A Framework for Soccer Video Processing and Analysis Based on Enhanced Algorithm for Dominant Color Extraction

Youness TABII youness.tabii@gmail.com
ENSIAS/GL
BP 713, Mohammed V University-Soussi
Rabat, 10000, Morocco

Rachid OULAD HAJ THAMI oulad@ensias.ma
ENSIAS/GL
BP 713, Mohammed V University-Souissi
Rabat, 10000, Morocco

Abstract

Video contents retrieval and semantics research attract a large number of researchers in video processing and analysis domain. The researchers try to propose structure or frameworks to extract the content of the video that's integrating many algorithms using low and high level features. To improve the efficiency, the system has to consider user behavior as well as develops a low complexity framework. In this paper we present a framework for automatic soccer video summaries and highlights extraction using audio/video features and an enhanced generic algorithm for dominant color extraction. Our framework consists of stages shown in Figure 1. Experimental results demonstrate the effectiveness and efficiency of the proposed framework.

Keywords: Video processing, Summary and Highlight, Finite state machine, Binarization, Text detection, OCR.

1. INTRODUCTION

Soccer video attracts a wide range of audiences and it is generally broadcasted for long hours. For most non-sport viewers and some sport fans, the most important requirements are to compress its long sequence into a more compact representation through a summarization process. This summarization has been popularly regarded as a good approach to the content-based representation of videos. It abstracts the entirety with the gist without losing the essential content of the original video and also facilitates efficient content-based access to the desired content.

In literature, many works have been proposed in the last decade concern sport videos processing and analysis. The sport videos are composed of play events, which refer to the times when the ball is in-play, and break events, which refer to intervals of stoppages in the game. In [1], the play-break was detected by thresholding the duration of the time interval between consecutive long shots; Xu et al. [2] segmented the soccer game into play-break by classifying the respective visual patterns, such as shot type and camera motion, during play, break and play/break transition.

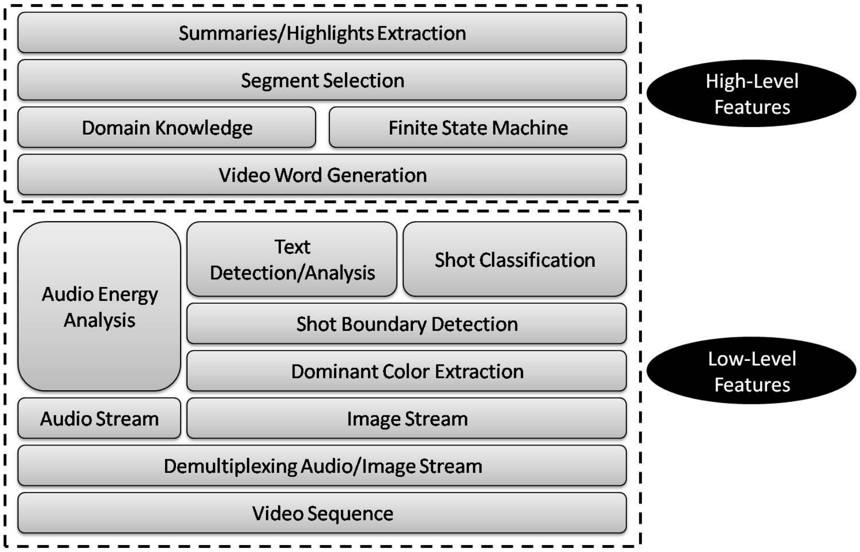

FIGURE 1: The proposed framework.

The highlight detection research aims to automatically extract the most important, interesting segments. In [3] introduced a generic game highlight detection method based on exciting segment detection. The exciting segment was detected using empirically selected cinematic features from different sports domains and weighted combine to obtain an excitement level indicator. Tjondronegoro et al. [4] detected the whistle sounds, crowd excitement, and text boxes to complement previous play-breaks and highlights localization method to generate more complete and generic sports video summarization.

Replay scenes in broadcast sports videos are excellent indicators of semantically mportant segments. Hence, replay scene detection is useful for many sports highlight generation applications. The characteristic of replay scenes are: 1) usually played in a slow-motion manner. Such slow-motion effect is produced or by repeating frames of the original video sequence. 2) Playing the sequence captured from a high- speed camera at the normal frame rate. Replay detection techniques have been proposed. In [5] used Bayesian Network together with six textual features extracted from Closed Caption to detect replay shot. Pan et al. [6] proposed to detect the transition effect, e.g. flying-logo, before and after the replay segments to detect replay. In [7] introduced a method to automatically discover the flying-logo for replay detection.

In the next of this paper, we will present each block/algorithm in our framework. We begin by representing the enhanced algorithm for dominant color extraction. Second, the shot boundary detection. Third, shots classification into defined classes. Fourth, the extraction of the audio descriptor. Fifth, the score box detection and text recognition, and as the final step, with domain knowledge we use finite state machine to extract the summaries and highlights.

2. DOMINANT COLOR EXTRACTION

The field region in many sports can be described by a single dominant color. This dominant color demonstrates variations from one sport to another, from one stadium to another, and even within one stadium during a sporting event.

In this section we will present our enhanced algorithm for dominant color extraction in soccer video. Adding that the algorithm can be used for other sports game like US football, golf and any sport games that have the play field color is green.

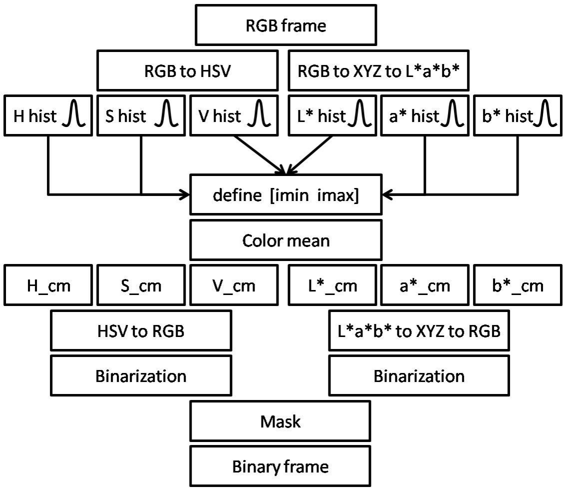

FIGURE 2: Dominant color stages.

The flowchart of the proposed algorithm is given in Figure 2. First, the system converts the **RGB** frames to **HSV** and to **L*a*b*** frames. After the conversion of frames, the system compute the histograms for each component (**H**, **S**, **V**, **L***, **a*** and **b***) using the following quantization factor: ***64*** bins for **H** and **L***, ***64*** bins for **S** and **b*** and ***128*** bins for **V** and **a***. Next, in order to remove noise effects that may result from using a single histogram index, the peak index ($\mathbf{i_{peak}}$), for each histogram is localized to estimate the mean value of dominant color for the corresponding color component. An interval about each histogram peak is defined, where the interval boundaries $\mathbf{[i_{min}, i_{max}]}$ correspond to the dominant color. Which given in Equations (1).

$$\begin{aligned} &\sum_{i=i_{\min}}^{i_{peak}} H[i] <= 2H[i_{peak}] \quad and \quad \sum_{i=i_{\min}-1}^{i_{peak}} H[i] > 2H[i_{peak}] \\ &\sum_{i=i_{peak}}^{i_{\max}} H[i] <= 2H[i_{peak}] \quad and \quad \sum_{i=i_{peak}}^{i_{\max}+1} H[i] < 2H[i_{peak}] \end{aligned} \tag{1}$$

After the interval boundaries are determined, the mean color in the detected interval is computed by Equation (2) for each color component, and we get $\mathbf{H_{cm}}$, $\mathbf{S_{cm}}$, $\mathbf{V_{cm}}$, $\mathbf{L^*_{cm}}$, $\mathbf{a^*_{cm}}$ and $\mathbf{b^*_{cm}}$.

$$colormean = \frac{\sum_{i=i_{\min}}^{i_{\max}} H[i]*i}{\sum_{i=i_{\min}}^{i_{\max}} H[i]} \tag{2}$$

Next step, the system convert the mean color of each color component into ***RGB*** space to binarize the frame using the Equation (3). Where I_R, I_G, I_B are the matrix of *Red*, *Green*} and *Blue* components in the ***RGB*** frame. **K**, $\mathbf{R_t}$, $\mathbf{G_t}$, $\mathbf{B_t}$ and $\mathbf{G_{th}}$ are the thresholds for binarzation. ***G(x,y)*** is the binarized frame.

$$G(x,y)=\begin{cases}1 \quad if & \begin{cases}I_G(x,y)>I_R(x,y)+K(G_{cm}-R_{cm})\\ I_G(x,y)>I_B(x,y)+K(G_{cm}-B_{cm})\\ \left|I_R-R_{cm}\right|<R_t\\ \left|I_G-G_{cm}\right|<G_t\\ \left|I_B-B_{cm}\right|<B_t\\ I_G>G_{th}\end{cases}\\ 0 \quad otherwise\end{cases} \tag{3}$$

The last step in the system, is the obtaining the final binarized frame, by using the two binarized frame comes from HSV space and from L*a*b space. We use the binarized frame with HSV space as base and the binarized frame with L*a*b as mask, to improve the quality of the final frame.

$$\begin{cases}if \quad \sum Ibin_{hsv} > T_{th*}N^2 \quad then \quad Ibin^N_{hsvLab}=Ibin^N_{hsv}\\ \qquad else \quad Ibin=Ibin^N_{Lab}\end{cases} \tag{4}$$

The Equation (4), shows the application of binarized frame with L*a*b space as mask into binarized frame with HSV space. Where ***N*** is the block size, ***T_{th}*** the threshold, **$Ibin_{hsv}$** is the binarized frame using HSV space, **$Ibin_{Lab}$** the binarized frame using L*a*b* and **$Ibin_{hsvLab}$** is the resulted binarized frame using the two colors spaces.

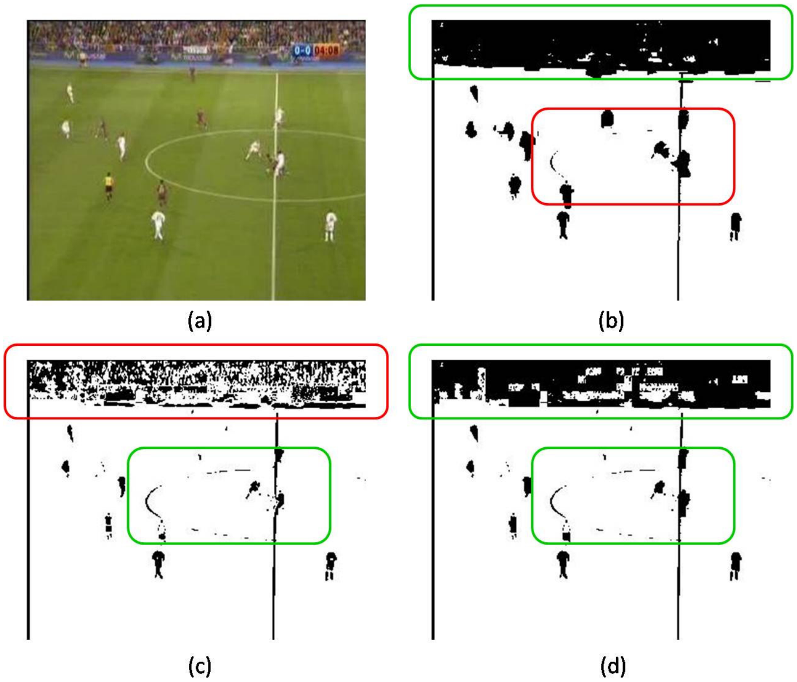

FIGURE 3: Dominant color extraction result.

Figure 3 show the result of dominant color extraction algorithm obtained in soccer video. Fig 3(a) it's the *RGB* frame, Fig 3(b) binarized frame comes from *HSV* space, Fig 3(c) binarized frame from L*a*b and Fig 3(d) it's the binarized frame using two color spaces and equation (4). This algorithm have generic behavior, we can apply it in US football and golf sport's game.

3. SHOT DETECTION

The shot is often used as a basic unit for video analysis and indexing. A shot may be defined as a sequence of frames captured by "a single camera in a single continuous action in time and space". The extraction of this unit (shot) still presents problems for sports video.

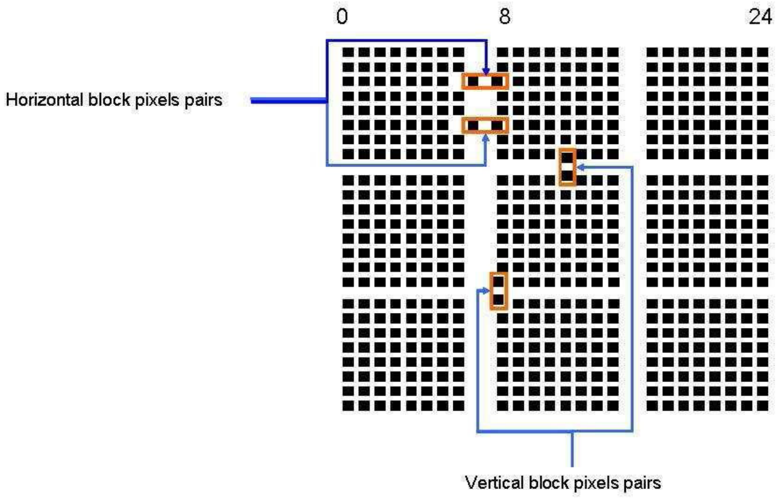

FIGURE 4: Adjacent pixels.

For shot detection, we use our algorithm based on Discrete Cosine Transform mutli-resolutions (DCT-MR) [8]. We binarize each I/P frames and we divide them into blocks of ***2^R*2^R*** (**R** is the resolution) and for every block we calculate the DCT then we calculate the vertical distance ***distV*** (Eq (5)) and the horizontal distance ***distH*** (Eq (6))between adjacent pixels in blocks (Figure 4), then the means of both distances ***distHV*** is computed using equation (7).

$$distV = \sum_{i=1}^{\frac{w-R}{R}} \sum_{j=1}^{h} \frac{\left| pixel_{Rij} - pixel_{R(i+1)j} \right|}{h(w-R)/R} \tag{5}$$

$$distH = \sum_{i=1}^{w} \sum_{j=1}^{\frac{h-R}{R}} \frac{\left| pixel_{iRj} - pixel_{iR(j+1)} \right|}{w(h-R)/R} \tag{6}$$

$$distHV = \frac{distH + distV}{2} \tag{7}$$

Where ***w*** and ***h*** are the width and the height of frame and ***R*** is the resolution. We compare the distance ***distHV*** with the threshold ***0.12*** in order to decide if there is shot change or not [8].

4. SHOT CLASSIFICATION

The shots classification usually offers interesting information for the semantics of shots and cues. In this algorithm we used our algorithm in [9], which is a statistical method for shots classification. The method is based on spatial segmentation using 3:5:3 dividing format of the binarized key frames extracted previously.

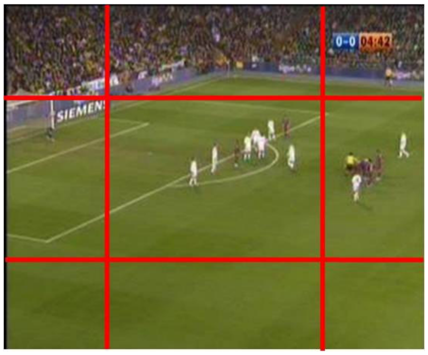

FIGURE 5: 3:5:3 division frame format.

The figure 5 shows the division 3:5:3 format of binarized frame. The shot classification algorithm in [9] gives very promising results in soccer video. We classify shot into four classes [9]: ***Long Shot***, ***Medium Shot***, ***Close-up Shot*** and ***Out Field Shot***.

5. AUDIO DESCRIPTOR

Adding to the color information in video track, the Audio track also bears important information that should not be ignored and, as well, gives the semantic to the video, especially in action and sport videos. In the case of soccer, when there is an important moment (goal, penalty, … etc), the voice intensity of the commentator's rises and falls proportionally with the action. This intensity of the audio track can be used as descriptor to more characterize the video.

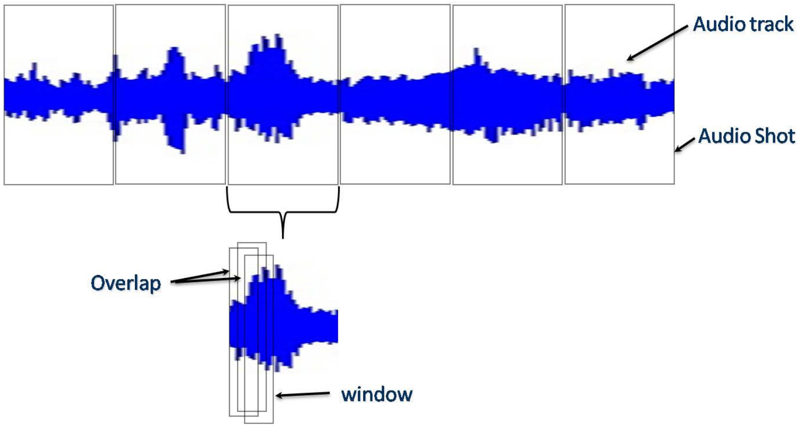

FIGURE 6: Windowing of the audio track.

In our method of soccer summaries/highlights extraction, we use the windowing algorithm with parameters: window size is ***10 seconds*** and the overlap parameter is ***2 seconds*** as shown in Figure 6; to compute the energy for each window in shot using Equation (8).

$$E_{shot,window} = \frac{1}{N}\sum_{i=1}^{N} S^2 \qquad (8)$$

In Equation (8), we compute the energy in each window of shot audio track, where **N** represents the number of samples in window of shot audio track and **S** is the set of samples in window of shot audio track.

After the energy computing, we extract the Maximum and the Minimum of energy of each shot; at the end of this step we get a vector of shot's Max and Min energy.

$$Max\_index = \arg Max(E_{shot}) \tag{9}$$

$$Min\_index = \arg Min(E_{shot}) \tag{10}$$

We make use this ***Max energy*** vector of shots to generate the candidate highlight of audio track. The *Max_index* and *Min_index* are used as indexes in XML file (section Audio Video XMLisation Block).

6. SCORE BOX DETECTION AND TEXT RECOGNITION

Text in video is normally generated in order to supplement or to summarize the visual content and thus is an important carrier of information that is highly relevant to the content of the video. As such, it is a potential ready-to-use source of semantic information.

Figure 7 show the stages of our algorithm for score box extraction and text recognition in soccer video. The first step consist of extraction a clip from the whole video, the length of the video clip extracted is **L**. The score box sub-block Extraction stage is the second step, that is based on soccer video editing. After, score box detection based on motion vector using Diamond Search (DS) algorithm. Next, the text detection in score box with binarization and number of morphology constraint (pre-processing). Finally, we use Optical Character Recognition (OCR) algorithm to generate an ASCII file contain the text detected in score box of soccer video.

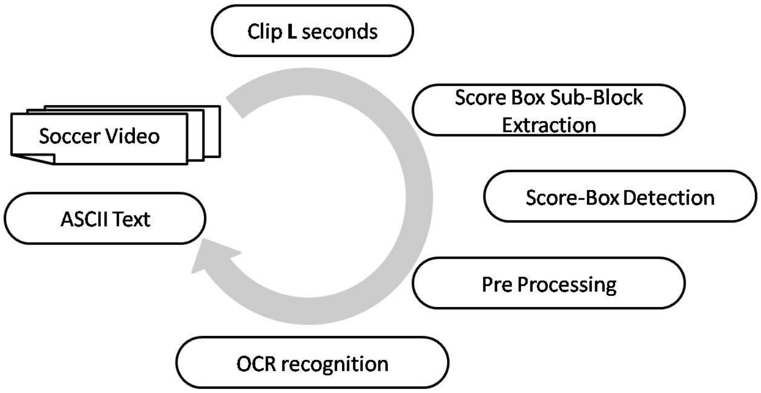

FIGURE 7: Score box detection flowchart.

In soccer video, the editors put the result of match and the names of the both teams on the top corners (left corner or right corner) (Figure 8). This box (teams' names and/or the result) remain superimposed in the video on the same place during the match.

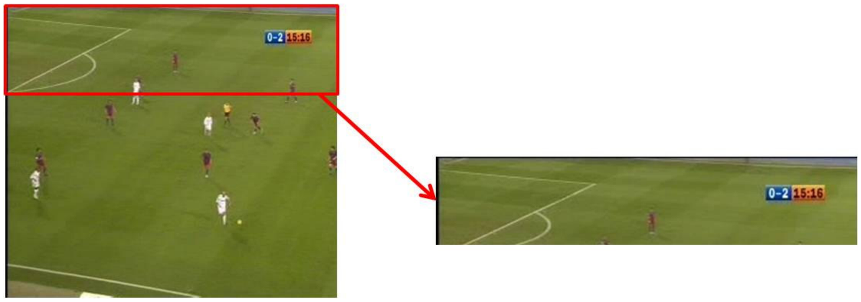

FIGURE 8: Sub-Block extraction.

With this style of video editing in soccer game, we extract the top block of frame which is about **30%** of screen video size (To optimize the search in DS algorithm), where we investigate the text superimposed in the video.

The property of stability of score box in the same place (corner) allows us to extract the score-box using motion vector (MV). To extract the score-box, we get ***L second*** (***L*** is the length of the clip in second for test) from the soccer video and we use the Diamond Search (DS) algorithm to extract motion vector of sub-block extracted previously (Figure 8). In our work we use DS algorithm with follows parameters: the Macro Block size is **8** pixels and the Search Parameter **p** size is **7** pixels.

To compute the motion vector of clip, we make a sampling of one from $\mathbf{S_{fr}}$ frames (we define $\mathbf{S_{fr}}$ in section of experiments). At the end, we get the motion vectors of all sub-blocks of all selected frames in sampling.

After getting the vector motion of all selected frames in clip of **L** seconds, we compute the 2D variance using Equation (11)

$$\sigma\_i_{t+1}{}^{t} = \sqrt{\frac{\sum_{l=1}^{M}\sum_{c=1}^{N}(MV_i(t+1)_{lc} - MV_i(t)_{lc})^{2}}{M*N-1}} \qquad (11)$$

Where **M** and **N** are the height and the width of the matrix **MV** respectively, and **i** refers to the samples number **i**.

$$\sigma\_i_{mean} = \frac{1}{2*NS_{fr}*p}\sum_{t=1}^{K-1}\sigma\_i_{t+1}^{t} \qquad (12)$$

In Equation (12), $\mathbf{NS_{fr}}$ represents the number of samples and **k** is the number of macro blocks in DS algorithm and **p** is the search parameter.

$$\begin{cases} if \quad \sigma^{i}{}_{mean} < T_{thd}\ then \quad i\_BlockStatic \\ else \quad i\_BlockNotStatic \end{cases} \qquad (13)$$

The last step in searching of the static macro blocks that are candidate belong to the score box, we use the Equation (13) for this purposes, where $\mathbf{T_{thd}}$ is the threshold.

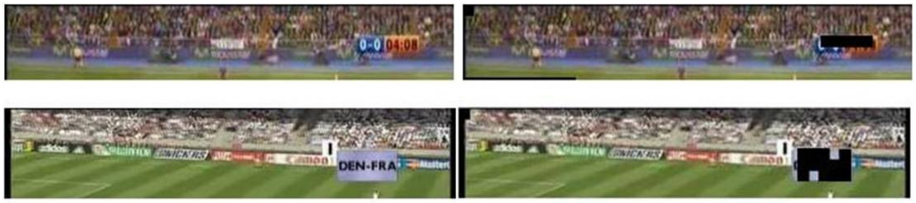

FIGURE 9: Score box detection result.

Figure 9 shows the result obtained for score box detection in clip video of **L = 3** second, sampling of **1** frame form **10** frames **S_{fr}=1/10**, the DS algorithm with parameters: macro block of **8** pixels, search parameter of **7** pixels, and **T_{thd}=0.2** in Equation (13).

In order to extract the totality of the score box and eliminate the border effect, we delete one macro block from each side of sub-block and we add one macro block form each side of score-box detected. The Figure 10 shows the final result obtained after this procedure.

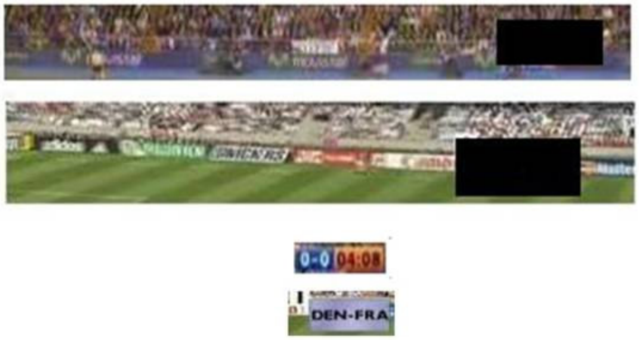

FIGURE 10: Score box extraction result.

After score-box extraction, we perform the binarization of the edge map in order to separate text-containing regions from the rest of the score-box using Otsu's global thresholding method as described in [10]. To reduce the noise and to connect loose characters form complete words, a number of morphological operations are performed as pre-processing. In Our case the morphological operations consist of the following steps:

- Step 1: 2x2 median filters.
- Step 2: 2x1 dilation.
- Step 3: 2x2 opening.
- Step 4: dilation in the horizontal direction using a 3x1.

The steps from 1 to 4 are repeated four times to avoid the problems mentioned above.

To complete our method and achieve the aims cited above, we used the freely optical recognition software know as ClaraOCR [11] to generate ASCII files.

7. AUDIO VIDEO XMLISATION

In this section we present a description of soccer video in XML file format. In sport video analysis, the Shot detection, shot classification, score extraction, key frames extraction and audio track analysis are very important steps for soccer video summaries and highlights extraction.

To give meaning to these steps, we present the video in a XML file format for future uses (section Summaries/Highlight Extraction). The structure adopted in our algorithm presented in figure bellow :

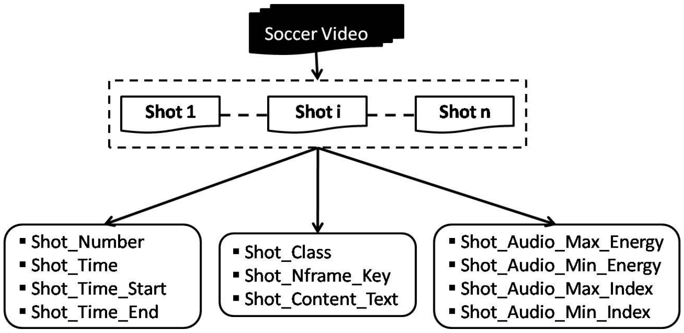

FIGURE 11: Adopted XML video file format.

In this structure we make used time as new factor, we shall use this factor in Finite State Machine (FSM) to generate the video summaries and highlights.

8. FINITE STATE MACHINE

In this section we present the finite state machine which modelizes soccer video. This finite state machine used with domain knowledge to generate dynamic summaries and highlights of soccer video (Figure 12).

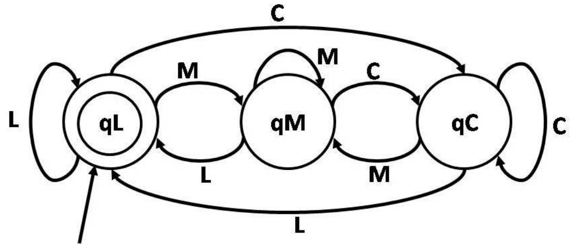

FIGURE 12: Adopted Finite State Machine for Soccer video (FSM-SC).

A Finite State Machine (FSM) is a 5-tuple $\mathbf{M} = \{\mathbf{P}; \mathbf{Q}; \boldsymbol{\delta}; \mathbf{q_0}; \mathbf{F}\}$, where **P** is the alphabet, **Q** is a set of state, $\boldsymbol{\delta}$ is a set of transition rules: $\boldsymbol{\delta} \subseteq (\mathbf{Q}*\sum \mathbf{U}\ \varepsilon*\mathbf{Q})$, $\mathbf{q_0} \in \mathbf{Q}$ is a set of initial (or starting) states and $\mathbf{F} \in \mathbf{Q}$ is a set of final states. In our case of soccer video, **P** is the alphabet which presents the summary, $\mathbf{Q} = \{\mathbf{qL};\mathbf{qM};\mathbf{qC}\}$, where **qL** is a Long shot, **qM** is a Medium shot and **qC** is a Close-up shot, $\boldsymbol{\delta}$ is the domain knowledge, $\mathbf{q_0} = \{\mathbf{qL}\}$, and $\mathbf{F} = \{\mathbf{qL}\}$.

Each video can be represented as a set of shots and scenes. The same, soccer video is a set of long, medium, close-up and out of field shots (for example: LLMMMCCMMLCMMMCCCLLLMLLCLMCCMLCM\dots).

Our Finite State Machine of soccer video (FSM-SC) will seek this **word** to find the **sub-words** that can be candidate as important moment (highlight) or can be as summary.

9. DOMAIN KNOWLEDGE

Domain knowledge is defined as the content of a particular field of knowledge, also the sum or range of what has been perceived, discovered, or learned.

Using the domain knowledge of soccer video which are the sum of the concepts and the steps followed by the most of editors of soccer match videos, we deduce a set of features specific to this domain, that allow as to more understand the structure of this type of videos.

- There is a limited number of shot views, Long view, Medium view, Close-up view and Out of field view.
- After every important moment in the match (goal, penalty,…), the editors make replays and/or slow motions.
- The cameraman always tries to follow all the actions: ball, goals, players, referee\dots etc.
- In order that TV viewers could see the action better, the cameraman puts the action in the middle of the screen.
- In case of goal, the cameraman keeps an eye on the striker player who reaches the goal.

10. SUMMARIES AND HIGHLIGHTS EXTRACTION

The last step of our algorithm of summaries and highlights extraction in soccer video is shown in Figure 13.

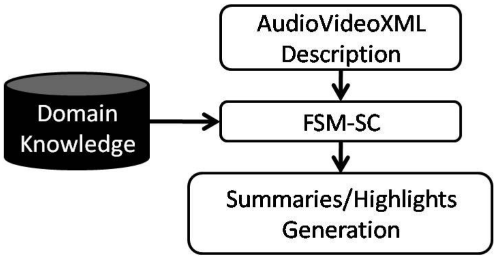

FIGURE 13: Last step in highlight extraction algorithm.

We make use the audio/video features: shot boundaries, shot classification, text in score box and Energy of audio presented in XML file and our FSM proposed to generate summaries and highlights of soccer video. Our framework use also the knowledge domain (set of rules) that helps the FSM to extract the video segments candidate to be summaries or highlights.

11. EXPERIMENTAL RESULTS

To show the effectiveness of our method that brings together the basic elements of the analysis of the video (shot detection and shot classification), we make use of five clips from five matches of soccer video. All the clips are in MPEG format, 352x288 size, 1150 kbps, 25 frames per second; dominant color ratio is computed on I- and P-frames.

Clip	Mpeg	Length	Goals	Fps	Total frames
clip soccer_1	2	10 min 22 sec	2	25	15551
clip soccer_2	2	13 min 06 sec	1	25	19652
clip soccer_3	2	16 min 55 sec	1	25	25380

TABLE 1: Clips information.

Table 1 shows a brief description of soccer video clips.
Where:

Clip Soccer_1: Match: FC Barcelona vs Real Betis.
Clip Soccer _2: Match: FC Barcelona vs Real Madrid.
Clip Soccer _3: Match: Cameroon vs Tunisia.

Clip	Summaries	Highlights	Detected Goals	Nothing	Missed
clip soccer_1	4	4	2	0	0
clip soccer_2	3	3	1	0	0
clip soccer_3	4	3	1	1	0

TABLE 1: Result of highlight and summaries extraction

.

Table 2 shows the obtained results of FSM, where **Summaries** presents the number of summaries generated by our framework, **Highlights** presents the number of important moment detected in clip, **Goals** is the number of goal detected in clip and **Nothing** presents summary of the clip in peaceful moments. Finally, **Missed** is the number of highlights missed by the framework in the clip.

12. CONSLUSION & FUTURE WORK

In this paper we present a full and efficient framework for highlights and summaries extraction in video soccer using audio/video features based on an enhanced and generic method for dominant color extraction.

The framework leaves much more for improvement and extension: there are other relevant low-level features that might provide complementary information and may help improve performance, such as camera motion, edge and higher-level object detectors.

13. REFERENCES

1. A. Ekin, A. M. Tekalp and R. Mehrotra. “Robust dominant color region detection with applications to sports video analysis”. In Proceedings of IEEE ICIP, vol. 1, pp. 21-24, 2003

2. P. Xu, L. Xie, S. Chang, A. Divakaran, A. Vetro and H. Sun. “Algorithms and system for segmentation and structure analysis in soccer video,” In Proceedings of IEEE ICME, pp. 928-931, 2001

3. A. Hanjalic. “Adaptive extraction of highlights from a sport video based on excitement modeling”. IEEE Trans. on MultiMedia, pp. 1114-1122, 2005

4. D. Tjondronegoro, Y.-P. Chen and B. Pham, “Highlights for more complete sports video summarization,” IEEE Trans. on Multimedia, pp. 22-37, 2004

5. N. Babaguchi and N. Nitta. “Intermodal collaboration: a strategy for semantic content analysis for broadcasted sports video”. IEEE ICIP, vol. 1, pp. 13-16, 2003

6. B. Pan, H. Li and M. I. Sezan; “Automatic detection of replay segments in broad-cast sports programs by detection of logos in scene transitions”. IEEE ICASSP, pp. 3385-3388, 2002

7. X. Tong, H. Lu, Q. Liu and H. Jin. “Replay detection in broadcasting sports videos”. ICIG, 2004

8. Y. Tabii and R. O. H. Thami. “A new method for soccer shot detection with multi-resolution dct”. COmpression et REprsentation des Signaux Audiovisuels (CORESA), 2007

9. Y. Tabii, M. O. Djibril, Y. Hadi and R. O. H. Thami. “A new method for video soccer shot classification,” 2nd International Conference on Computer Vision Theory and Applications(VISAPP), pp. 221-224, 2007

10. C. Wolf and J. M. Jolion. “Extraction and recognition of artificial text in multimedia documents”, Technical report (2002)

11. ClaraOCR download website : http://www.claraocr.org