ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 1, January 2013

A Comparative based study of Different Video-Shot Boundary Detection algorithms

Mr. Ravi Mishra, Dr. S. K. Singhai, Dr. Monisha Sharma

Abstract—A shot in a digital video sequence may be defined as a set of images (frames) from a single camera. A shot boundary is determined when one shot changes to another shot. A scene is a collection of one or more shots focusing on one or more objects of interest. Video shot boundary detection is a fundamental step for content based video analysis. Various methods have been proposed and developed to detect boundaries in video sequence which uses information from feature extraction, color histogram, boundary histogram, pixel differences, motion features. Some novel shot boundary methods have developed which use features above and can detect cut and gradual shot simultaneously and distinguish them. More recent algorithms designed explicitly to detect specific complex editing operations such as fades and dissolves are taken into account, and their ability to classify the types and locate the boundaries of such edits are examined. The algorithms' performance is measured in terms of hit rate, number of false hits, and miss rate for hard cuts, fades, and dissolves over a large and diverse set of video sequences
This paper gives introduction and a comparative study of all those algorithms.

Keywords: motion vectors, color histogram, segmentation, tracking, pixel difference, etc.

I. INTRODUCTION

Video shot boundary detection is the initial and fundamental step of indexing, browsing and retrieval applications. Shot boundary detection is the most basic video segmentation task. It also referred as video temporal segmentation. General objective is to segment a given video sequence into is constituent shots and to identify the different transitions between adjacent shots in the sequence. The first step of any video processing method is to segment the input video into elementary shots. A shot is defined as the continuous frame from start to the end of recording in the camera. It shows continuous image in sequence. There are two types of transitions that can occur between shots, they are abrupt change and gradual transition. Shot detection identify and classify this different shot transitions in the sequence.

Shot boundary detection is the basic and difficult problem in content-based video retrieval, because it determines the results of retrieval directly. The problem has been studied deeply and numerous technologies have been proposed, for example, the template matching method, color histogram matching method, edge matching method and model based method[1,2,3,4] etc. In many of these methods, the feature is selected to reflect the dissimilarity between two successive frames, and then all the dissimilarity among the successive frames is computed and compared to the given threshold, if the dissimilarity is high, it is considered that the content of the frame change, and the corresponding frame is the boundary of the shot. Otherwise it is considered that the frames belong to the same shot, and there is not boundary during these frames.

Although advances have been made in the recent studies, the following problems haven't be solved and they have a relatively high false detection rate: (1) It is always difficult when detecting gradual shot boundary for the little dissimilarity between two successive frames; (2) Missed detection or false detection is always brought when using a fixed threshold, too big threshold will result in missing some shot boundary, and too small threshold will regard erroneously flash lamp, the motion of camera and object as shot boundary[5].

Various algorithms have been proposed based on information theory, graph theory, object segmentation and tracking, robust three-dimensional Tracking, local feature post refinement, motion activity descriptor. These algorithms are based on information obtained from video like features, color histogram, pixel differences, motion features, etc. Section II gives introduction and general method of all these algorithms, section III presents a comparative study of these algorithms then conclusion and references.

Manuscript received Jan, 2013.

Mr.RaviMishra,Electrical&Electronics ngineering,Dr.CVRAMAN Universit,*Bilaspure,India,9893856609*
Dr.S.K.Singhai,Electrical Engg.CSVTU/ Govt. Engg. CollegeyBilaspur,India,7752260089
Dr.Monisha Sharmae, Electronics & Telecommunicatione, CSVTU/ SSCET/ , Bhilai Nagar, India, 9425564794.

II. VARIOUS VIDEO SHOT BOUNDARY DETECTION ALGORITHMS

1.Based on Graph Theory

In this algorithm, first the feature of color is extracted then the dissimilarity of video frames is defined. Then the video frames are divided into several different groups through performing graph-theoretical algorithm. This algorithm can detect cut and gradual shot and distinguish them. According to cut and gradual changes have the different characters on the two successive frames belong to different groups, it detect cut and gradual shot.

This method uses HSV color model to extract feature. Divide hue space H into 8 parts, while divide saturation space S and brightness space V into 3 parts. Constructing a feature vector with one dimension, composing HSV into such a vector that: l = 9H +3S +V. The discontinuous frame difference between frames i and j can be gained.

$$D(H_i, H_j) = (\sum_{k=0}^{71} | H_i(k) - H_j(k) |^2)^{\frac{1}{2}}$$

His normalized histogram of frame i and j and D is the frame difference between i and j.

In this algorithm, the videos in the frame are first divided into groups using graph tree method. This process first computes video frame differences by frame histograms where video frames can be regard as vertices of the graph and the frame difference can be regard as the weight of edge accordingly. Then cut the edges whose weights exceed threshold, if the edges whose weights don't exceed threshold but the difference of the two vertices number of the edge exceed 30, then cut the edge, and compute the minimum distance between the two vertices of the edge, if the minimum distance is less than threshold, connect the two vertices which the distance of the two vertices is minimum. Then according to abrupt and gradual changes, flash lamp, motion of camera and object whose different characters at successive frame differences of the different groups, detect abrupt and gradual changes exactly.

2. Based on Information theory

In this algorithm, first we extract the features of color and texture by wavelet transform then we define the dissimilarity based on the mutual information of color feature and the co-occurrence mutual information of texture feature. Wavelet transform provides a method to analyze image in different scales, and the high frequency information describe mainly the texture feature, the low frequency information describe mainly the color feature. In this algorithm, the wavelet coefficient is computed by discrete wavelet transform and the Daubechies-4 is adopted for its advantage of orthogonality, time-frequency compactly supported and low complexity.

To measure the dissimilarity, the information entropy and mutual information is used. In frames, pixel with different gray level is random, and its appearance probability is independent. The values of entropy and mutual information depend on not the gray level itself, but the probability of the gray level. The translation, rotation, and transformation in frames change only the space position in the frame, and the value of gray level change a little, so the mutual information between two successive frames is defined to describe the similarity and it is not sensitive to the motion of object. But sometimes, two frames uncorrelated completely have high mutual information and always lead to missed detection, which is because the mutual information is independent to the position of the pixel. To avoid missed detection, the co-occurrence entropy of texture information is used. The co-occurrence mutual information indicates the information about direction, continuous pixel and the change of amplitude.

It is required to use threshold to measure the dissimilarity between two successive frames to detect shot. To detect the cut shot, we compare the difference (d) with the threshold Th, if the value of d is higher, i.e. the difference between the tth and the (t+1)th frames is significant, there must be a shot change, if the value of d is lower, there is not a shot change.

3. Robust Three-dimensional Tracking

This method detects shot boundary that relies on robust tracking of salient features. By using simultaneous localization and mapping (SLAM), we are able to track objects in a scene by modeling the relative 3D positions of the features as well as the camera. In doing so, we rely on the notion that certain aspects of an image may change within a shot, but if we are unable to find any features to track between two frames, we must be observing a new shot boundary.

One of the most popular techniques proposed recently for localization using a single camera is the MonoSLAM framework. The method presents an efficient approach capable of localizing a single monocular camera and simultaneously estimating the relative 3D structure of the environment seen by the camera, all with real time performance. Our approach is to detect boundaries by robustly tracking certain objects present in the scene

This method begins by finding 16 individual salient image regions and initializing them into the probabilistic map. In this, the number of tracked regions is 16 and the individual region size is 17×17 pixels for a video with 640×480 resolution. The search for features is limited to the inner rectangle of the image plane like we only search for features within the inner 620×400 pixels.

The reason for this is to attempt and evade common locations for logos and text present in videos that

sometimes span across shot boundaries. If we take the entire image plane into consideration, the feature detector will attempt to locate and track the text located on the screen. At each iteration, the system updates the camera parameters as well as the feature vectors for all successfully tracked features. If the system fails to successfully track a given feature at any point in time, that feature is marked as one to be possibly deleted. If the same feature is unsuccessfully tracked for 15 consecutive frames, it is removed from the map and deleted from the system. If at any point the algorithm marks all of the visible features to be possibly deleted, then we assume a shot boundary has occurred.

4. Object segmentation and tracking

This algorithm combines three main techniques: the partitioned histogram comparison method, the video object segmentation and tracking based on wavelet analysis. The partitioned histogram comparison is used as the first filter to effectively reduce the number of video frames which need object segmentation and tracking. Then image segmentation algorithm for video object segmentation by the wavelet analysis and simultaneously partition and class parameters estimation (SPCPE) is used. A class is characterized by a statistical description and consists of all the regions in a video frame that follows this description; while a partition is an instance of a class. SPCPE is an unsupervised video object segmentation method by simultaneously partition and class parameters estimation.

This algorithm starts segmentation from the video frame of the first candidate shot boundary and calculates the corresponding parameters. Using these class parameters and the data, a new partition is estimated. Both the parameters and the class parameters are iteratively refined until there is no further change in them. The initial segmentation impacts on the efficiency of object segmentation and result of the ultimate segmentation.

The basic idea of the object tracking is that the discrete objects are extracted from the segmentation results of the video frames, and their bounding boxes and centroids are obtained. Then objects are tracked by calculating the differences between objects in the adjacent frames.

The segmented frame is scanned either row-wise or column-wise. If the number of rows (columns) is less than the number of columns (rows), then row-wise (column-wise) is used, respectively. Within one scanning process, we can obtain the centroids, locations and sizes of all objects in a video frame. It calculate the distance of the centroid, location, and size of the correlative object between two adjacent frames respectively, when it is less than a certain threshold value, then they are regard as the same object.

5. Local feature post refinement

This algorithm is based on mean shift procedure and feature definition. In order to reduce computational cost, the input video is first segmented based on salient change of Regional Histogram Vector distance by pair-wise feature similarity. Video frames can be treated as colored images, thus a video shot is a bundle of consecutive images. Each frame image corresponds to a data point during the mean shift smoothing procedure.

To apply mean shift kernel density estimations, the difference metric between frame images are need to be defined.

The mean shift smoothing is a typical kernel smoothing method and is widely used in image smoothing area. It smooth image noise and retain import image structures such as edges. In general, the mean shift discontinuity preserving smoothing is a density gradient estimation algorithm. Due to the similarity between image edges and video shot boundaries, it is natural that mean shift can be used for video segmentation. For flashlights and fast background change which cause the false detection, a post refinement strategy using local feature analysis is used.

Local descriptors are widely used and well suited to recognition and matching. The advantage of local descriptors is that they are robust to background clutter and other content changes. Among many local descriptors, Scale-invariant feature transform shows the best performance. The features are invariant to image scale and rotation, which can provide robust matching.

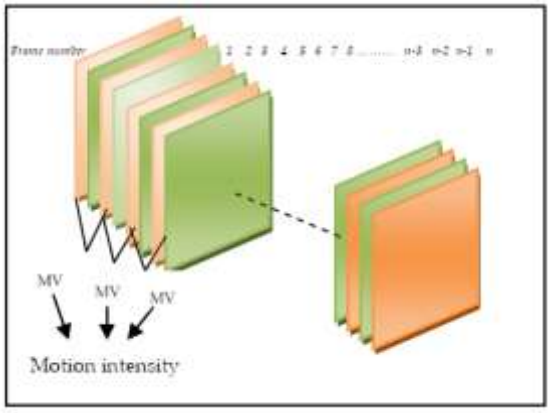
6. Motion activity descriptor

Motion is a salient feature in video, in addition to other typical image features such as color, shape and texture. The motion activity is one of the motion features used to describe the level or intensity of activity, action, or motion. 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 is expressed by an integer in the range (1-5) and higher the value of intensity, higher the motion activity. The motion features are extracted using the motion vectors.

To extract motion intensity in the uncompressed domain, we use the block matching algorithm to extract the motion intensity feature. To extract the motion intensity, the motion of the moving objects has to be estimated first. This is called motion estimation. The commonly used motion estimation technique in all the standard video is the block matching algorithm (BMA).

Block matching is used to retrieve an initial estimate of the image displacement.

The basis of any video segmentation method consists in detecting visual discontinuities along the time domain. The main idea underlying the methods of segmentation schemes is that images in the vicinity of a transition are highly dissimilar. It then seeks to identify discontinuities in the video stream. The general principle is to extract a comment on each image, and then define a distance between observations. First we extract the motion vectors between image i and image i+2 then calculates the intensity of motion, we repeat this process until reaching the last frame of the video and comparing the difference between the intensities of successive motion to a specified threshold. The idea can be visualized in fig.



III. COMPARATIVE STUDY OF SHOT DETECTION TECHNIQUES

1. Graph Theory

Although many advances have been made, false detection is always brought when using successive frame differences. It will regard erroneously flash lamp, the motion of camera and object as shot boundary. This algorithm can solve the problem effectively. This algorithm has a high accurate rate in both cut shot detection and gradual shot detection. This algorithm is proved to have better efficiency in video shot detection. It is robust to flash lamp and motion of camera and object.

For abrupt changes, the video frame differences of abrupt shot change will be larger, if successive frame differences of the different groups will exceed threshold, we consider there exists a shot boundary, otherwise it may be caused by motion of camera and object.

But it only utilizes the video feature of frame color histogram. If combined with text features, the shot detection performance would improve greatly.

2. Information Theory

It is simply to calculate the dissimilarity between two successive frames, but a threshold is required to determine a cut. It depends on the video type and the kind of shot, for example, to a far shot, the moving object is small, and the threshold should be set a low value, and to a close shot, the moving object is big and it represent a important part of the frame, the threshold will associate with a high value. So improperly setting of the threshold will result missed detection or false detection, at the same time, some video include both far shot and close shot, for example the news or the movie. If the threshold is set fixedly, it is difficult to detect the shot boundary accurately.

This algorithm based on the knowledge of information theory, the mutual information and co-occurrence mutual information is used to color feature and texture feature respectively, and then calculate the dissimilarity. This algorithm can detect cut and gradual shot simultaneously and distinguish them accurately.

3. Robust Three-dimensional Tracking

This extract salient features from a video sequence and track them over time in order to estimate shot boundaries within the video. This algorithm is used to detect abrupt transitions in videos. MonoSLAM uses as input a video stream from a single monocular camera. This data is no different from the data available in a single video of an observed scene, and thus the MonoSLAM algorithm can technically be applied to videos as well.

The main idea of our approach is to use this framework to track objects in the scene. If at any point the algorithm fails to track all of the previously-observed objects in the scene, we assume that a shot boundary has been detected. The primary reason for choosing MonoSLAM for this problem is its ability to successfully localize a monocular camera in a three dimensional environment.

This approach also accounts for features being occluded for a short amount of time. Due to the dynamics of most videos, even static features may be occluded, by either a change in camera angle, or other dynamic objects located between the feature an the camera. If the object is successfully observed and tracked after being occluded, it is no longer considered as one to be possibly deleted until the system fails to track it once again.

The main limitation currently is the system's inability to track independently moving objects. If the system only chooses to track features on moving objects, these features will eventually fail tracking and be deleted from

the system. If all the features fail at times relatively close to each other, the system may falsely detect a shot boundary. The second limitation is the relatively inefficient performance of such a system. By increasing the size of each feature and the number of features tracked by the algorithm, approach is slowed down considerably.

4. **Object segmentation and tracking** :

This algorithm is used for uncompressed video data. It combines three main techniques: the partitioned histogram comparison method, the video object segmentation, and tracking based on wavelet analysis. The partitioned histogram comparison method detects, as far as possible, candidate shot boundaries by a strict threshold. Unsupervised object segmentation algorithm based on wavelet can automatically extract the object mask map of video frame near candidate shot boundaries. By measuring the similarity between the two successive object mask maps, most of shot boundaries are detected. Then, through the object tracking analysis, the

The output taken with different videos is shown below: Two different video of a temple and sunrise has been taken through camera fur further processing. then the frames are extracted from the videos and different algorithms for shot boundary detection will be applied. In the output waveform spikes shows the boundary between two shots.

Video showing Hard cut:- Frames:-

Frame 30 Frame 31 Frame 32

Frame 33 Frame 34

Figure 3(a) :- frames for hart cut video(temple & sunrise)

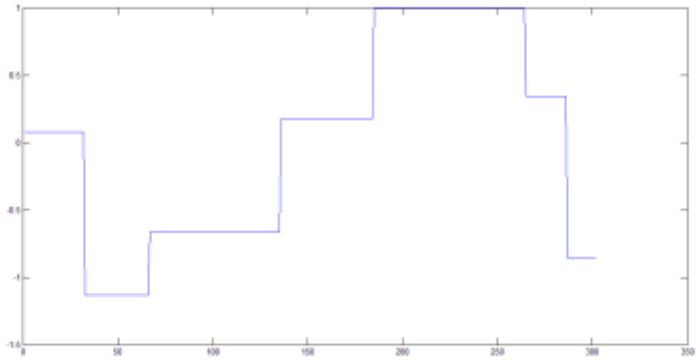


Figure 3(b):- output for Graph Theory (temple video)

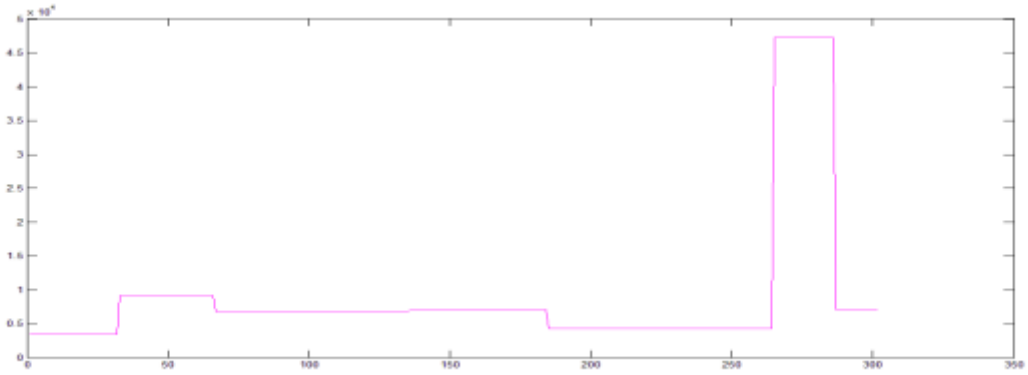


Figure 3(c):- output for Information Theory (temple video)

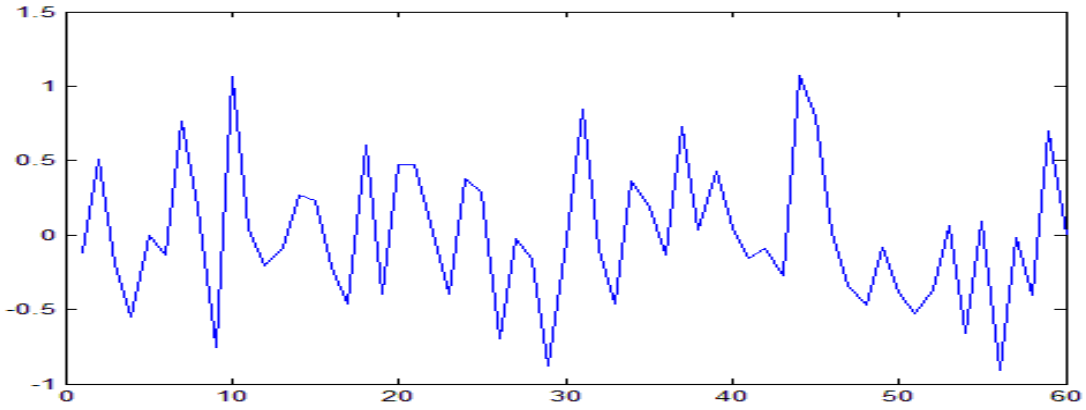


Figure 3(d)- output for Robust Three-dimensional Tracking (sunrise video)

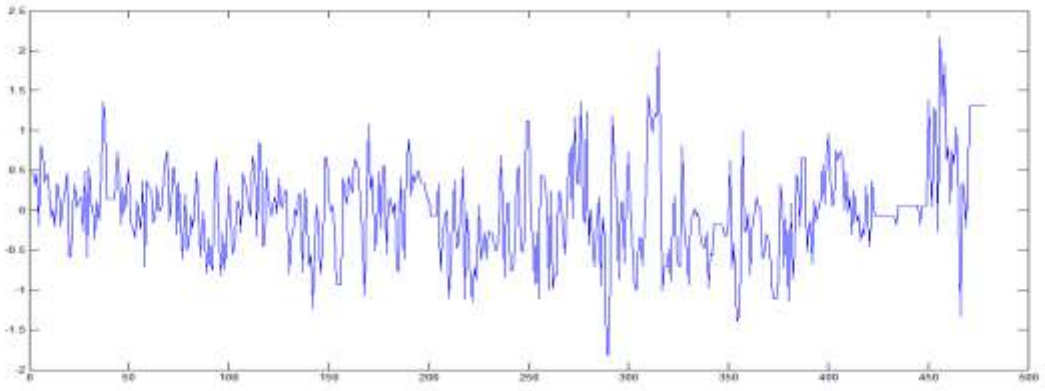


Figure 3(e)- output for Object segmentation and tracking (sunrise video)

candidate shot boundaries are further confirmed. The partitioned histogram comparison method as the first filter effectively reduced the number of video frames which need object segmentation and tracking. So the efficiency of the whole method is enhanced. The unsupervised video object segmentation and tracking based on wavelet analysis conquer the impact which the change of the low-level feathers, the motion of object and camera, and the quality of video bring in matching process.

5. Local feature post refinement

The algorithm uses feature space kernel smoothing to segment video into shots. The method is demonstrated to have high accuracy in both cuts and gradual transition detection. Video frames can be treated as colored images, thus a video shot is a bundle of consecutive images. Each frame image corresponds to a data point during the mean shift smoothing procedure. To apply mean shift kernel density estimations, the difference metric between frame images are need to be defined.

6. Motion activity descriptor

Motion activity uses motion vectors to detect motion features. To extract the motion intensity, the motion of the moving objects has to be estimated first. This is called motion estimation. The commonly used motion estimation technique in all the standard video codecs is the block matching algorithm (BMA). Block matching is used to retrieve an initial estimate of the image displacement. To obtain a dense displacement field, matching with adaptive block sizes was implemented. In this typical algorithm, a frame is divided into blocks of M × N pixels or, more usually, square blocks of N2 pixels. Then, each block undergoes translation only with no scaling or rotation. The blocks in the first frame are compared to the blocks in the second frame. Motion Vectors can then be calculated for each block to see where each block from the first frame ends up in the second frame.

Use of Motion activity detect shot boundary more accurately. It gives better performance and can be use for real time implementation.

IV. CONCLUSION

In this paper, we have discussed various shot detection algorithms which are very useful and which can detect gradual and shot simultaneously. In future, all these can combine with other information also to detect shot accurately.

REFERENCES

[1] Bezzera, F. Leite (2007) . "Using string matching to detect video transitions". *Pattern Anal*, Vol. 10, No. 1, pp.45–54.

[2] Zhang Hongjiang, Kankanhalli A and Smoliar S W(1993). "Automatic Partitioning of Full-motion Video".Multimedia Systems, Vol. 1, No. 1, pp.10-28.

[3] Zabih R, Miller J and Mai K (1999). "A Feature-based Algorithm for Detecting and Classifying Production Effects". *Multimedia Systems*, Vol. 7, No. 2, pp.119-128.

[4] Gargi, U. Kasturi R and Strayer S H (2000). "Performance characterization of video-shot change detection methods". *IEEE Trans. Circuits Syst. Video Technol*, Vol. 10, No. 1, pp.1-13

[5]Hanjalic A(2002). "Shot Boundary Detection: Unresolved and Resolved". *IEEE Trans. Circuits Syst. Video Technol*, Vol. 12, No. 2, pp.90-105.

[6]"Video shot boundary detection using motion activity descriptor", Abdelati Malek Amel, Ben Abdelali Abdessalem and Mtibaa Abdellatif, april 2010.

[7]Yufeng Li, Zheng Zhao, "A Novel Shot Detection Algorithm Based on Information Theory", IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, 2008.

[8] Arturo Donate and Xiuwen Liu, "Shot Boundary Detection in Videos Using Robust Three-Dimensional Tracking", IEEE 2010.

[9] Wenzhu Xu, Lihong Xu, "A Novel Shot Detection Algorithm Based on Graph Theory", IEEE 2010.

[10] Shouqun Liu Ming Zhu Quan Zheng, "Video Shot Boundary Detection with Local Feature Post Refinement", 2008 IEEE.

[11] Xin-Wen Xu, Guo-Hui Li, Jian Yuan, "A shot boundary detection method for news video based on object segmentation and tracking", proceedings of the seventh international conference on machine learning and cybernetics, kunming, 12-15july 2008.

[12] Don Adjeroh, M. C. Lee, N. Banda, and Uma andaswamy, "Adaptive Edge-Oriented Shot Boundary Detection," Hindawi Publishing Corporation EURASIP Journal on Image and Video Processing Volume 2009, Article ID 859371, 13 pages doi:10.1155/2009/859371, 18 May 2009.

[13] Cheol, K., Cheon, Y., Kim, G., Choi, H.: Robust scene change detection algorithm for flashlights. In: Proceedings of International Conference on Computational Science and Its Applications (ICCSA), Kuala Lumpur, Malasiya, pp. 1003—1013, 26—29 Aug 2007.

[14] So, C., Liao, H., Fan, K., chen, L.: A motion-tolerant dissolve detection algorithm. In: IEEE Trans. Multimed. 7(6), 1106—1113 (2005).

[15] Xo, Y., De, X., Tengfei, G., Aimin, W., Congyan, L.: 3-DWT based motion suppression for video shot boundary detection. In: Khosla, R., et al. (eds.) Springer-verlag, KES 2005, LNAI, vol. 3682, pp. 1204—1209 (2005).

[16] Selenick, 1W., Baraniuk, R.G., Kingsbury, N.G.: The dual tree complex wavelet transform. IEEE Signal Process. Mag. 2(6), 123—151 (2005).

[17] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibilty to structural similarity. In: IEEE Trans. Image Process. 13(4), 600—612 (2004).

[18] Wang. Z.. Simoncelli. E.P.: Translation insensitive image similarity in complex wavelet domain. In: Proc. IEEE Inter. Conf. Acoost. Speech Signal Process. II, 573—576 (2005).