International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.7, No.4 (2014), pp.381-388
http://dx.doi.org/10.14257/ijsip.2014.7.4.36

Contourlet Transform Based Shot Boundary Detection

Pamarthy Chenna Rao[1] and M. Ramesh Patnaik[2]

[1]*Research scholar, Dept. of Instrument Technology*
Andhra University
[2]*Assistant Professor in Dept. of Instrument Technology,*
Andhra University, Vizag, Andhra Pradesh, India

Abstract

Shot boundary detection is used to summarize the video in the video content management system so that particular content or required scene can be extracted from a given video and also observers or users can easily find summary of the given video in short time instead of time spending in seeing full length video. Also shot boundaries can be used for video indexing, browsing, and retrieval. So to extract such key frames (shot boundaries) from a given video needs a robust technique with high accuracy and low error rate. In this research work, we proposed a robust technique to find shot boundary using feature vectors, which are obtained from features calculated for each sub-band in the contour let transform domain. The feature vector has the energy, standard deviation, and histogram similarity as the features of each sub-band. The experimental results proved that this novel and robust method had more accuracy rate and low error rate.

Keywords: *Contourlet, Key frame, Histogram, Energy, Standard Deviation, Euclidian distance*

1. Introduction

Increase in video on the Internet demands fast and efficient techniques for managing and indexing of such contents and also users or observers want to save their valuable time in knowing summary of the full video or want to extract required event without watching and browsing full videos and even though they are not interested in the entire movie. This demands a technique which provides video abstraction and summarization that produces video summaries highlighting only the relevant contents of the video while preserving the continuity of the video. This summarized video helps the user to evaluate whether it is interesting or not [1] and also gives quick summary about that video. This kind of summarization finds its applications in security like CCTV footage browsing, entertainment and military areas [2]. Video summarization techniques are classified into three broad categories: (i) internal: analyze information sourced directly from the video stream, (ii) external: analyze information not sourced directly from the video stream and (iii) hybrid: analyze a combination of internal and external information [3]. Shot boundary detection and key frame extraction words are interchangeable used in this paper in order to achieve video summary.

A shot is defined as an unbroken sequence of frames recorded from a single camera [4]. This kind of shot forms the building block of a video. The main objective of key frame detection is to segment the video stream into multiple shots and then key frames can be extracted. This key frame is the main frame of the given shot, which intern gives information about that shot. Generally Key frame refers the starting frame of the shot also termed as intra

ISSN: 2005-4254 IJSIP

frame, which can represent the salient content of the shot. Depending on the content complexity of the shot one or more key frames can be extracted from a single shot

Since key frames give an overview about the full video for video indexing, browsing, and retrieval. This drastically reduces amount of data required for video indexing and forms the framework to dealing with video content [4]. To achieve this task different author used different methods. Uchihashi [9] uses YUV color histogram to find the key frames. Lagendijk and Iacob [10] divides the each frame into rectangles with size depends on local structure and YUV color histogram will be applied on each rectangle in order to extract key frame. Ratakonda [11] uses color histogram differences between two frames and then generates a summary of the video by clustering these key frames. Ferman and Tekalp [12] select key frames based on fuzzy-clustering algorithm.

Due to the importance of key frame extraction, many research efforts has been given in [4], and the progress in research has been made in this area, but existing approaches have their own advantages and disadvantages and some of the algorithms are computationally expensive and few may not effectively capture the major visual content and some captures the key frames with less accuracy rate and high error rates. Where accuracy rate is defined as the ratio of extracted number of correct key frames to actual number of key frames. Error rate is the ratio of number of extracted non matching key frames to the actual number of key frames.

The proposed method uses Contourlet transform [5] and energy, standard deviation and histogram similarity features are extracted from each sub band and feature vector is formed by using these feature values and Euclidean distance metric is used on feature vectors of two adjacent frames in order to extract key frame based on threshold value. The proposed work also proved that Contourlet transform is good enough to extract the key frames and also proved that the method is having high accuracy rate and low error rates with selected feature vector and distance measure technique on all kinds of videos.

The work in this paper has been divided into few sections, Section II gives the overview of the proposed work. Section III describes advantages and importance of Contourlet transform. Section IV explains how to extract features and forming of feature vector and by using these vectors with Euclidean distance for key frame extraction. Section V shows the experimental results and then conclusion in Section VI.

2. Overview of the Proposed Work

The block diagram of proposed work is shown in Figure 1. Key frame extraction process involves, framing the input video and apply the Contourlet transform on each input frame with the decomposition parameter [4 3 3], this results in 32 high frequency sub bands and one low frequency sub band. As a next step, energy and standard deviation features are computed for each sub band and form a feature vector by using the energy and standard deviation features, where feature vector length is 66. Now these feature vectors are used to decide whether the given frame is key frame or not. To compare these feature vectors of two consecutive frames Euclidean distance is used, which computes the distance between two feature vectors. Finally, the resulted distance is compared with the predetermined threshold to classify whether the frame is key frame or not. This threshold is fixed after testing on few training videos. If the distance is above threshold then the frame is classified as key frame, otherwise the frame belongs to same shot. Sometimes even in same shot content may be rotated in single shot, and then this may results in redundant key frames. The above process is continued till the video completes and resultant key frames give s the video summarization.

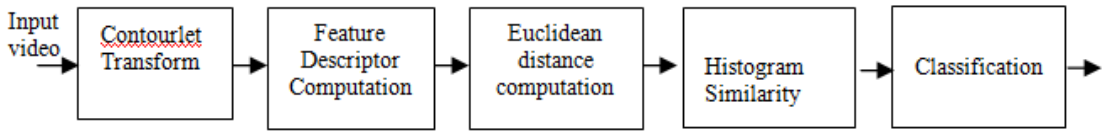

Figure 1. Block Diagram of Key Frame Extraction Process

3. Contourlet Transform

The Contourlet construction was chosen to obtain a sparse expansion of the images to produce a piecewise smooth contours. This Contourlet has the advantage that wavelet couldn't achieve because wavelets are lack of directionality where as Contourlet can find directionality. Wavelets are only good at catching point discontinuities but don't capture geometrical smoothness of the contours [6]. Due to this inefficiency in wavelet transform, Contourlets were developed. The Contourlet transform offers a high degree of directionality and anisotropy in addition to wavelet properties like multi scale and time frequency localization property. Contourlet transform has its basis functions oriented at power of two's number of directions. The basic block diagram is shown below.

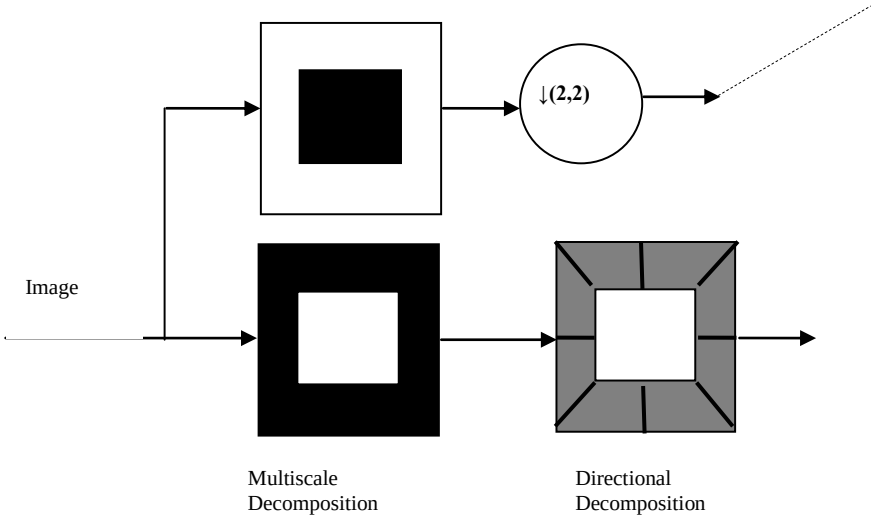

Figure 2. Contourlet Filter Bank

The Contourlet transform is implemented using directional filter banks which decompose the image into several directional sub-bands at multiple scales. Contourlet achieves this by combining directional filter banks with Laplacian pyramid at each scale. This cascade structure, multi scale and directional decomposition are independent of each other. The Figure 2 shows an example frequency partition of the Contourlet transform. For further details about Contourlet transform, please read [6, 7].

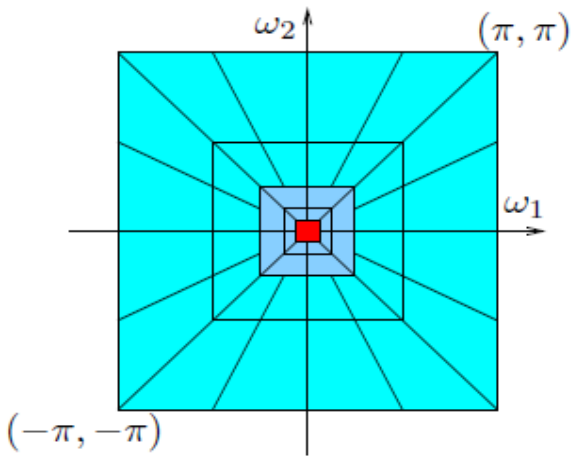

Figure 3. Example Frequency Partition Using Contourlet Transform

4. Feature Vector Computation and Classification

This section gives some overview about computation of feature vectors and how they used to extraction key frame extraction and histogram similarity removes misclassified key frames.

4.1. Feature Vector Computation

Feature is an attribute which describes an image in other form. Combining of these features results in a feature vector. Different methods use different features, where as in our algorithm we used the energy value and stand deviation of each Contourlet domain directional sub-band as a features.

In order to calculate average Energy from each sub-band, formula one will be used

$$E(s,k)=\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left|W_{S,K}(m,n)\right| \quad (1)$$

Where E(s,k) denotes the average energy of the sub-band with scale s and direction k as indexes, and M,N represents row and column number of sub-band coefficients[8].

This paper uses energy as one feature and standard deviation as another feature in our feature vectors. Standard deviation can be computed on each sub band.

These two features not robust to illumination change but this can be mitigated by normalization methods. This system increases its robustness against rotation by using histogram similarity using color histogram intersection in formula (3) as post processing step in order to eliminate non key frames.

$$d(h,g)=\frac{\sum_{A}\sum_{B}\sum_{C}\min(h(a,b,c),g(a,b,c))}{\min(|h|,|g|)} \quad (2)$$

4.2. Classification by Euclidean Distance

Two feature vectors similarity is measured by using different techniques out of those some simple techniques are by measuring distance between two feature vectors. This paper choses

Euclidean distance as a metric to measure distance between two feature vectors. This distance has performance improvement over Manhattan (L1), Weighted-Mean–Standard deviation (WMV), Euclidean (L2), Chebychev (L), Mahalanobis, Canberra, Bray-Curtis, Squared Chord, Squared Chi-Squared and Kull-back Leibler. Kokare compared the nine measures except Kull-back distance (KLD). This computed Euclidean distance between two consecutive frames will be compared against threshold in order to classify whether the frame is key frame or not. If the distance is greater than threshold then the frame will be classified as key frame else it not key frame i.e. part of the same shot.

5. Results

This results section shows strength of the proposed work with the examples in terms of accuracy and error rates. Key frames are decided using foresaid process and compared with our previous work. The following figures give the comparison between two methods and its merits than previous work.

Figure 4. Input Sports Video Frames

Figure 5. Key Frame Extraction Using Our Old Method

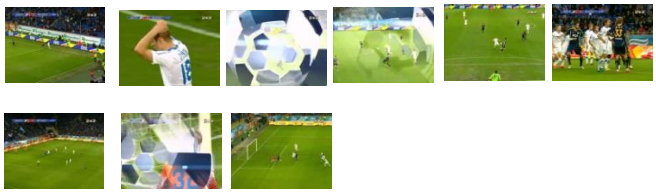

Figure 6. Key Frame Extraction Using Our New Proposed Method

In our experiment we used Contourlet transform toolbox and chosen the Laplacian pyramid and DFB filter are “pkva” type. In this Contourlet, the decomposition parameter are [4 3 3], this means that the numbers of directional sub-bands are 16, 8, 8. Adding the low

frequency sub-band, so the total number of sub-bands is 33. The paper chose two features for each sub-band so the dimension of feature vector is 66 for single image.

This work also tested on different videos i.e. sports video (foot ball video), cartoon video and ordinary video, Even this is tested on news videos. Figure 7 shows Euclidean distances between feature vectors which were calculated on a video consist of 36 frames.

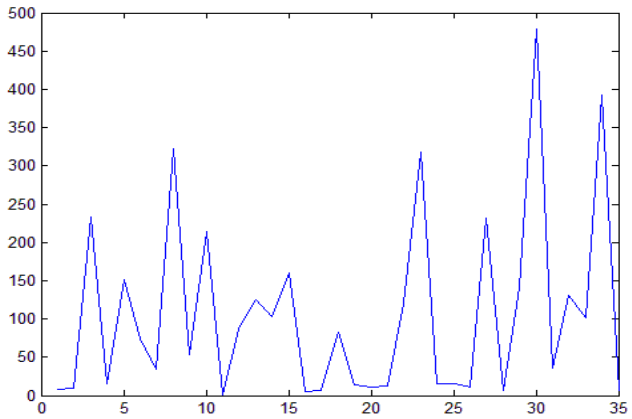

Figure 7. Sample Distance Values

And also Figure 7 shows the minimum values that means those two frames are belongs to same shot and at the same time maximum values means those two frames surely belongs to different shots so that second frame will be decided as key frame, to improve the accuracy we select threshold value and compared every distance with the threshold in order to classify a frame as key frame or not.

The previous work in this paper classified 16 frames as key frame from 36 total frames of football clip. But actual number of key frames by human observation is 9. Observation is made on the content of the frame. Out of 16 frames some frames of slightly rotated frames of previous frame, so these redundant frames are eliminated by using our new techniques such as histogram similarity as a post processing method.

The work also carried out on different videos and measured accuracy and error rates by using old and new techniques. Accuracy rate is a ratio of retrieved key frame to actual number of key frames and error rate is a ratio of number of non key frames to the actual number of key frames. Accuracy rate varies from 0 to 1 but if its value is one then that method said to be good. Error rate varies from 0 to maximum (infinity), ideally it should be zero for good algorithms.

The below table shows the error and accuracy rates on different videos and using our two techniques.

Table 1. Accuracy Rates and Error Rates using Proposed Method on Different Videos

Input video	Total frames	Key frames	Extracted Key frames		Extracted Non key frames		Accuracy rate		Err rate method	
			Old	New	Old	New	Old	New	Old	New

Cartoon video	38	18	18	18	0	0	1	1	0	0
Football	36	9	16	9	5	3	0.6	1	0.4	0.25
Movie with rotated content	136	29	27	29	2	0	0.93	1	.06	0
News video	150	8	8	8	0	0	1	1	0	0
Space video	98	11	10	11	3	1	0.9	1	0.27	.09

6. Conclusion

This paper proposed a robust key frame extraction method using Contourlet transform and energy and standard deviation as features in feature vectors and histogram similarity as post processing step in order to eliminate mis-classified key frames. The experimental results proved that this techniques show s good accuracy rate and low error rate in key frame extraction process.

References

[1] D. M. Russell, "A design pattern based video summarization technique: moving from low – level signals to high – level structure", in IEEE Proc. of the 33rd Hawaii International Conference on system, **(2000)**.
[2] Z. Li, G. M. Schuster, A. K. Katsaggelos and B. Gandhi, "Rate – distortion optimal video summary generation", IEEE Trans. On Image Processing, vol. 14, no. 10, **(2005)** October.
[3] Y. Li, S.-H. Lee, C.-H. Yeh and C.-C. J. Kuo, "Techniques for Movie Content Analysis and Skimming", IEEE Signal Processing Magazine, **(2006)** March.
[4] Z. Y. Rui, T. S. Huang and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering", In: ICIP'98, IEEE Computer Society, **(1998).**
[5] M. N. Do and M. Vetterli, "The Contourlet Transform: An Efficient Directional Multiresolution Image Representation", IEEE Trans. on Image Processing, vol. 14, no. 12, **(2005)** December, pp. 2091-2096.
[6] D. D. Po and M. N. Do, "Directional multi scale modeling of images using the Contourlet transform", Statistical Signal Processing, 2003 IEEE Workshop on, **(2003)** September 28-October 1, pp. 262-265.
[7] M. N. Do and M. Vetterli, "The Contourlet transform: An efficient directional multiresolution image representation", IEEE Trans. Image Process, vol. 14, no. 12, **(2005)** December, pp. 2091-2106.
[8] X. Chen, G. Yu and J. Gong, "Contourlet-1.3 texture image retrievalsystem", IEEE International Conf on wavelet analysis and pattern recognition, **(2010)**, pp. 49–54.
[9] S. Uchihashi, J. Foote, A. Girgenson and J. Boreczky, "Video manga: Generating semantically meaningful video summaries", In Proceedings of ACM Multimedia 99, **(1999)** October 30-November 5, pp. 383–392, Orlando, FL.
[10] S. M. Iacob, R. L. Lagendijk and M. E. Iacob, "Video abstraction based on asymmetric similarity values", In Proceedings of SPIE Conference on Multimedia Storage and Archiving Systems IV, vol. 3846, **(1999)** September, pp. 181–191, Boston, MA.
[11] K. Ratakonda, I. M. Sezan and R. J. Crinon, "Hierarchical video summarization", In Proceedings of SPIE Conference Visual Communications and Image Processing, vol. 3653, **(1999)** January, pp. 1531–1541, San Jose, CA**.**
[12] A. M. Ferman and A. M. Tekalp, "Two-stage hierarchical video summary extraction to match low-level user browsing preferences", IEEE Transactions on Multimedia, vol. 5, no. 2, **(2003)**, pp. 244–256.