On-line Key Frame Extraction and Video Boundary Detection using Mixed Scales Wavelets and SVD

Assma Azeroual, Karim Afdel, Mohamed El Hajji and Hassan Douzi.

Abstract—A video is a set of successive frames (images), one minute of a video stream can contain 1500 frames, but just some of them are the most representative, these frames are called key frames. The huge number of video frames requires a high computational cost on time and memory. Hence it's necessary to find new techniques to improve the video processing like video indexing, video retrieval and video summary especially when the real-time computing is required. In this context, this paper proposes a novel technique to extract key frames and detect video boundary based on dominants blocks of Faber-Schauder wavelet coefficients in mixed scales representation and Singular Value Decomposition (SVD). The reason behind using dominants blocks is that local features like contours or edges are unique to each frame, thus, they can act as a signature of the frame. These contours and its near textures contain an important concentration of dominant coefficients which are used to select the dominant blocks. Any substantial change in a video frame will result in a change of their edges and the neighboring textures of these edges, therefore an important change in the dominants blocks. Then this frame is considered as a key frame and represent the beginning of a new shot. The dominant blocks of every frame are computed, then feature vectors are extracted from the dominant blocks image of each frame and arranged in a feature matrix. After that, Singular Value Decomposition is used to calculate sliding windows ranks of those matrices. At the end, the computed ranks are traced to extract key frames of a video. The experimental results indicate that the proposed method is robust against a large range of digital effects used during shot transition and detect effectively the video shots and key frames.

Keywords—Key Frame Extraction, Shot Detection, Faber-Schauder wavelet, SVD, Mixed Scales Representation.

I. INTRODUCTION

THE proliferation of digital media (audio, image and video) is creating a pressing need for proper techniques to deal with the cost of signal processing, particularly for video processing which demand more time and memory. With the development of information technology, there has been a huge increase in video data which requires efficient techniques for retrieval, indexing and storage of this data.

Video is a rich and convenient way to get information due to advanced and friendly multimedia techniques available [1]. A great amount of video content can be found on-line as well as on devices of almost every user. It is a challenge to handle, index, sort of these contents without manual human help [2].

A video is composed by a series of basic units called frames (images) at a certain rate, for the human eye the rate at which it can distinguish images is 20 FPS (frame per seconds). Hence, films or television programs are projected at an average rate of 30 FPS. All the frames on a video have the same size and the time is equal between each two frames. The series of interrelated consecutive frames taken continuously by a single camera and representing a continuous action in time and space are called video shot [3]. These shots are joined together by editing operations which can contain transition effects or not. There are two kinds of shot changes namely, abrupt changes and gradual transitions. Abrupt changes usually result from camera breaks, while gradual transitions are produced with artificial editing effects, such as fading, dissolve and wipe [3].

A. Azeroual and K. Afdel are with Computer Systems and Vision Laboratory, Faculty of Science, Agadir,Morocco.

M. El Hajji and H. Douzi are with IRF-SIC Laboratory, Faculty of Science, Agadir, Morocco.

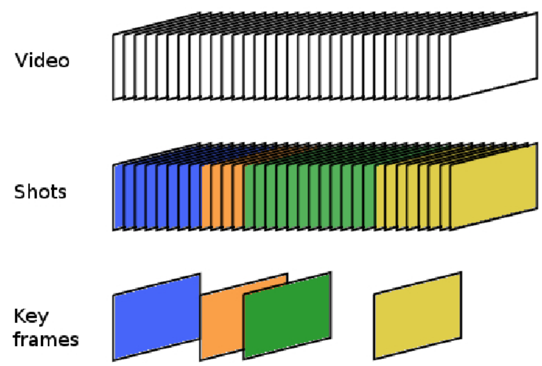


Fig. 1. The structure of a video

One frame is sufficient to present the important informations of a shot, this frame is called key frame, the figure 1 illustrate the structure of a video. Key frames hold the most important content of the video and thus are representative of the video [4] [5].

Advances in digital content distribution and digital video recorders, made the digital content recording easy. However, the coast on time and memory is expensive when we work on the full video frames especially for the real-time applications where missed real-time deadlines result in performance degradation rather than failure [6]. Furthermore in the case of video summarization the user may not have enough time to watch the entire video. Therefore, many research works have been done about the key frame extraction to perform well video processing like video summarization,video annotation, creating chapter titles in DVDs,video transmission, video indexing, and prints from video [7]. The nature of video gives a solution to those problems, as videos usually contains redundant information which can be removed to perform video processing.

Many methods have been presented in the literature for key frame extraction [5] [8] [9], but most of them are computationally expensive [10]. A set of techniques compute the difference between consecutive frames based on some criteria like color histogram, intensity histogram [11], or color features and Singular Value Decomposition [12] [13], these techniques chose a frame as a key frame if this difference is less or greater then a threshold [14], but those methods are available for the abrupt transitions and not for gradual transitions.

The approaches based on color feature attempt to calculate an histogram of video frames presented in Red-Green-Blue (RGB) color space, or Hue-Saturation-Value (HSV) color space, or other color space, after that, comparisons between frames histograms take place to chose the key frames. However this method is computationally expensive and sensitive to brightness and camera color effects. Hence, this method can give a false alarm on key frame or extract some frames presenting a non important information.

Some techniques cluster frames based on some resemblance measure, then chose one frame from each cluster as a key frame, other techniques extract the interesting objects and events in a video to find the semantically pertinent key frames.

Several visual features are used to select key frames, however the techniques used are too complex or have a bad quality of key frame extraction. To address these problems, this paper proposes a novel key frame extraction algorithm based on Faber-Schauder Discrete Wavelet Transform (FSDWT) and SVD.

The algorithm extracts the block dominant image features of each video frame and constructs a 2D feature matrix. Then the matrix is factorized using SVD. Finally key frames are extracted based on the traced rank. The advantages of the algorithm are the low computational requirements, the robustness against the gradual transitions and non-sensitivity to brightness.

II. Background And Theory

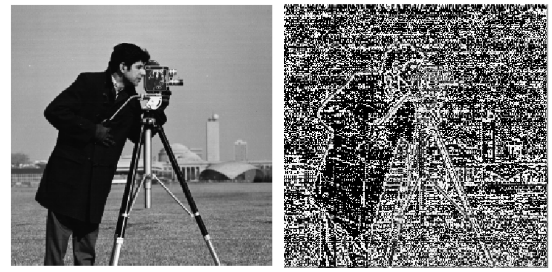
A. FSDWT

The choice of Faber-Schauder wavelet transform is motivated by the following. First it is easy to generate wavelet functions that has nice mathematical properties in image processing like the fact that they can be used as multiscale edge detectors. Second The FSWT has a simple lifting scheme formulation with only arithmetic operations and no boundary processing and it preserves the range of pixel values after transformation. Finally the FSWT is well adapted to edge detection and image characterization by extrema wavelet coefficients [15].

The FSDWT is a mixed scales representation of an integer wavelet transform [15], the figure 2 shows the mixed scales represent of the cameraman image. It based on the Lifting Scheme [16] without any boundary treatment.

We can consider an image as a sequence $f^0 = (f^0_{m,n})_{m,n \in Z}$ of Z^2, transform FSWT is done in three steps as shown in the figure 3:

- Splitting : We split the sequence f^0 into two sets of samples $f^{1,0} = (f^0_{2k+1})_{k \in Z}$ and $g^{1,0} = (f^0_{2k})_{k \in Z}$.
- Predicting : We predict the odd coefficients from a linear combination of the neighboring even coefficients $g^1 : g^1_k = g^{1,0}_k - P(f^{1,0})(k)$ for $k \in Z$ and $P(f^{1,0})(k) = \frac{1}{2} f^{1,0}_k + \frac{1}{2} f^{1,0}_{k+1}$.
- Updating : f^1 is a low-pass filter of f^0 and is obtained by updating $f^{1,0}$ with $g^1 : f^1_k = f^{1,0}_k - U(g^1)(k)$ for $k \in Z$. For FSWT there is no updating operator and f^1 is simply an interpolation of $f^0 : f^1 = f^{1,0}$.
- For finite size signal we repeat the lifting scheme on the coarser signal until we obtain only one sample f^N.



(a) Original Image (b) Mixed scales representation of the original image

Fig. 2. Mixed scales representation.

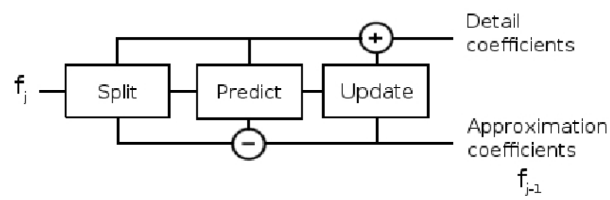


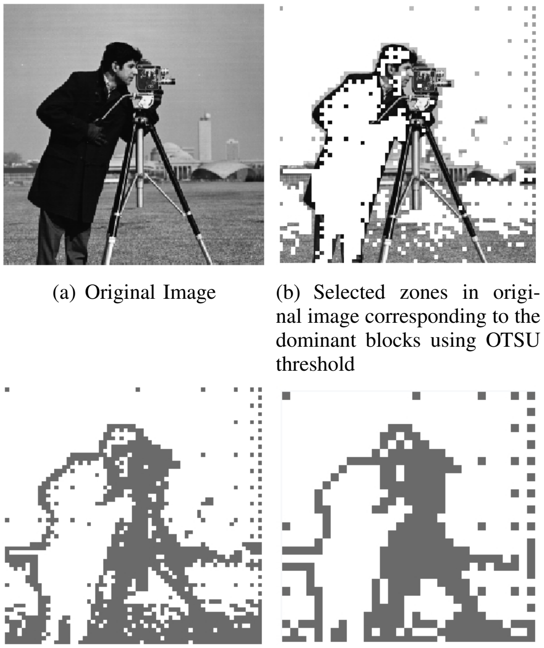
Fig. 3. Lifting Scheme

The lifting Scheme of the FSDWT [15] is given by the following algorithm:

$$\begin{cases} f^0 = f_{ij} \quad for\ i,j \in Z \\ for \quad 1 \le k \le N \\ f^0_{ij} = f^{k-1} \\ g^k_{ij} = (g^{k1}_{ij}, g^{k2}_{ij}, g^{k3}_{ij}) \\ g^{k1}_{ij} = f^{k-1}_{2i+1,2j} - \frac{1}{2}(f^{k-1}_{2i,2j} + f^{k-1}_{2i+2,2j}) \\ g^{k2}_{ij} = f^{k-1}_{2i,2j+1} - \frac{1}{2}(f^{k-1}_{2i,2j} + f^{k-1}_{2i+2,2j+2}) \\ g^{k3}_{ij} = f^{k-1}_{2i+1,2j+1} - \frac{1}{4}(f^{k-1}_{2i,2j} + f^{k-1}_{2i,2j+2} + \\ f^{k-1}_{2i+2,2j} + f^{k-1}_{2i+2,2j+2}) \end{cases} \quad (1)$$

Textured regions and contours are efficiently detected by FSDWT. It redistributes the image contained information which is mostly carried in the dominant coefficients. To facilitate the selection of theses dominant coefficients in all subbands, we use mixed-scales representation which puts each coefficient at the point where its related basis function reaches its maximum. So, a coherent image can be visually obtained with edges and textured regions formed by dominant coefficients. These regions are represented by a high density of dominant coefficients. They present more stability for any transformation keeping visual characteristics of the image [15].

In [17] [18], El Hajji and al. use standard deviation of mixed scales DWT coefficients σ_1 and local deviation σ_2 for given



(a) Original Image

(b) Selected zones in original image corresponding to the dominant blocks using OTSU threshold

(c) The dominant blocks using OTSU threshold

(d) The dominant blocks using standard deviation of mixed scales DWT coefficients

Fig. 4. Comparison between the standard deviation of mixed scales DWT coefficients method and OTSU threshold method.

8x8 block as a rule to detect a dominant block: if $\sigma_2 \geq \sigma_1$ then the block is dominant. For more precision and to fix automatically the threshold used in the algorithm we use the OTSU threshold [19] in the place of standard deviation of mixed scales DWT coefficients and 4x4 blocks. The dominant coefficient blocks are located around the image contours and textured zones near to contours, as shown in Figure 4. The original image is presented in figure 4-a, then the figure 4-d was obtained by assigning a gray color to the positions of the image's pixels corresponding to the dominant blocks using standard deviation of mixed scales DWT coefficients, we remark that this presentation is not precise. The figure 4-c is obtained by assigning a gray color to the positions of the image's pixels corresponding to the dominant blocks using OTSU threshold, this presentation is more precise than the other one in figure 4-d. Finally the figure 3-b presents the zones of image 4-a associated with the dominant blocks presented in figure 4-c.

- In the first step we compute the Faber-Schauder DWT coefficients.
- In the second step we divide the image into 4x4 blocks.
- In the third step we calculate the local deviation of each block.
- Finally we compare the local deviation to the OTSU threshold α. If $\sigma \geq \alpha$ a block is considered dominant, otherwise this block contain a big density of coefficients which are related to image contours and textured zones near to contours.

B. SVD

The decomposition into singular values is based on a linear algebra theorem which tells us that any m x n matrix A with m $\geq$ n can be factored as in (2) where U is an m x m orthogonal matrix, V^T is the transposed matrix of an n x n orthogonal matrix V, and S is an m x n matrix with singular values on the diagonal.

$$A = USV^T \tag{2}$$

The matrix S can be presented as in (3). For i = 1, 2, 3,...,n, σ_n are called Singular Values of matrix A.

$$\mathbf{A} = \begin{bmatrix} \sigma_1 & 0 & ... & 0 \\ 0 & \sigma_2 & ... & 0 \\ \vdots & \vdots & & \vdots \\ 0 & ... & 0 & \sigma_n \\ 0 & ... & 0 & 0 \end{bmatrix}, \tag{3}$$

and $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n$

There are many properties of SVD from the viewpoint of image processing applications :

- The singular values of an image have very good stability, that is, when a small perturbation is added to an image, its Singular values do not change significantly [20].
- Each singular value specifies the luminance of an image layer while the corresponding pair of singular vectors specifies the geometry of the image [20].
- Singular values represent intrinsic algebraic image properties [20].
- Singular values represent the image energy, and we can approximate an image by only the first few terms.
- The first term of singular values will have the largest impact on approximating image, followed by the second term, then the third term, etc.

III. Proposed Method

The proposed method for extraction of key frames is based on FSDWT and SVD. In [12], W. Abd-Almageed uses a sliding-window SVD approach based on Hue- Saturation-Value (HSV) color space of video frame. However, this approach is sensitive to change of frame brightness and frame color. To solve these problem we use the dominant blocks of a video frame in the place of his HSV presentation. The dominant blocks are located at the frame contours and textures around; they characterize uniquely the frame and give us a good precision when we extract the key frames.

Firstly, we convert the video to gray color, after that we compute the dominant blocks of each video frame. Then we select the dominant blocks zones in frame using the OTSU threshold α.

Secondly, an histogram H^t of length l (the number of histogram bins) is computed for the video frame at time t, next build a Nxl feature matrix X^t for every frame at time $t > N$ as shown in (4), N is a window width and can be the maximum number of frames used in a transition in the video.

$$X^t = \begin{bmatrix} H^t \\ H^{t-1} \\ \vdots \\ H^{t-N+1} \end{bmatrix}, and \qquad t = N, ..., T \tag{4}$$

X^t is a matrix feature varying in the time, presenting the feature of the current frame and previous N-1 frames and T is the total number of video frames. Thirdly, we use SVD to factorize the matrix X^t as shown in (5):

$$X^t = USV^T \tag{5}$$

Let the singular values be $S_1, S_2, ..., S_N$, with S_1 being the maximum singular value. The rank $r^t of X^t$ is the number of S_i that satisfy the condition as shown in (6):

$$\frac{S_i}{S_1} > \tau \tag{6}$$

τ is a user-defined threshold limiting the number of key frame extracted according to the precision liked.

Tracing the computed ranks over time, we can draw two scenarios. The first one, if the rank of the current feature matrix, X^t , is greater than the previous one, X^{t-1} , and then the visual content of the current video frame is different than the content of the previous frame, since the first singular values present the most informations contained in X^t , hence the increase in number of singular values that satisfy the condition (6) means that there is a change in the content of X^t , otherwise the current frame is enough different to be considered as a key frame. The second scenario, if the rank of the current feature matrix, X^t , is smaller than the rank of previous matrix, X^{t-1} , and then the visual content of the video has been stable.

Finally, we have two conclusions. First, the frame at which the rank $r^t = 1$ and $r^{t+1} > r^t$, is the ending of shot. Second, between two consecutive shots, the frame at which the rank is maximum is extracted as a key frame and presents the start of shot. The algorithm is illustrated in figure 5.

The algorithm is initialized with the first N frames that are used to compute $X^t = N$,then the main algorithm loop starts at N+1.

IV. Experimental Results

The results of the proposed key frame extraction algorithm are presented in this section. We used C++ and OpenCV library to implement the shot boundary detection and key frame extraction algorithm. A video soccer of 5253 frames was used to validate the proposed approach.

With frame size 320 x 240 and frame rate 30 fps. The algorithm produces the correct key frames. For the video in our example as shown in figure 6, the number of frames dissolve effect transition is 3 to switch from frame number 505 to frame number 509, at the frame 506 the rank = 1 and the rank of the frame number 507 is 5, so the frame number is the ending of shot, then the rank increases to 3 at frame number 509 which is the key frame. The algorithm selects a stable key frame even if it was a dissolve transition. The algorithm extract 67 key frame from 5253, some of them are shown in figure 7.

TABLE I
Experimental results

Method	key frame number	Average recall	Precision
Our method	67	98.41%	92.53%
Liu Feipeng method	36	60.34%	97.22%
W.Abd-Almageed method	81	83.87%	64.19%

The performances are evaluated based on (7) and (8). Using a window of width N = 6 and threshold 0.05, we obtained an average recall of 97.05 % and a precision of 98.50 % and 1.25 % of video frames are extracted as a key frames. The Comparative results of the key frame extraction with the methods in ([12], [11]) is illustrated in Table I.

$$Recall = \frac{Correct}{Correct + Missed} \tag{7}$$

$$Precision = \frac{Correct}{Correct + FalseAlarms} \tag{8}$$

V. Conclusion

In this paper, a video key frame extraction and boundary shot detection algorithm is proposed. In the proposed approach a Faber-Schauder dominant blocks of each video frame is computed to construct a feature matrix. Then a sliding window SVD is used to compute the rank of the current feature matrix. By tracing the computed rank we can detect the end of shot and the start of shot which can be extracted us a key frame.

Experimental results shows that our algorithm is robust against the transition effects like dissolve one used in some videos like sports ones. More experiments should be done to replace the threshold using in the phase of computing rank, by a threshold fixed automatically.

Acknowledgment

This work was supported by the Centre National pour la Recherche Scientifique et Technique (CNRST), funded by the Moroccan government.

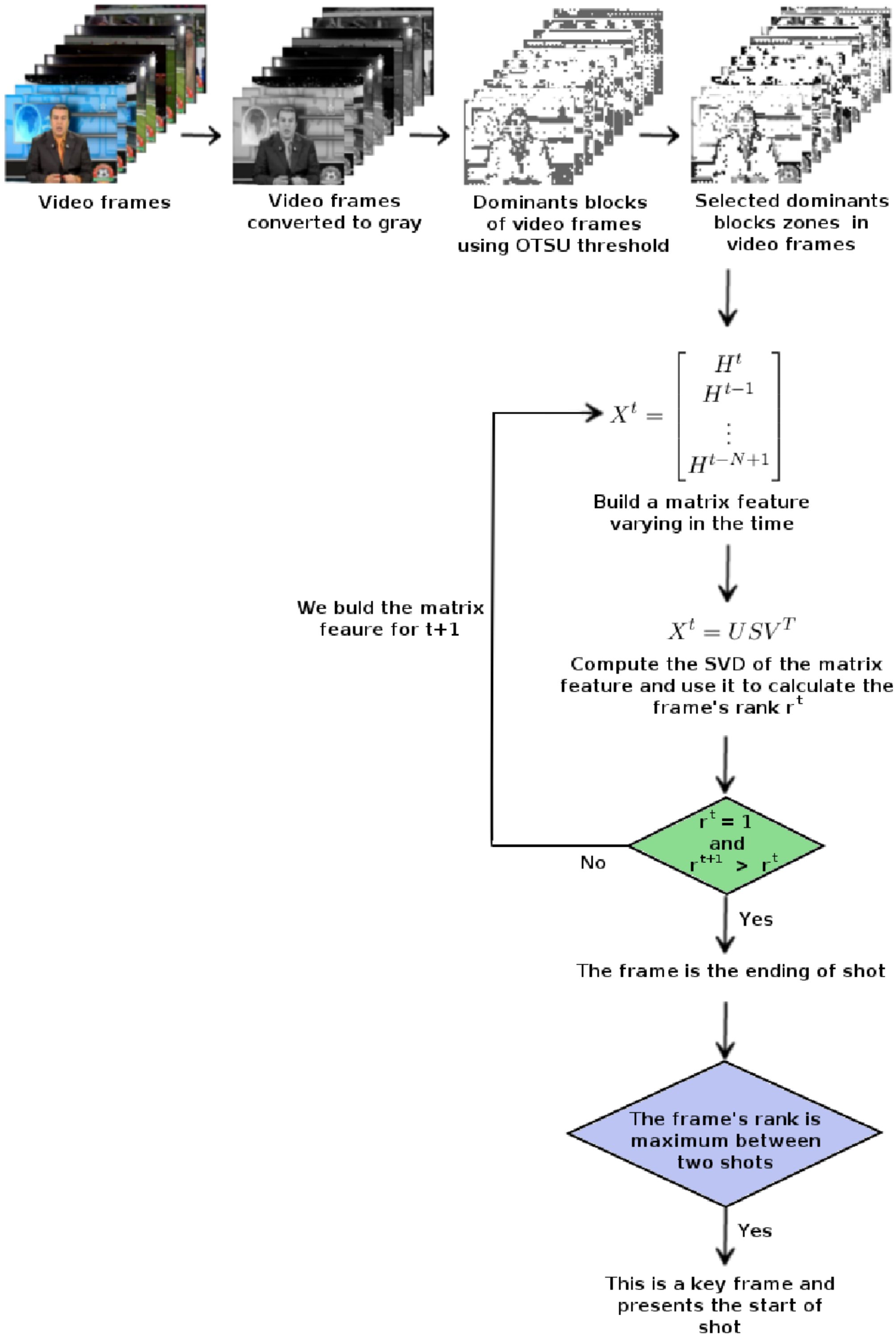


Fig. 5. The new approach algorithm

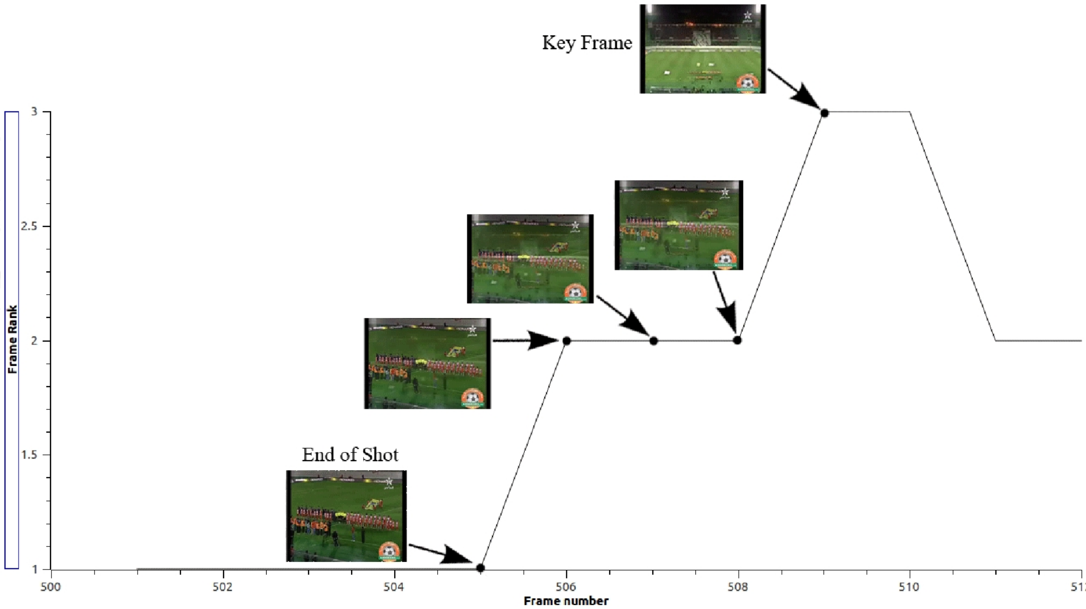


Fig. 6. Dissolve effect from frame number 505 to frame number 509.

Fig. 7. Some video key frames, we obtain 67 key frames from a video of 5253 frames

REFERENCES

[1] S.H. Yen, H.W. Chang, C.J. Wang, C.W. Wang, Robust News Video Text Detection Based on Edges and Line-Deletion, WSEAS Transactions on Signal Processing, Issue 4, Volume 6, October 2010, pp. 186-195.
[2] G. Szcs, Index picture selection for automatically divided video segments, International Journal of Computers, Volume 8, 2014, pp. 183-192.
[3] Guozhu Liu, Junming Zhao, Key Frame Extraction from MPEG Video Stream,Third International Symposium on Information Processing, 2011.
[4] G. Ciocca, R. Schettini, Innovative algorithm for key frame extraction in video summarization, Journal of Real-Time Image Processing, vol. ,pp. 6988, 2006.
[5] B.T. Truong, S. Venkatesh, Video abstraction: a systematic review and classification, ACM Transactions Multimedia Computing, Communications and Applications, vol. 3, 2007.
[6] R. Dobrescu, M. Dobrescu, D. Popescu, Parallel image and video processing on distributed computer systems, WSEAS Transactions on Signal Processing, Issue 3, Volume 6, July 2010, pp. 123-132
[7] C. T. Dang, M. Kumar, H. Radha, Key Frame Extraction from Consumer Videos Using Epitome, Image Processing (ICIP), 19th IEEE International Conference on. pp. 93-96, September 2012.
[8] A.G. Money, 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) 121143.
[9] Y. Li, S.-H. Lee, C.-H. Yeh, C.-C. Kuo, Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques, IEEE Signal Processing Magazine 23 (2) (2006) 7989.
[10] M. Furini, F. Geraci, M. Montangero, M. Pellegrini, STIMO: STIll and moving video storyboard for the web scenario, Multimedia Tools and Applications 46 (1) (2010) 4769.
[11] Video Boundary DetectionPart 1 Abrupt Transition and Its Matlab Implementation, http://www.roman10.net/video-boundary-detectionpart-1-abrupt-transitions-and-its-matlab-implementation/.
[12] W. Abd-Almageed, Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing, Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on , vol., no., pp.3200,3203, 12-15 Oct. 2008.
[13] A.V.Kumthekar, Prof. J.K. Patil, Key frame extraction using color histogram method, International Journal of Scientific Research Engi neering and Technology (IJSRET) Volume 2 Issue 4 pp 207-214, ISSN 22780882, july.2013.
[14] R.M. Jiang, A.H. Sadka, D. Crooks (Eds.), Advances in video summarization and skimming, M. Grgic, K. Delac, M. Ghanbari (Eds.), Recent Advances in Multimedia Signal Processing and Communications, 231, Springer Berlin, Heidelberg, 2009, pp. 2750.
[15] H. Douzi, D. Mammass, F. Nouboud, ”Faber-Schauder wavelet transformation application to edge detection and image characterization,” Journal of Mathematical Imaging and Vision Kluwer Academic Press, pp 91-102 ,Vol. 14(2),2001.
[16] W. Sweldens, ”The lifting scheme: A construction of second generation wavelets,” SIAM Journal on Mathematical Analysis, vol. 29, no.2, pp. 511546, 1998.
[17] M. El Hajji, H. Douzi, D. Mammas, R. Harba, F. Ros, A New Image Watermarking Algorithm Based on Mixed Scales Wavelets, J.Electron. Imaging. 21(1), 013003 (Feb 27, 2012).
[18] M. Hajji , H. Douzi , R. Harba, Watermarking Based on the Density Coefficients of Faber-Schauder Wavelets, Proceedings of the 3rd international conference on Image and Signal Processing, July 01-03, 2008, Cherbourg-Octeville, France.
[19] N. Otsu,A threshold selection method from grey scale histogram, IEEE Trans. on SMC, Vol. 1,pp. 62-66, 1979.
[20] K. Bhagyashri,Joshi M. Y. ,Robust Image Watermarking based on Singular

Assma AZEROUAL In 2012 she received the Master on Computer Systems and Networks from The University of IBN ZOHR Morocco. Since December 2012 she prepares Ph.D on Computer Systems and Vision.

Karim AFDEL In 1994 he received the Doctorat (French Ph.D) from the University of Aix Provence France in Computer Engineering, Analysis and Medical Image Processing. Since 1995 he is Professor at the University of Agadir, Morocco. His research interests are mainly on Computer Vision and Machine Learning.

Mohamed EL HAJJI In 2012 he received the Doctorat (Moroccan Ph.D) from The University of IBN ZOHR Morocco in Computer Science and Watermarking. Since 2012 he is Assistant Professor in Regional Center for Careers in Education and Training-Agadir.

Hassan DOUZI In 1992 he received the Doctorat (French Ph.D.) from The University of Paris IX (Dauphine) in wavelets application to seismic inversion problem. Since 1993 he is Professor at the University of Agadir, Morocco. His research interests are mainly on wavelet transforms applied to image and signal processing.