Video Shot Detection and Key Frame Extraction Using Faber Shauder DWT and SVD

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

Abstract—Key frame extraction methods select the most representative frames of a video, which can be used in different areas of video processing such as video retrieval, video summary, and video indexing. In this paper we present a novel approach for extracting key frames from video sequences. The frame is characterized uniquely by his contours which are represented by the dominant blocks. These dominant blocks are located on the contours and its near textures. When the video frames have a noticeable changement, its dominant blocks changed, then we can extracte a key frame. The dominant blocks of every frame is computed, and then feature vectors are extracted from the dominant blocks image of each frame and arranged in a feature matrix. Singular Value Decomposition is used to calculate sliding windows ranks of those matrices. Finally the computed ranks are traced and then we are able to extract key frames of a video. Experimental results show that the proposed approach is robust against a large range of digital effects used during shot transition.

Keywords—Key Frame Extraction,Shot detection, FSDWT, Singular Value Decomposition.

I. Introduction

THE rapid development of digital video capture and editing technology led to increase in video data, creating the need for effective techniques [5] for video applications and analysis. Advances in digital content distribution and digital video recorders, has caused digital content recording easy. However, the coast on time and memory is expensive when we work on the full video frames. Furthermore in the case of video summarization the user may not have enough time to watch the entire video. Therefore, many research works has been done about the key frame extraction to perform well video processing like video summarization, creating chapter titles in DVDs, video indexing, and prints from video [6]. Key frames, also called representative frames, are defined as the most informative frames that capture the major elements in a video in terms of content [4].

Many methods exist for key frame extraction. However, many limitations are noticed like the expensive computing and the digital video effects (abrupt transition and gradual transitions).

In this paper, we present a novel key frame extraction algorithm based on Faber Shauder Discrete Wavelet Transform (FSDWT) and Singular Value Decomposition (SVD). The algorithm extracts the block dominant image features of each video frame and constructs a 2D feature matrix. Then we factorize the matrix 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.

A. Azeroual and K. Afdel are with Computer Systems and Vision Laboratory, Faculty of Science, Agadir,Morocco(e-mail:assma.azeroual edu.uiz.ac.ma).

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

II. Background And Theory

A. FSDWT

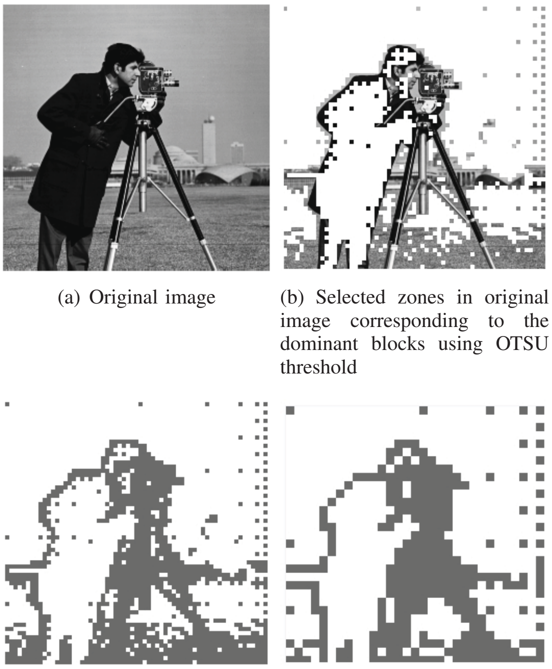
The FSDWT is a mixed scales representation of an integer wavelets transform [7]. It based on the Lifting Scheme [8] without any boundary treatment.

The Scheme lifting of the FSWT [6] is given by the following algorithm:

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

Textured and contours regions 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 sub bands, 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 [1].

In [1], El Hajji and al. use standard deviation of mixed scales DWT coefficients σ_1 and local deviation σ_2 for given 8x8 block as a rule to detect a dominant block: if $\sigma_2 \geq \sigma_1$ then the block is dominant[2]. For more precision and to fix automatically the threshold used in the algorithm we use the OTSU threshold [9] in place of standard deviation of mixed scales DWT coefficients and 4x4 block .The dominant coefficient blocks are located around the image contours and textured zones near to contours, as shown in Fig. 1. The original image is presented in Fig. 1 (a), then the Fig. 1 (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 Fig. 1 (c)



(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. 1. Comparison Between the Standard Deviation of Mixed Scales DWT Coefficients Method and OTSU Threshold Method.

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 Fig. 1 (d). Finally the Fig. 1 (b) presents the zones of image in the Fig. 1 s(a) associated with the dominant blocks presented in Fig. 1 (c).

The algorithm for detecting dominant coefficient blocks can be presented as following:

- In the first step we compute the Faber Schauder DWT coefficients.
- In the second step we divide the image in to 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.

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

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 [10].
- Each singular value specifies the luminance of an image layer while the corresponding pair of singular vectors specifies the geometry of the image [10].
- Singular values represent intrinsic algebraic image properties [10].
- 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.

C. Proposed Method

In [3], 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 this 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.

Secondly, an histogram Ht of length l 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.

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

where T is the total number of video frames. Otherwise X^t is a matrix feature varying in the time, presenting the feature of the current frame and previous N-1 frames.

Thirdly, we factorize the matrix X^t using SVD 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. 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 Fig. 2.

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.

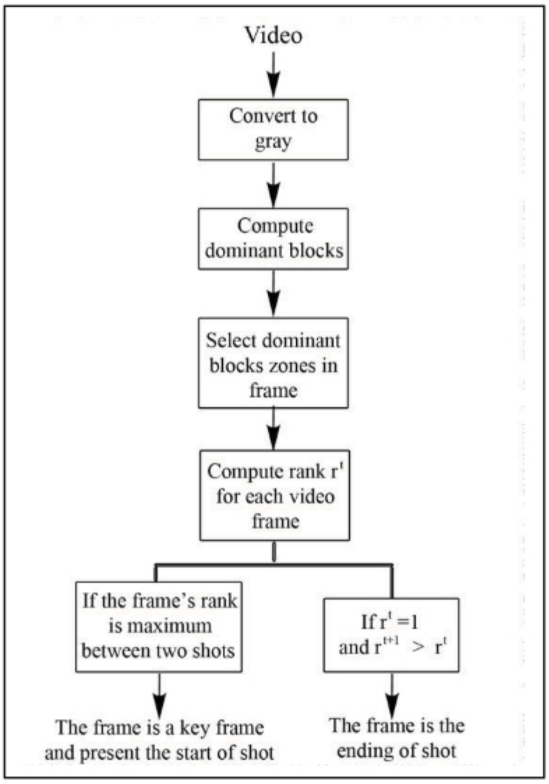

Video
Convert to gray
Compute dominant blocks
Select dominant blocks zones in frame
Compute rank r^t for each video frame
If the frame's rank is maximum between two shots
If r^t =1 and r^{t+1} > r^t
The frame is a key frame and present the start of shot
The frame is the ending of shot

Fig. 2. The new approach algorithm.

III. 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 Fig. 3, 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 2, 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 Fig. 4.

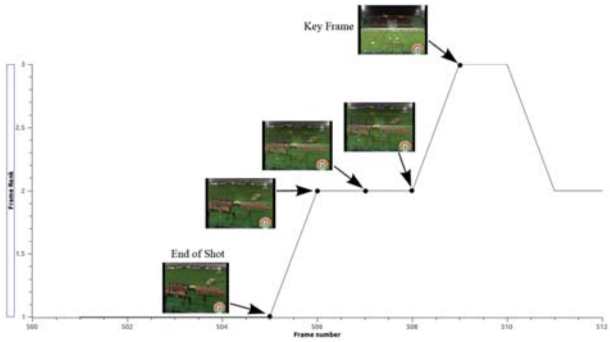


Fig. 3. Dissolve effect from frame number 505 to frame number 509.

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.

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

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

Fig. 4. Some video key frames.

IV. Conclusion

In this paper, a video key frame extraction and boundary shot detection algorithm is proposed. In the proposed approach a Faber-Shauder 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.

REFERENCES

[1] 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).

[2] 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.

[3] 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.

[4] S. Lei, G. Xie, G. Yan, A Novel Key-Frame Extraction Approach for Both Video Summary and Video Index , ScientificWorldJournal. 2014 Mar 16;2014:695168.

[5] B. T. Truong, Venkatesh, Video abstraction: A systematic review and classification, ACM Trans. Multimedia Comput. Commun. Appl. 3, 1, Article 3, Feb. 2007.

[6] 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.

[7] 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.

[8] W. Sweldens, ”The lifting scheme: A construction of second generation wavelets,” SIAM Journal on Mathematical Analysis, vol. 29, no.2, pp. 511546, 1998.

[9] N. Otsu, A threshold selection method from grey scale histogram, IEEE Trans. on SMC, Vol. 1, pp. 62-66, 1979.

[10] K. Bhagyashri, Joshi M. Y. ,Robust Image Watermarking based on Singular Value Decomposition and Discrete Wavelet Transform, Nanded 2010 IEEE.

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.