Key Frame Extraction Using MPEG-7 Motion Descriptors[1]

R. Narasimha[§], A. Savakis**, R. M. Rao* and R.De Queiroz***
[§] School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA 30332-0250
*Department of Electrical Engineering, **Department of Computer Engineering
Rochester Institute of technology, 79 Lomb Memorial Drive, Rochester, NY 14623 USA
***Xerox Corporation

ABSTRACT

We address the problem of key frame extraction in the compressed domain that is of great importance in content-based system applications. A novel MPEG-7 motion activity descriptor is discussed that is a combination of temporal and spatial descriptors. These descriptors represent both temporal motion intensity as well as spatial distribution of motion activity. It is assumed that the *apriori* information about the shot boundaries is available. The temporal descriptors are obtained by classifying the shots into five different intensity levels based on fuzzy membership functions. A high value of intensity indicates high activity and a low value of intensity indicates low activity. The spatial descriptors are obtained using motion vectors. The individual frames are characterized into spatial regions depending on the change in motion activity between successive frames. The main motivation behind this approach is to pick those frames as key frames that have maximum centralized spatial activity and high motion intensity. The motion intensity and spatial distribution are then fed to a neural network that decides the key frames based on maximum temporal activity and centralized spatial distribution. Results illustrate that the proposed approach is computationally less intensive once the network is trained and works much better than selecting the first frame and middle frame of the shot as key frame for a wide range of video sequences.

1. INTRODUCTION

The rapid increase in multimedia content, which is readily available in digitized format both, on the Internet and many other sources, has led to the problem of efficient management of these contents. The MPEG-7 standard provides a set of standardized tools to describe multimedia contents and also efficiently mange these resources on the Internet[1]. The standard specifies a set of descriptors and description schemes. One of the applications that this paper deals with is the use of motion descriptors to extract the key frames from a video sequence in the compressed domain to make possible content based access such as retrieval from multimedia databases, video browsing and summarization. Key frames provide an abridged representation of the original video sequence, serving a multitude of applications depending on the needs of the user. Key frames can provide a low bandwidth representation of the video sequence, can serve as pointers to the desired portion of the video content or can be used in video indexing application[2].

Earlier attempts for key frame extraction have been mainly using color features. A naive approach to key frame extraction was to choose the first frame of the shot as a key frame [3]. However, this method fails in case of high intensity shots. The key frame should have little overlap in the content so as to provide maximum information[2]. A more robust method to key frame selection based on color histogram was proposed by Zhang *et al*[4]. Key frame extraction based on maximizing the distance in feature space was adopted by Yeung *et al* [5]. Methods based on clustering approaches to key frame extraction have been studied by Zhuang *et al*[6] and Sun *et al*[7].

All the above approaches do not take into consideration the motion features and are computationally intensive, which can easily accessed in the compressed domain. Wolf has proposed an algorithm to extract key frame based on optical flow[8]. An improvement over Wolf's method was proposed by Divakaran *et al*[2], which makes use of the MPEG-7 intensity of motion activity descriptor as a measure of summarizability of the video sequence. A descriptor for spatial distribution of motion activity has been studied by Divakaran and Sun[9].

The main motivation of the proposed approach is that key frame should have maximum motion since high activity frames can depict maximum content and least overlap. The other criteria for key frame selection being that the spatial activity is localized at the center of the frame rather at the corners. Thus in the proposed algorithm the key frame selection is based on both the motion intensity descriptor as well as the spatial activity descriptor. The decision is done

[1] This work was supported in part by a grant from the Xerox Foundation.

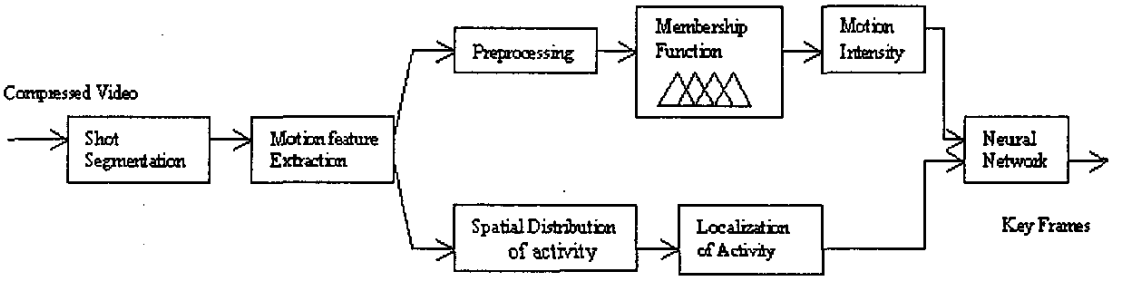

Figure 1. Block diagram of key frame extraction using MPEG-7 descriptors

using a neural network model that takes both descriptors into consideration.

Sections are organized as follows. Section 2 gives a brief view of the MPEG-7 motion descriptors and explains the proposed approach of extraction of temporal and spatial descriptors. Results and comparison with earlier approaches are provided in section 3. Conclusion and future work are drawn in section 4.

2. Extraction Of Temporal And Spatial Descriptors

A. *MPEG-7 Motion Descriptors*

The proposed algorithm uses two of the MPEG-7 motion descriptors, namely intensity of motion activity and spatial distribution of activity. The magnitude of the motion vectors represents a measure of intensity of motion activity that includes several additional attributes that contribute towards the efficient use of these motion descriptors in a number of applications. Intensity of Activity[10] is expressed by an integer in the range (1-5) and higher the value of intensity, higher the motion activity. Spatial distribution of activity [9] indicates whether the activity is spread across several regions or confined to one large region and therefore depicts the number of active regions in a frame.

B. *Extraction of Motion Descriptors*

The block diagram of the proposed key frame extraction algorithm is shown in Figure 1. The shot segmentation is performed using the twin-comparison algorithm proposed in [11], which detects gradual transitions such as dissolve, wipe, fade-in and fade-out. Therefore, this algorithm minimizes the subsequent errors that may have caused due to imperfect shot segmentation. The motion features are extracted using the motion vectors. Block motion techniques are employed to extract the motion vectors.

Suppose $x(i,j)$ and $y(i,j)$ denote the motion vectors in x and y directions for a given frame, where (i,j) indicates the block indices. We use the method defined in [12] to determine the spatial activity matrix, $Z(i,j)$,

$$Z(i,j)=\begin{cases} R_{xy}(i,j) & if\ R_{xy}(i,j) \geq avg(R_{xy}(i,j)) \\ 0 & otherwise \end{cases} \quad (1)$$

where $R_{xy}(i,j)=\sqrt{\left(x(i,j)^2+y(i,j)^2\right)}$ and is defined as the activity matrix. The average of activity matrix for each frame is given by $avg(R_{xy}(i,j))=\frac{1}{MN}\sum_{i=0}^{M}\sum_{j=0}^{N}R_{xy}(i,j)$. Here M and N denote the size of each frame. This method suppresses the low activity blocks and leaves the high activity blocks unaltered. The setup is shown in Figure 1. The intensity of motion for each frame is determined as follows

$$I_n=\frac{1}{MN}\sum_{i=0}^{M}\sum_{j=0}^{N}Z(i,j) \quad (2)$$

where n is the frame index. The preprocessing block normalizes the intensity values to the range (0-1). Then the intensity values (I_n) are fed into fuzzy membership function. A gaussian bell membership function was used as shown in Figure 2. Intensity values are classified into five categories namely very low, low, medium, high and very high activity. The output of the single input/output fuzzy system is used to classify the intensities into five different categories depending on the fuzzy rules that capture the whole intensity range[12] as depicted in Figure 2. The surface plot of the fuzzy membership function is shown in Figure 3, where x-axis represents the input and y-axis represents the output. High value of intensity indicates high activity whereas low intensity indicates low activity. Since the membership functions overlap, the activity levels are better classified into their respective intensity levels and is a robust approach than most of the hard thresholding approaches.

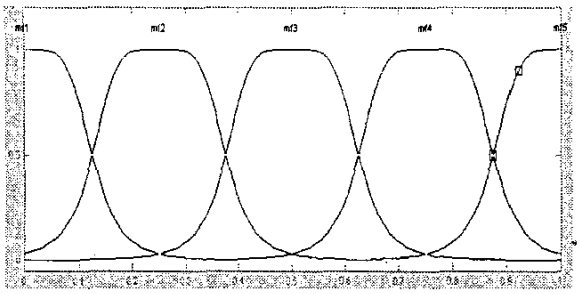

Figure 2. Fuzzy membership functions

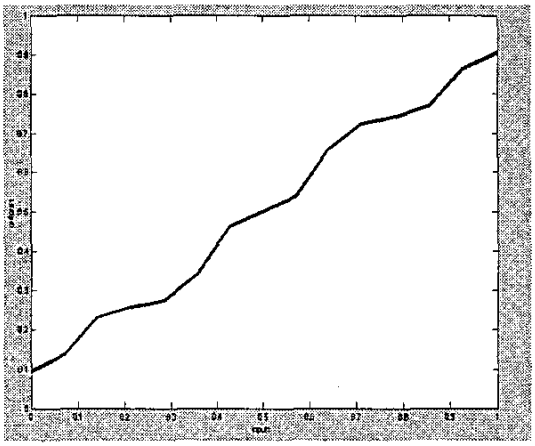

Figure 3. Surface plot of the fuzzy membership function.

C. *Spatial distribution of activity*

Generally, the activity in a frame occurs mostly at the center and therefore those frames are better candidates for key frames than those whose activity is localized at the corners. The proposed approach is shown in Figure 4, where i denotes the frame index and k denotes the number of frames in a shot. To determine the spatial distribution of activity, the spatial activity matrix is divided into nine non-overlapping regions. This procedure can be visualized as a sliding windowing approach that slides from top-left portion of the spatial activity matrix to bottom-right scanning nine spatial regions as depicted in Figure 5. The partitioning is limited to nine regions due to the computational constraints of the neural network. The spatial activity matrix values in each region are summed up to give an average spatial motion distribution in each region. Then the method localizes the spatial distribution activity in each frame to that region that depicts the maximum activity. The outcome of this approach is a matrix of nine-motion activity values that are given as an input to the neural network as shown in the block diagram in Figure 1. The idea can be visualized in Figure 6 where the subject is always centralized and the frames depict more description of activity at the center.

D. *Neural Network Model*

A backpropagation neural network with 10 input nodes, 30 hidden nodes and one output node was used to select the key frame. The input to the network is a vector that consists of the spatial distribution (9 elements) and the motion intensity information (one element) per frame. The transfer functions used in the neural network were logarithm sigmoid and tangent sigmoid functions. The network is trained in such a way that in the training phase, only those frames are picked as key frames that have maximum centralized spatial distribution and highest motion intensity. Video sequences selected from different categories that included movies, news sequences and sports sequences were chosen as training and test sequences and were presented to the network in a random fashion. The training phase had about 15 video sequences of varied contents and the network was trained to pick a key frame per shot boundary based on the imposed criteria.

3. RESULTS

The system performance was evaluated using test set video sequences that included movies such as terminator, sports, news sequences and other video sequences. Since the news reading and other talking head sequences did not depict much change in the motion intensity, although the spatial distribution was at the center of the frame, any frame in the sequence can be chosen as the key frame. The approach does not take into consideration camera pan, but since the key frames are chosen based on the maximum center spatial activity and maximum motion intensity, the approach performs satisfactorily even in the case of camera pan. The performance of the proposed method is compared with the existing methodology of picking the middle frame (MF) of a given shot as a key frame and the comparison with selecting the first frame as key frame is not provided due to its limitations is the case of high intensity shots as mentioned earlier. The results are summarized in Table 1 where the proposed method picked key frames that were visually more descriptive based on the criteria specified in about 40% of the total number of shots. In 56.76% of the cases there were no difference between the two approaches. The key frame comparison for the cases where the proposed model performed better is shown in Figure 6. The interesting conclusion in all the selected key frames is that there is always maximum description in terms of intensity and spatial distribution at the center of the frame and performs better in the case of high activity sequences such as sports and action movies.

4. CONCLUSION AND FUTURE WORK

The general assumption of the above method is that the frame that has maximum motion intensity will have minimum overlap in content, which is valid in the case of key

frame extraction for video summarization. The algorithm demonstrates that the frames with maximum centralized spatial distribution are better candidates for key frames. The algorithm performs better than earlier key frame selection approaches such as selecting first frame and middle frame of the shot as key frames in the case of high activity sequences. The use of a neural network results in computationally intensive training, but once the weights have been adjusted the computational burden is reduced significantly.

5. REFERENCES

[1] http://www.cselt.it/mpeg/workingdocuments.htm, MPEG-7 Visual Committee Draft.

[2] A. Divakaran, R. Regunathan, and K. A. Peker, "Video Summarization Using Descriptors of Motion Activity: A Motion Activity Based Approach to Key-Frame Extraction from Video Shots," *Journal of Electronic Imaging*, vol. 10, pp. 909-916, October 2001.

[3] B. Shahraray and D. C. Gibbon, "Automatic Generation of Pictorial Transcripts of Video Programs," *Multimedia Computing and Networking 1995*, vol. Proc. SPIE 2417, Feb 1995.

[4] H. Zhang, J. Y. A. Wang, and Y. Altunbasak, "An integrated system for content-based video retrieval and browsing," *Pattern Recognition*, vol. 30, pp. 643-648, 1997.

[5] M. M. Yeung and B. Liu, "Efficient matching and clustering of video shots," *Proceedings of IEEE ICIP*, pp. 338-341, 1995.

[6] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive Key Frame Extraction Using Unsupervised Clustering," *Proc. of IEEE Int Conf on Image Processing (ICIP)*, pp. 866-870, 1998.

[7] X. Sun, M. S. Kankanhalli, Y. Zhu, and J. Wu, "Content-Based Representative Frame Extraction for Digital Video," *ICMCS*, pp. 190-193, 1998.

[8] W. Wolf, "Key Frame selection by motion analysis," *ICASSP 96*, pp. 1228-1231, 1996.

[9] A. Divakaran and H. Sun, "Descriptor for spatial distribution of motion activity for compressed video," *Proceedings of SPIE on Storage and Retrieval for Media Databases 2000*, vol. 3972, pp. 24-28, Jan 2000.

[10] S. Jeannin and A. Divakaran, "MPEG-7 visual motion descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, pp. 720-724, 2001.

[11] H. J. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partioning of full-motion video," *Multimedia Systems: Springer-Verlag*, vol. 1, pp. 10-28, 1993.

[12] D. A. X. Sun and B. S. Manjunath, "A Motion Activity Descriptor and Its Extraction in Compressed Domain," *IEEE Pacific-Rim Conference on Multimedia (PCM)*, vol. LNCS 2195, pp. 450-453, October 2001.

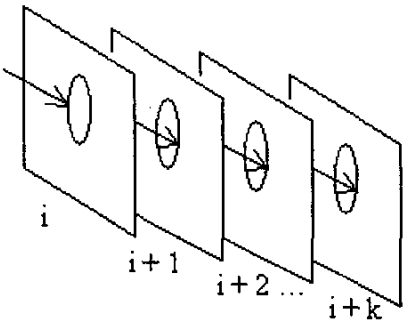

Figure 4. Pictorial representation of the proposed approach to localize the spatial distribution of activity

LT	CT	RT
LC	CC	RC
LB	CB	RB

Figure 5. Partioning of the frame into nine regions, where T-Top, L-Left, C-Center, R-Right and B-bottom.

	Percentage
MPEG 7 descriptor method	40.54
Middle Frame (MF) method	2.7
No difference	56.76

Table 1. Comparison of percentage scores of the proposed model and the MF method for different test sequences.

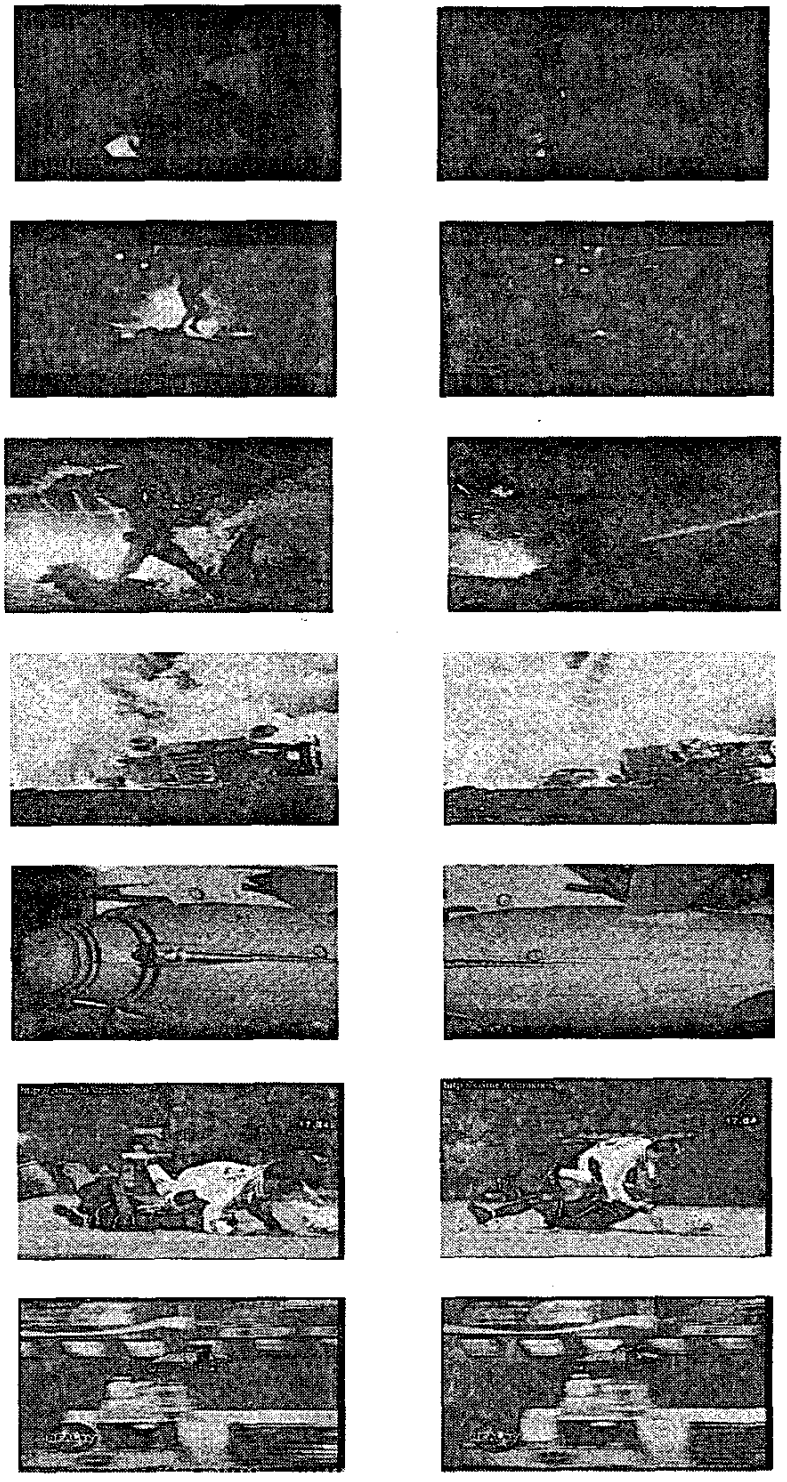

Figure 6. Comparison of the proposed key frame extraction system (left) versus middle frame extraction method (right).