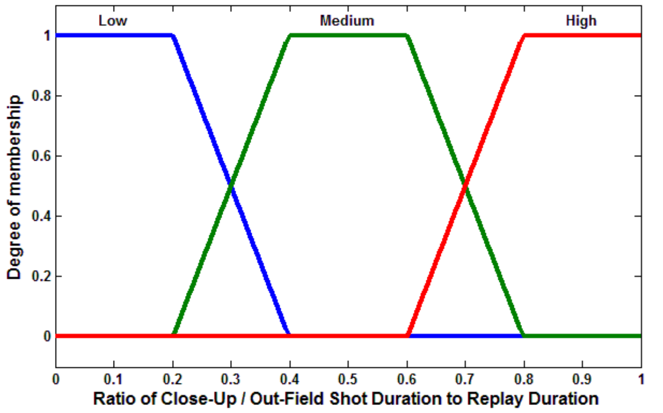

Figure 5. Membership Function of the Second Input: the Ratio of the Total Length of the Close-Up or Out-Filed Shots during the Replay to the Length of a Replay ($P_{view\ type}$)

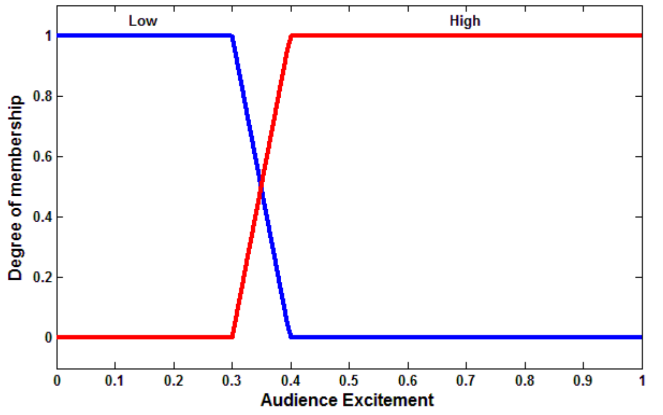

Figure 6. Membership Function of Third Input: Audience Excitement before a Replay (AE)

Copyright ⓒ 2015 SERSC

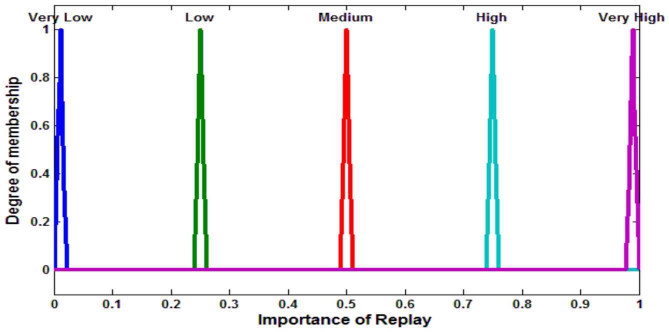

Figure 7. Membership Function of Output: Importance (score) of a Replay ($Importance$)

Table 1. Rules for the Proposed FIS

Rule #	If-Clause			Then-Clause
	Input 1 (L_{Replay})	Input 2 ($P_{view\ type}$)	Input 3 (AE)	Output ($Importance$)
1	Short	-	-	Very Low
2	Medium	Low	-	Very Low
3	Medium	Medium	Low	Very Low
4	Medium	Medium	High	Low
5	Medium	High	Low	Low
6	Medium	High	High	Medium
7	Long	Low	Low	Low
8	Long	Low	High	Medium
9	Long	Medium	Low	Medium
10	Long	Medium	High	High
11	Long	High	Low	High
12	Long	High	High	Very High

2.8. Post-processing

The proposed FIS estimates the importance of each replay but does not summarize the soccer video explicitly. In the final stage of highlight detection and video summarization, the important segments of the video are selected and separated. This stage is named post-processing.

At the post-processing stage, some shots before each replay are selected and combined with the corresponding replay to compose the summarized video. The number of selected shots before a replay is related to the length and importance of the corresponding replay. The total length of a highlight clip is denoted by L_{Clip} and equals:

$$L_{Clip} = L_{Replay} + L_{Play} \quad (5)$$

Where L_{Play} is the total length of the selected shots (in seconds) before the replay. Some constraints are applied to determine L_{Play}. The first constraint is applied by:

$$L_{Play}^{min} = max\left(10, \left(L_{Replay} \times C_{Play}\right)\right) \quad (6)$$

Where L_{Play}^{min} is the minimum length of L_{Play} and always must be equal or greater than 10 seconds. Also, a coefficient named C_{Play} is used to compute L_{Play}^{min}. C_{Play} is related to the importance of the corresponding replay and computed by:

Copyright ⓒ 2015 SERSC

$$C_{Play} = \min\Big(1.4\,,\max\big(0.5\,,(2 \times Importance)\big)\Big) \tag{7}$$

According to the above equation, C_{Play} varies in the range of $[0.5,1.4]$ as a piece-wise linear function. When $Importance < 0.25$, the occurred event is not important; therefore, $C_{Play} = 0.5$ and L_{Play}^{min} is equal to the half length of the replay. When $0.25 \leq Importance \leq 0.7$, the occurred event may be important and C_{Play} is changed in the range of $[0.5,1.4]$ monotonically. When $Importance > 0.7$, the occurred event is important; therefore, $C_{Play} = 1.4$. The constants in the above equation are obtained according to the expert knowledge.

Now, the minimum length of L_{Play} is computed but the final length of L_{Play} depends on the view type of the first shot. The first shot of the play part of each highlighted segment must be long-shot or medium-shot, because each important event in the soccer video starts with a long-shot or medium-shot that shows the scene of the event. In the proposed method, the start point of each highlight clip is same as the start point of the first shot with far or medium view while the length of the play part is equal or greater than L_{Play}^{min}. The proposed post-processing method for highlight detection and determination of the start/stop point of a highlight is demonstrated in Figure 8.

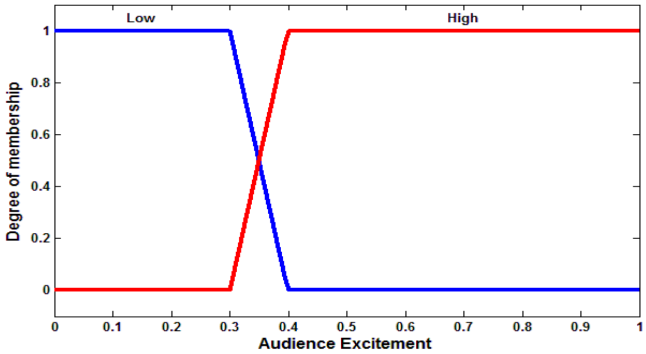

Figure 8. Illustration of the Post-Processing Method Proposed for Determination of the Boundary of a Highlight

3. Experimental Results

Experiments are performed on some broadcast soccer videos related to FIFA World Cup 2010, South Africa. The selected matches are listed in Table 2. The total duration of this dataset is about 10 hours. All videos are HDTV (1280×720) and the video frame rate is 25 fps. The frame size of all videos is down-sampled to 320×180 for increasing the processing speed.

Table 2. List of Videos used for the Experimental Results

Video #	Description	Video #	Description
1	Germany vs. Argentina (1st Half)	2	Germany vs. Argentina (2nd Half)
3	Germany vs. England (1st Half)	4	Germany vs. England (2nd Half)
5	Germany vs. Spain (1st Half)	6	Germany vs. Spain (2nd Half)
7	Germany vs. Uruguay (1st Half)	8	Germany vs. Uruguay (2nd Half)
9	Greece vs. Argentina (1st Half)	10	Greece vs. Argentina (2nd Half)
11	Slovakia vs. Italy(1st Half)	12	Slovakia vs. Italy (2nd Half)

Evaluation of highlight detection and video summarization methods are usually difficult because definition of a suitable criterion for this evaluation is complicated

Copyright ⓒ 2015 SERSC

[6]. Evaluation methods for highlight detection and summarization are divided into two broad categories: subjective and objective. Subjective methods are based on the vote of individuals about the quality of the summarized video, while objective methods evaluate performance of a given technique based on checking the appearance of a specific object or event in the highlights or the summarized video. In this paper, both of the objective and subjective evaluations are considered while the key event for objective evaluation is the goal event.

3.1. Experiments on Logo Detection and Replay Detection

Replay detection is one of the main parts of our method and is based on logo detection. For logo detection, a CART is learnt by some positive and negative training samples. Positive and negative samples are selected from Video #1. To this end, 186 frames that contain logo image are selected manually as positive samples and 2000 frames are selected randomly as negative samples. Some logo images are shown in Figure 9.

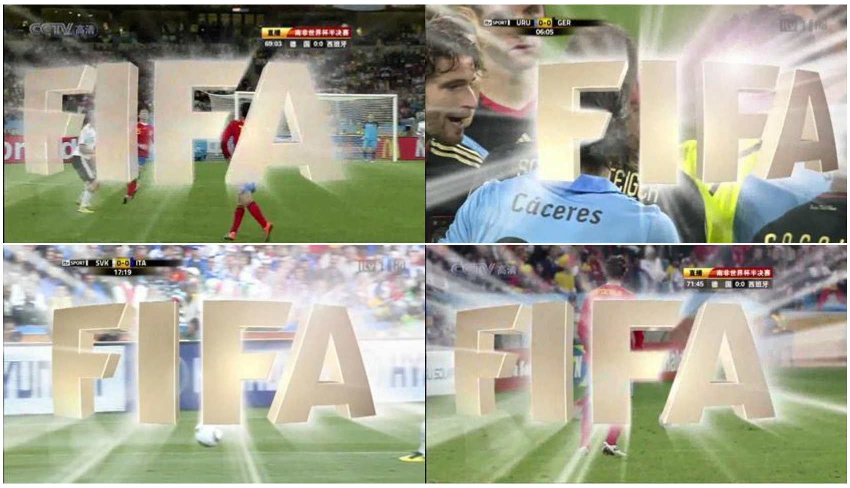

Figure 9. Sample Logo Images from our Video Data Set

Because the negative samples are more than the positive samples, the CART is biased. This means that the False Positive Rate (FPR) of the classifier is less than the False Negative Rate (FNR). This bias is acceptable and desirable for us, because we want to decrease the rate of false logo detection (FPR). Detection of start/stop point of a replay is based on detection of the logo in some consequent frames; therefore, if we lose some logos during a transition that contains the logo, the total efficiency of the replay detection will not be affected strongly. In other words, the proposed method for replay detection is robust to miss detection of logos. Because the logo appears as a gradual transition in our video data set, the training logo samples are selected during a period of time that the whole logo appears in the frame.

The 3D RGB histogram of all training samples is calculated and used to train the CART. In the experiments, the CART could learn the training image perfectly (without any mistakes in the training phase). For detection of start/stop point of the replay, a sliding window is used which is 5 frames long. The Length of the sliding window is equal to 33% of the total length of a logo transition (15 frames). If at least 3 frames out of 5 frames contain the logo, the center of the sliding window is detected as the start/stop point of the replay.

According to the experiments, the logo detection algorithm has some miss detections and false detections. However, accuracy of our proposed method for replay detection is

Copyright ⓒ 2015 SERSC

98.5%. In other words, only 5 replays out of 334 replays are not detected. The main reason for an error in replay detection is an error in logo detection.

3.2. Experiments on View Recognition

The experiments on view recognition are performed on Video #1 and #2. The confusion matrix resulted from the experiments are shown in Table 3. According to this table, the average accuracy of view recognition is 95.0%.

Table 3. Confusion Matrix for View Recognition

Predicted Type Actual Type	Close-Up / Out-Filed	In-Field	Recognition Rate
Close-Up / Out-Filed	307	19	94.2%
In-Filed	28	581	95.4%

3.3. Objective Evaluation of the Proposed Method

Objective evaluation is based on the detection of the goal event which is the most important event in the soccer video. The proposed system segments the highlights and determines the importance of each one; therefore, we expect that the most important highlights contain the goal event. It means that the most important highlights should contain the goal event. Thus, the highlights are sorted based on their importance. Then, the sorted highlights are investigated to determine whether the goal event has occurred in a highlight or not. In the best case, if there are G goal events in the video, the first G highlights with maximum importance would contain these goal events.

In general, if there are G goal events in the video, H highlights have to be considered to find all G goals where $H \geq G$. In other words, the number of hits to find all the goals is H; in the best case, H is equal to G. The ratio of H to G is one of the objective evaluation measures used in this paper. This evaluation measure is named average hit and noted by $H_{avg} = \frac{H}{G}$. In the best case, H_{avg} is equal to 1 and in other cases, H_{avg} is greater than 1.

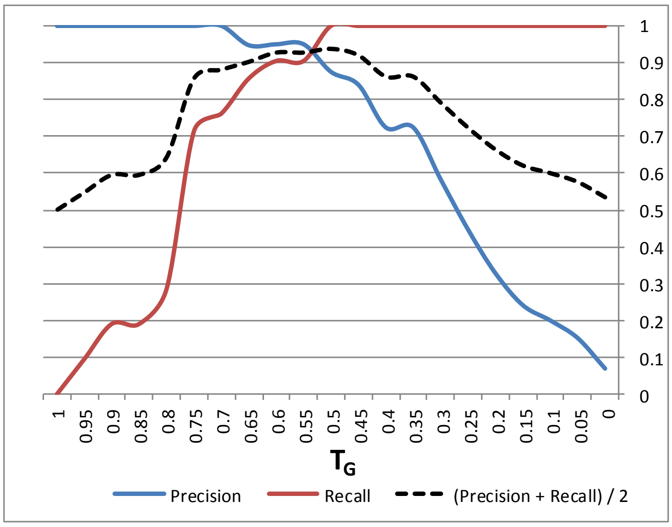

Figure 10. Changes in Precision, Recall and Average of Precision and Recall Due to Changes in T_G for Goal Event Detection

Copyright ⓒ 2015 SERSC

Another evaluation method is based on goal event detection which is performed by applying a threshold (T_G) to the importance of each highlight. In this method, after applying T_G, some goal events may be missed and some non-goal event may be detected. Thus, we can change T_G in the range of [0,1] and calculate precision and recall for each value of T_G. Generally, the best value of T_G depends on the application. To find the best threshold, we can plot the precision and recall for different thresholds (Figure 10). According to the results depicted in Figure 10, the maximum average value of precision and recall is achieved by applying $T_G = 0.5$. In this case, the precision and recall of 84.6% and 100% are achieved respectively. In a compromised status, it seems that using $T_G = 0.55$ leads to a better result. In this case, the precision and recall of 95.2% and 90.9% are achieved respectively.

According to the above description of the proposed objective evaluation for highlight detection in soccer video, the results depicted in Table 4 are achieved using $T_G = 0.55$.

Table 4. Results of Goal Event Detection

Video #	Number of Goal Events	Number of Detected Highlight	H_{avg}	False Positive	False Negative
1	1	31	1	0	0
2	3	28	1	0	0
3	3	34	1	0	1
4	2	33	1	0	0
5	0	17	1	0	0
6	1	23	1	0	0
7	2	27	1	0	1
8	3	32	1	0	0
9	0	23	1	0	0
10	2	19	1	1	0
11	1	39	1	0	0
12	4	28	1	0	0
Total	22	334	-	1	2

According to the results shown in Table 4, H_{avg} is 1 for all videos in datasets. It means that highlights that contain goal events have higher importance with respect to other highlights of the video. The average number of highlights per video is about 28 and the average number of goal events per video is 1.83. This means that the average probability of occurrence of goal event in a highlight is less than 6.6%. In this case ($T_G = 0.55$), precision and recall are 95.2% and 90.9% respectively. Thus, the mentioned results are very valuable considering that the probability of occurrence of goal is very low.

The highlight wrongly detected in Video #10 is related to the 2nd half of the match between Greece and Argentina. In this highlight, the goal keeper was injured. This part of the video was detected as a significant highlight because the cameraman showed the event in a long replay with many close-up shots. Although audience excitement was low, L_{Replay} and $P_{view\ type}$ were considered more important than the audience excitement. Thus, this event was detected as a goal event incorrectly. Some informative frames related to this event are shown in Figure 11. The importance of highlights related to this match is shown in Figure 12.

In Video #3, a highlight contains goal event but it is missed by the proposed approach. Although the replay of the missed goal event in Video #3 is relatively long (about 30 seconds) and audience excitement before replay is high, the ratio of the total length of the close-up or out-filed shots during the replay to the length of replay is medium. Therefore, the importance of this highlight is estimated as medium (0.5) by FIS. Some sample frames of this goal event are shown in Figure 13. Importance of highlights related to Video #3 is shown in Figure 14.

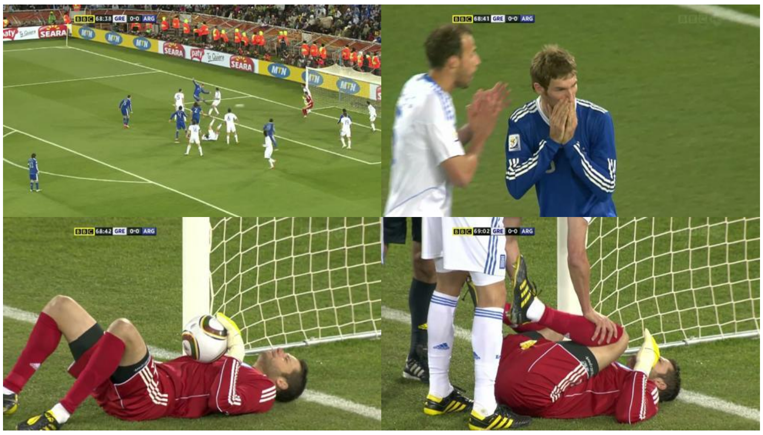

Figure 11. Some Informative Frames Related to an Event Incorrectly Detected as a Goal Event in Video #10

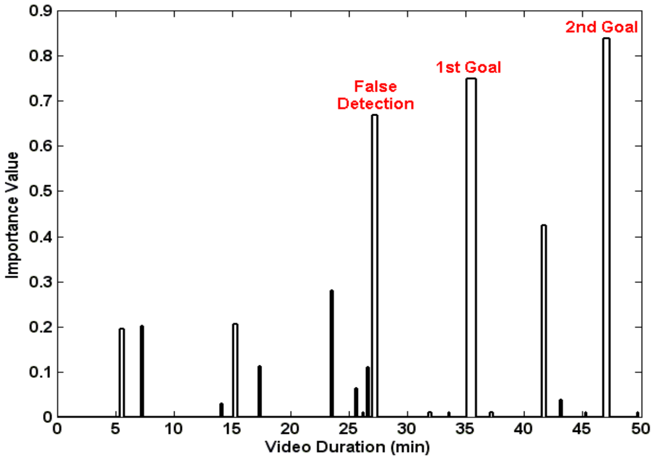

Figure 12. Importance of Highlights Related to Video #10 that Considered a Highlight, which Incorrectly Detected as a Goal Event

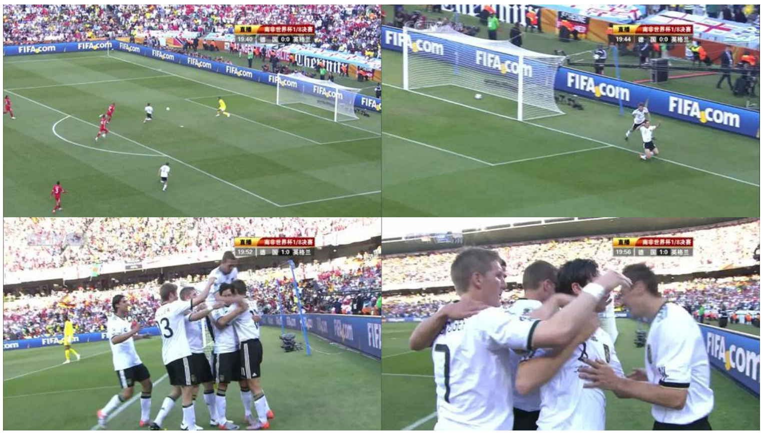

Figure 13. Some Sample Frames Related to a Missed Goal Event in Video #3

Copyright ⓒ 2015 SERSC

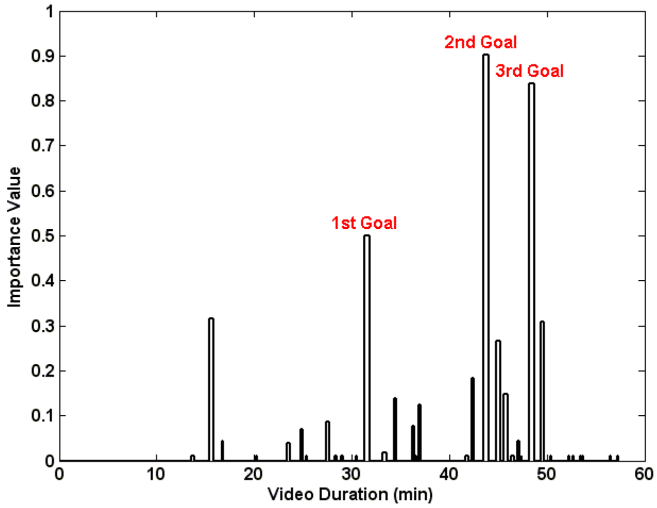

Figure 14. Importance of Highlights Related to Video #3 that Considered an Undetected Goal Event. The First Goal Event was Not Detected after Applying the Threshold

3.4. Subjective Evaluation of the Proposed Method

In the subjective evaluation experiments, 6 individuals have watched each full video and all the highlights of the video. There are three options for summarization: (1) long-time summarization (about 15 minutes long), (2) short-time summarization (about 5 minutes) and (3) goal-event summarization (only goal events, if any). To compose the long-time and short-time summarization, the highlights are sorted based on their importance, and then they are selected in order of importance while the total length of the selected highlights satisfies the time constraint. Finally, the selected highlights are put together in order of time to compose the final summarized video. For goal-event summarization, only highlights with importance higher than 0.55 is selected. Therefore, the goal-event summarization may be empty.

The summarized videos are composed by different summarization options and are shown to the individuals. Then, we ask the individuals to score the quality of summarization in the range of [0,1]. The scores assigned to each summarized video are presented in Figure 15.

According to the results shown in Figure 15, the overall quality of summarization in goal-event level is the best. These results show the scores of goal-event summarization are higher than other summarization types except for Video #3, #7 and #10. In addition, the overall quality of long-time summarization is better than short-time summarization. This shows that the selected highlights for short-time summarization are not as important for individuals. It seems that the main reason for this problem is related to the inability of the extracted feature to describe the contents of highlights properly. In other words, accuracy of the proposed method for detection of general highlights and goal events are very good, but the system is weak to estimate the actual importance of non-goal events.

Copyright ⓒ 2015 SERSC

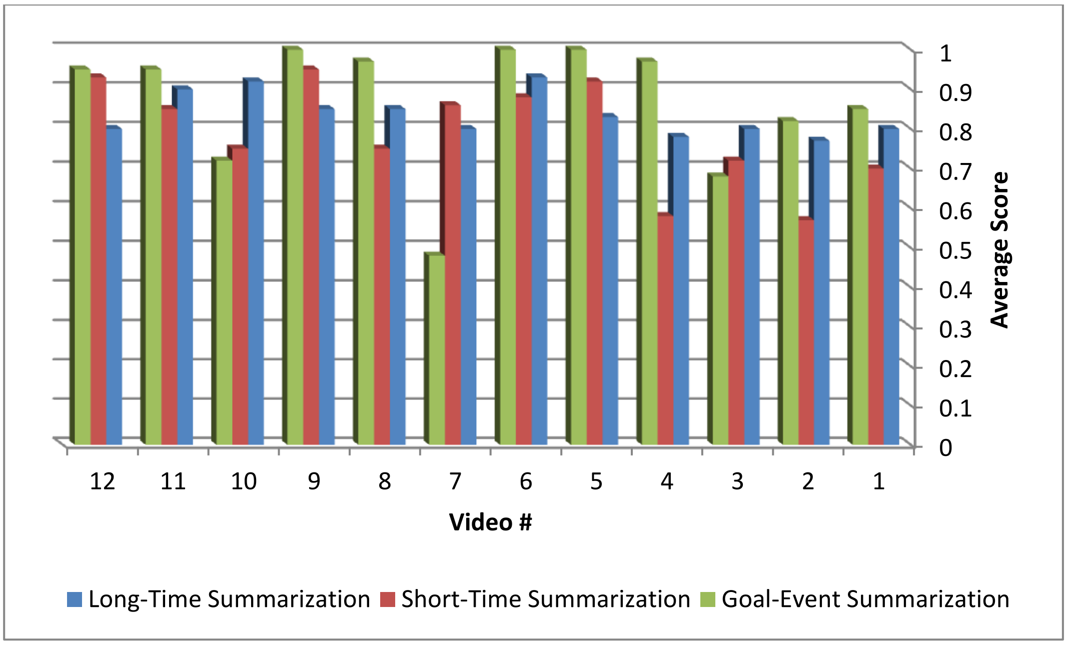

Figure 15. Scores Assigned to the Summarized Videos by Individuals

3.5. Evaluation of Execution Time

In this section, execution time for each part of the proposed method is evaluated. The proposed method extracts the features by an on-demand approach to increase the processing speed. The proposed method contains some functions that are called based on an on-demand approach. In other words, some functions are called for all frames and some functions are called rarely. Thus, we introduce two terms to calculate the execution time of each function: (1) average processing time per frame (t_f) and (2) average number of calls per second (n_s).

Shot boundary detection is the most frequent task performed on every frame. Because of the frame rate of our video is 25 fps, the functions related to shot boundary detection are called 25 times per second. Other functions related to logo detection, replay detection, view type recognition, audience excitement and highlight detection are called occasionally based on the content of the video. In Table 5, t_f and n_s are reported for the main functions of the proposed method.

Table 5. Execution Times of Different Parts of the Proposed Method

Function Name	t_f (mili-second)	n_s
RGB Histogram	5	25
Shot Boundary Detection	1	25
3D RGB Histogram	255	0.14
Logo Detection	2	0.14
Replay Detection	4	0.02
HSV Color Conversion	38	0.01
View Recognition	1	0.01
Short-Time Energy	1	0.16
Audience Excitement	6	0.01
Fuzzy Inference System	5	0.01
Post Processing	53	0.01

The proposed method is implemented in MATLAB R2011a on Windows XP SP3. The experimental results are given using a personal computer with an AMD 4200+ Dual Core CPU and 4 GB DDR2 memory. Although we use a dual core processor,

Copyright ⓒ 2015 SERSC

we do not use parallel computing toolbox of MATLAB. Therefore, all the parts reported in Table 5 are executed sequentially. However, MATLAB may use two cores of the processor simultaneously to execute low-level instructions of a given part automatically.

According to the experiments, the proposed method processes about 130 frames per second. However, the average processing time is less than 8 mili-seconds per frame, but in the worst case, the processing time is about 260 mili-seconds per frame. The worst case occurs when the algorithm wants to detect the logo during the frames of a gradual shot. In this case, calculating 3D RGB histogram for logo detection is very time consuming.

3.6. Comparative Results

Performing a subjective comparison and evaluation between different methods is very difficult. Therefore, we compare our proposed method with other methods for goal-event detection. Performance of the proposed method in goal event detection is at par with the recent achievements such as [14, 15, 18]. Primary processing, such as shot boundary detection and view type recognition, is the same in these systems but their mid- and high-level processing are different.

In [14], a method was proposed to summarize soccer videos based on cinematic and object-based features. This method extracts some features including slow-motion detection, penalty box detection, and referee detection. Additionally, it detects goal events to present different summaries. This system can summarize three types of videos: (1) all slow-motion segments, (2) all goal events, and (3) slow-motion segments classified based on appearance of a specific object. The system can generate type 1 or 2 summaries in real-time but cannot do it for type 3 summaries, because they are computationally complex. In this system, precision of goal event detection is low.

In [15], after video partitioning, logo and caption regions are detected to determine informative segments of the video. Then, other features such as vertical goal posts, goal net and audio loudness are detected. This system can detect goal and attack events in a soccer video.

In [18], a hierarchical structure, like a decision tree, was used for event detection in soccer video. In the proposed structure, a top-down method was used for video partitioning, highlight detection, view type recognition, as well as detection of player, goal keeper, and referee. Finally, goal events and save events were detected in the video. In this method, a priori knowledge was used to extract the association between the events.

Although using a standard dataset would provide fair comparison results, but there is not any standard dataset for broadcast soccer video analysis. Thus, the results reported in this section were achieved by reimplementation of the mentioned methods on our dataset. Table 6 shows that our proposed method outperforms other methods of goal event detection.

Table 6. Comparison Results between the Proposed Method and some Recent Methods for Goal-Event Detection

Method	Precision	Recall
Method in [14]	81.2%	60.0%
Method in [18]	83.3%	90.9%
Method in [15]	100%	77.3%
The proposed method ($T_G = 0.5$)	88.0%	100%
The proposed method ($T_G = 0.55$)	95.2%	90.9%

Although these methods based on heuristic rules achieved higher precision and recall rates, the methods presented in [16, 17], especially [17], are more general. These methods are adaptive to the training samples but they usually suffer from limited number of training samples. Therefore, the methods developed based on the machine learning approaches achieve lower accuracies than those developed based on the heuristic rules.

4. Conclusions and future works

In this paper, a fast highlight detection and scoring method is proposed for broadcast soccer video summarization. The proposed method partitions video to some highlights and then extracts low-level and mid-level features using an on-demand feature extraction approach to estimate importance of each highlight. Importance of each highlight is estimated using a FIS. According to the level of summarization determined by the user, highlights with higher importance are selected to compose the summarized clip.

The proposed method is evaluated by objective and subjective evaluations. Objective evaluation shows that the proposed method detects goal events with high precision and recall rate. However, there are a few errors in goal event detection, but all the highlights that contain goal events have higher importance than other highlights.

The main advantage of the proposed system is on-demand feature extraction. At first, only a few simple features are extracted from all frames to investigate the content of frames roughly. Then, more complicated features are extracted from the video gradually based on the results of the previous processing. In the last stage of processing, importance of each highlight is estimated by a FIS and finally, high rank highlights are selected to compose the summarized video. Average processing time of the proposed method is less than 8 mili-seconds per frame, but logo detection during the frames of a gradual shot is the worst case which takes about 260 mili-seconds per frame. Therefore, the proposed method is very fast and processes about 130 frames per second. Unlike many recent soccer video summarization systems, the proposed system offers acceptable quality with different options for broadcast soccer video summarization in real-time.

Another advantage of the proposed method is that it considers the continuity of the highlights. In other words, because the selected shots for a highlight are consequent, semantic continuity of highlight is preserved. However, some redundancy remains in highlight, especially in shots related to replay.

In the proposed method, highlight detection depends on the style of editor who has edited the video for broadcasting. If the editor decides to put a replay for an event in the broadcast video, the proposed method detects this event as a highlight and considers importance of the event; otherwise, it does not detect this event as a highlight and there is no chance that it appears in any summarization level for this event. Additionally, replay length and view type of shots are affective for calculating the importance of a highlight. Although very important events in soccer match (such as goal and penalty) have almost same importance for everyone, other events may have different significance in the mind of different individuals. For example, corner kick may be more important than free kick for a given person, but for another person, it may not be the same. In this case, highlight is detected but the importance of highlight may be estimated inaccurately. Inaccurate importance estimation of a highlight usually occurs for not very important events such as free kick, corner kick, foul and shooting to goal. Thus, the proposed method for highlight detection and summarization is affected by interest and style of the editor. Experimental results show that individuals enjoy more from goal-event summarization, which is the most important event of a soccer match. The scores are lower for long-time and short-time summarizations, especially for short-time summarization. This means that the proposed approach suffers from the weakness of the extracted feature which does not describe the content of highlights completely.

Copyright ⓒ 2015 SERSC

For future works, we can use other mid-level audio-visual features. For example, detection of referee whistling and goal mouth are some mid-level features that may increase the quality of highlight detection and summarization. These features can describe the content of the video more efficiently and therefore, the importance of highlights can be estimated more accurately.

To reduce the effects of editor's interests in highlight detection, content of all shots have to be investigated. In this case, highlight detection will be more independent to the interests of editor and probably the quality of summarization will increase, but the computational complexity of the system will also be increased.

Applications of the proposed highlight detection and summarization method are wide. It can be used for users who do not have enough time to watch the whole soccer video. Thus, the proposed method can detect highlights in real-time according to the user interest for rate of summarization. Additionally, this system can be used as a value added service on mobile networks for soccer fans who cannot watch soccer matches on TV.

Also, the proposed method may be used as a pre-processing tool for high-level processing of soccer videos. For example, it may be used for detection and recognition of important events occurred in highlights. If such a system is used as a pre-processing tool, detection and recognition of important events can be performed on off-the-shelf hardware in real-time.

Acknowledgements

The first author would like to thank all individuals who evaluate and score the summarized video patiently, especially Mr. Abbas Kahkhaei Javan and Mrs. Mahboobeh Kahkhaei Javan.

References

[1] T. Lotfi, M. Bagheri, A. A. Darabi and S. Kasaei, An Efficient Content-Based Video Coding Method for Distance Learning Applications, Scientia Iranica, 16, 2, **(2009)**.

[2] Z. Xiong, X. S. Zhou, Q. Tian, Y. Rui and T. S. Huang, Semantic Retrieval of Video - Review of Research on Video Retrieval in Meetings, Movies and Broadcast News, and Sports, IEEE Signal Processing Magazine, 23, 2, **(2006)**.

[3] M. H. Sigari, S. A. Sureshjani and H. Soltanian-Zadeh, Sport Video Classification using an Ensemble Classifier, Iranian Machine Vision and Image Processing, **(2011)**, November, Tehran, Iran.

[4] J. K. Aggarwal and M. S. Ryoo, Human Activity Analysis: A Review, ACM Computing Surveys, 43, 3, **(2011)**.

[5] Z. Niu, X. Gao and Q. Tian, Tactic Analysis based on Real-World Ball Trajectory in Soccer Video, Pattern Recognition, 45, 5, **(2012)**.

[6] A. G. Money and H. Agius, Video Summarisation: A Conceptual Framework and Survey of the State of the Art, Journal of Visual Communication and Image Representation, 19, 2, **(2008)**.

[7] V. Vijayakumar and R. Nedunchezhian, A Study on Video Data Mining, International Journal of Multimedia Information Retrieval, 1, 3, **(2012)**.

[8] A. Amiri and M. Fathy, Hierarchical Keyframe-based Video Summarization Using QR-Decomposition and Modified k-Means Clustering, EURASIP Journal on Advances in Signal Processing, **(2010)**.

[9] V. Kiani and H. R. Pourreza, Flexible Soccer Video Summarization in Compressed Domain, International eConference on Computer and Knowledge Engineering, **(2013)**, Mashhad, Iran.

[10] Y.-F. Ma, X.-S. Hua, L. Lu and H.-J. Zhang, A Generic Framework of User Attention Model and Its Application in Video Summarization, IEEE Transactions on Multimedia, 7, 5, **(2005)**.

[11] M. Y. Eldib, B. S. A. Zaid, H. M. Zawbaa, M. El-Zahar and M. El-Saban, Soccer Video Summarization using Enhanced Logo Detection, International Conference on Image Processing, **(2009)**, November, Cairo, Egypt.

[12] E.-J. Kim, G.-G. Lee, C. Jung, S.-K. Kim, J.-Y. Kim and W.-Y. Kim, A Video Summarization Method for Basketball Game, Pacific Rim Conference on Multimedia, **(2005)**, November, Jeju Island, Korea.

[13] Z. Zhao, S. Jiang, Q. Huang and G. Zhu, Highlight Summarization in Sports Video Based on Replay Detection, International Conference on Multimedia and Expo, **(2006)**, July, Toronto, Canada.

[14] A. Ekin, A. M. Tekalp and R. Mehrotra, Automatic Soccer Video Analysis and Summarization, IEEE Transactions on Image Processing, 12, 7, **(2003)**.

[15] H. M. Zawbaa, N. El-Bendary, A. E. Hassanien and T.-h. Kim, Event Detection Based Approach for Soccer Video Summarization Using Machine learning, International Journal of Multimedia and Ubiquitous Engineering, 7, 2, **(2012)**.

[16] M.-S. Hosseini and A.-M. E. Moghadam, Fuzzy Rule-based Reasoning Approach for Event Detection and Annotation of Broadcast Soccer, Applied Soft Computing, 13, 2, **(2013)**.

[17] D. W. Tjondronegoro and Y.-P. P. Chen, Knowledge-Discounted Event Detection in Sports Video, IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 40, 5, **(2010)**.

[18] M. H. Kolekar, K. Palaniappan, S. Sengupta and G. Seetharaman, Semantic Concept Mining Based on Hierarchical Event Detection for Soccer Video Indexing, Journal of Multimedia, 4, 5, **(2009)**.

[19] H. Zhou, A. H. Sadka, M. R. Swash, J. Azizi and U. A. Sadiq, Feature Extraction and Clustering for Dynamic Video Summarisation, Neurocomputing, 73, 10-12, **(2010)**.

[20] E. Spyrou, G. Tolias, P. Mylonas and Y. Avrithis, Concept Detection and Keyframe Extraction using a Visual Thesaurus, Multimedia Tools and Applications, 41, 3, **(2009)**.

[21] M. Chatzigiorgaki and A. N. Skodras, Real-Time Keyframe Extraction Towards Video Content Identification, International Conference on Digital Signal Processing, **(2009)**, Santorini, Greece.

[22] S. Frintrop, E. Rome and H. I. Christensen, Computational Visual Attention Systems and their Cognitive Foundations: A Survey, ACM Transactions on Applied Perception, 7, 1, **(2010)**.

[23] F. Ahmadi, M. H. Sigari and M.-E. Shiri, A Rank based Ensemble Classifier for Image Classification using Color and Texture Features, Iranian Conference on Machine Vision and Image Processing, **(2013)**, Iran.

[24] H. Pan, B. Li and M. I. Sezan, Automatic Detection of Replay Segments in Broadcast Sports Programs by Detection of Logos in Scene Transitions, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), **(2002)**, May, Orlando, FL, USA.

[25] L. Wang, B. Zeng, S. Lin, G. Xu and H.-Y. Shum, Automatic Extraction of Semantic Colors in Sports Video, International Conference on Acoustics, Speech, and Signal Processing, **(2004)**, May.

[26] F. Gasparini and R. Schettini, Skin Segmentation using Multiple Thresholding, Internet Imaging VII, **(2006)**.

[27] M. H. Sigari, H. Soltanian-Zadeh, V. Kiani and H. R. Pourreza, Counterattack Detection in Broadcast Soccer Videos using Camera Motion Estimation, International Symposium on Artificial Intelligence and Signal Processing, **(2015)**, Iran.

Authors

Mohamad-Hoseyn Sigari received BS and MS degrees in computer engineering from Ferdowsi University of Mashhad and Iran University of Science and Technology, in 2006 and 2008 respectively with first rank honor. At present, he is a PhD. candidate in the school of electrical and computer engineering, college of engineering, University of Tehran, Tehran, Iran. His research interests are machine vision, semantic video analysis, biometrics and applications of machine vision in science and industry.

Hamid Soltanian-Zadeh received BS and MS degrees in electrical engineering: electronics (with honors) from University of Tehran, Tehran, Iran in 1986 and MSE and PhD degrees in electrical engineering: systems and bio-electrical sciences from the University of Michigan, Ann Arbor, Michigan, USA, in 1992. He is currently a full Professor and a founder of Control and Intelligent Processing Center of Excellence (CIPCE) in the school of electrical and computer engineering at the University of Tehran. He is also a senior scientist and head of medical image analysis group in the department of radiology, Henry Ford Health System, Detroit, Michigan, USA. His research interests include medical imaging, signal and image processing, pattern recognition and machine vision.

Copyright ⓒ 2015 SERSC

Hamid-Reza Pourreza received BS degree in electrical engineering from Ferdowsi University of Mashad, Iran, in 1989. He received MS. and PhD. degrees in electrical engineering and computer engineering from Amirkabir University of Technology, Iran, in 1992 and 2003 respectively. Now, he is an associate professor in the department of computer engineering, Ferdowsi University of Mashad, Iran. He is establisher and head of machine vision lab and one of establishers of Eye Images Analysis Research Group (EIARG) at Ferdowsi University of Mashhad. His research interests are signal and image processing, machine vision and intelligent transportation systems.

Copyright ⓒ 2015 SERSC