A Fusion Approach of Video Shot Boundary Detection and Video Annotation: A Review

Dhruvi K Patel
Department of Civil Engineering
Ipcowala Institute of Engineering and Technology (IIET), Dharmaj

Abstract— Now a day, large volume of digital videos has become available online to the masses and people are interested in searching videos containing specific information. Video image processing is a technique to handle the video data in an effective and efficient way. It is one of the most popular aspects in the video and image based technologies such as surveillance. Shot change boundary detection is also one of the major research areas in video signal processing. The insufficiency of labeled training data for representing the distribution of the entire dataset is a major obstacle in automatic semantic annotation of large-scale video database. Semi-supervised learning algorithms, which attempt to learn from both labeled and unlabeled data, are promising to solve this problem. In this paper represent a review on video shot boundary detection and video annotation.

Key words: Video Shot Boundary Detection, fades, dissolves, wipe, multimodal, Video Annotation automatic video annotations, domain specific

I. INTRODUCTION

The advances in digital video technology and the ever increasing availability of computing resources have resulted in the last few years in an explosion of digital video data, especially on the Internet. However, the increasing availability of digital video has not been accompanied by an increase in its accessibility. This is due to the nature of video data, which is unsuitable for traditional forms of data access, indexing, search and retrieval, which are either text-based or based on the query-by-example paradigm. Therefore, techniques have been sought that organize video data into more compact forms or extract semantically meaningful information [1]. Such operations can serve as a first step for a number of different data access tasks, such as browsing, retrieval, genre classification and event detection. Here we focus not on the high level video analysis tasks themselves. but on the common basic techniques that have been developed to facilitate them. These basic tasks are shot boundary detection and condensed video representation. Shot boundary detection is the most basic temporal video segmentation task, as it is intrinsically and inextricably linked to the way that video is produced. It is a natural choice for segmenting a video into more manageable parts, and thus it is very often the first step in algorithms that accomplish other video analysis tasks, one of them being condensed video representation, described below. In the case of video retrieval, a video index is much smaller and thus easier to construct and use if it references whole video shots rather than every video frame. Since scene changes almost always happen on a shot change, shot boundary detection is indispensable as a first step for scene boundary detection. Finally, shot transitions provide convenient jump points for video browsing. Condensed representation is the extraction of a characteristic set of either independent frames or short sequences from a video.

II. SHOT BOUNDARY DETECTION

Shot boundaries can be broadly classified into two types: abrupt transition and gradual transitions. Abrupt transition is quick transition from one shot to the subsequent shot. Gradual transition occurs over multiple frames, which is produced through requisition of more detailed editing result involving numerous frames. Gradual transition can be further classified into fade out/in (FOI) transition; dissolve transition, wipe transformation, and others transformation, as per the characteristics of the different editing effects. They are required for further video analysis such as.(a) Person tracking, identification (b) High level feature detection[2]

A. Fade transition:

This is a shot transformation with the first shot slowly pass away (fade out) before the second shot gradually appears (fade in).

B. Dissolve transition:

This is a shot transformation or transition with the first shot step by step disappearing while the second shot slowly appears. In this case, the last few frames of the pass away shot momentarily overlap with the first few frames of the appearing shot.

C. Wipe transition:

This is a set of shot change methods, where the appearing and pass away shots exist at the same time in different dimensional regions of the in-between video frames. One scene slowly enters across the view while another gradually leaves.

D. Other transition types:

There is a number of innovative special outcome methods used in motion pictures. They are very rare and difficult to detect.

They provide the cue about high level semantics:

a) In video making each transition type is selected carefully to support the content & context.
b) For example, dissolve occur much more often in feature films & documentaries, while wipe usually occur in news, sports & shows.

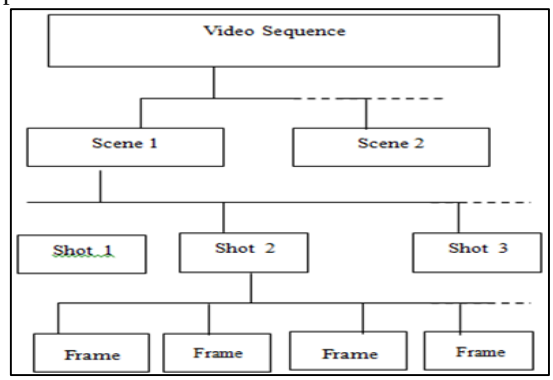

Fig. 1: Video Structure

Fig. 1: (a) Dissolve image

Fig. 1: (b) Fade - Out image

Fig. 1: (c) Fade - In image

Fig. 1: (d) Wipe image

III. VIDEO ANNOTATION

Image annotation is an active field of research that serves as a precursor to video annotation in numerous ways. Video features are often inspired and sometimes directly borrowed from image techniques and many methods for image indexing are also easily applied to video. Here we survey some of the most relevant static image annotation literature including modern trends in the field and adaptations of techniques for static image annotation to video. In the following literature the covered topics include emerging and state of the art feature extraction techniques specifically designed for video. We review image features, indexing techniques, and scalable designs that are particularly useful for working with web-scale video collections The annotation is the basis for the detection of video's semantic concepts and the construction of semantic indices for videos. The following are the approaches for video annotation (a) Statistic-based approach (b) Rule or knowledge-based approach (c)Machine learning-based approach Video annotation is very important for video management, such as video retrieval. Despite continuous efforts in inventing new annotation algorithms, the annotation performance is usually unsatisfactory, and the annotation vocabulary is still limited due to the use of a small scale training set. The effectiveness of proposed method is analyzed by valuating the precision-recall of test videos. Most of the current existing video annotation systems are video scenario based. Notes can be added to the time segments on a video timeline. A user can also view the video clip, mark a time segment, playback the segment, or attach his/her written notes to the segment. All of the annotation information is in the video level and will be mixed together, which makes it very difficult on semantic video retrieval. That is, the users cannot effectively and easily get what they want. Resolving this problem is our main objective. In addition, a semantic video annotation tool at least should support the following functionality (1) Divide a video into a number of scenes(2)Divide a scene into a number of frames(3)Develop a unified schema for video annotation(4)Annotate a scene and a frame solely. Video annotation is the task of associating graphical objects with moving objects on the screen. In existing interactive applications, only still images can be annotated, annotations can easily be attached to moving objects in the scene by novices with minimal user effort. video object annotations being used in any field in which video is produced or used to communicate information. Below show video annotation example.[4]

Fig. 2: Video Annotati

IV. RELATED WORK

Mr. Ravi Mishra, Dr. S. K Singhai, Dr. Monisha Sharma by This paper presents a new robust and efficient paradigm capable of detecting transitions in AVI videos is tested. In this research work they proposed a newly develop method in conjunction with the advantageous feature of predefined method. they frame our task as an outlier detection problem during different occurring transitions. For non-real time videos they used DTWT and nodal analysis concept for efficient result and better accuracy. They test different video types and observe their result and measure performance for various design parameters. Thus their system produces favorably good result compared to various existing approaches. The method they used for frame difference calculation i.e., histogram difference and standard deviation

of pixel intensities using contrast change parameters show high accuracy in detecting abrupt and gradual transition respectively. For real time video shot boundary detection they prepare Graphical User Interface in which three push buttons has been configured as "START VIDEO","STOP VIDEO" and "EXIT". START VIDEO button is used for video recording. Data acquisition functionality is used to record the video with the help of local webcam connected to the system. Now they are monitoring the live video in the mean while they are checking the difference between the previous image frame and current image frame by using a comparison parameter. Now they are extracting the structural features of an image by using DTCWT on the sub bands of the image. If the comparison parameter is increase beyond its limit they save that image and display the image in second window means if there is a huge motion difference comparison parameter will vary more which indicate shot boundary.[5]

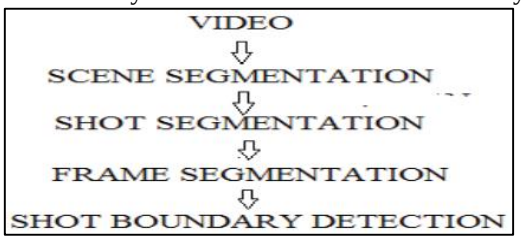

Fig. 3: Structure analysis [5]

A. Proposed methodology

1) Non-real time video shot boundary detection

For non-real time video shot boundary detection They used test videos having edit effects like fade, cut etc. First step towards the approach is to convert video into its frame. The elements of frames are known as node or vertices and the path joining the node with other node to form a tree pattern is called edge. A graph is always represented by G = (V,E). With this node They form a tree pattern based on the number of available nodes. The dissimilarities between the tree pattern is examined by DWT so form results in frame difference. This difference is calculated in terms of histogram difference for abrupt change which is insensitive to changes in color, luminance because there is a sudden change between two adjacent frames so no similarity exist there. In case of gradual, we see minute changes in luminance, color, and motion of both camera and background which is very frequent. So for this we use standard deviation of pixel intensities in combination with contrast change feature, to calculate frame difference. It deals with color, intensity feature including motion and luminance characteristics. Proposed algorithm is tested for several videos e.g. sports, animated, wildlife, cartoon, action, movies etc. Lastly the calculated difference is compared against a reference level known as threshold value. In the proposed algorithm we chose adaptive thresholding. Adaptive threshold measure the average discontinuity within a temporal domain.

2) Real time video shot boundary detection

Live video shot boundary detection will be implemented in Graphical User Interface in which three push buttons has been configured as "START VIDEO","STOP VIDEO" and "EXIT". Functions are used to get the live video streaming from the connected camera of computing system (laptop). Functions are used to configure the input settings of data acquisition for video streaming Start button is used to start the video from camera for getting the data acquisition data from the camera and display on the first axes on the window of GUI. Calculate the sensitivity and modifying the video error after that we are updating the video frames. Sensitivity will be calculated and video error has been modified after that video frames are updated. Function has been used to save the snapshot image and show on the axes of the GUI continuously. To check the motion variation of video if the motion has huge variation means if the difference of current image frame and previous image is more than expected then store the image in the data base and display in the second axes window. The structural features of an image are extracted by using DTCWT on the sub bands of the image to determine the shot boundary detection by using functions.

Fang Liu, Yi Wan Made High speed video analysis often requires efficient shot boundary detection to break a video into meaningful continuous segments. In this paper propose new techniques to improve the efficiency in terms of both processing speed and detection accuracy upon the state-of-the-art methods. Specifically, they use the HSV color space instead of the traditional RGB color space. This turns out to produce more robust detection results. In addition, each image frame is subsampled based on which a frame distance is defined. It turns out that such processing significantly lowers the computational complexity while yielding the same level of detection accuracy. Simulation results confirm the advantages of the proposed techniques over the state-of-the-art methods. Specifically, they use the value of V to compute the frame distance on a subsampled image. The advantage of this method is that its computational complexity is very low, calculation speed is faster, and the accuracy is higher. The experimental work shows that method is practical and easy to use.[6]

B. Proposed methodology

1) Cut Transition Detection

For an ideal CT, the distance of adjacent two frames calculated by should be large. On the contrary, for a shot, the distance of adjacent two frames should be small. Hence necessary conditions need to be proposed for detecting CT as follows:

D t, 1 > T c1 --------- (1)

D t - 2, 1 < T c2 n D t-l, l < T c2

n D t + l , l < T c2 n D t+2,1 < Tc2 --------- (2)

Where Tel and Tc2 are the thresholds. On one side, by (1) the CT is detected as much as possible. On the other hand, by (2) strict conditions, the interferences such as the glisten and camera shake are effectively excluded. They set the thresholds Tel = 4.9, Tc2 = 1.9, if the equation (1) and (2) is satisfied, a CT is found.

2) Gradual Transition Detection

GT is a slow transformation by multiple frames. The speed of transformation has a great relationship with the frame rate. Hence segment the video sequence into segments of length p to study. The step length that use to calculate the D t,p is p/3, in which p is the frame rate of the video sequence. The criteria of the GT are shown as following

1) The value of D t,p should satisfy the following equation:
   D t,p > T_{g1} -------- (1)
2) The following equation guarantees that the value of D t,p is maximum in range [t - 2p/3, t + 5P/3].

$$D_{t,p} = \max_{-2 \le i \le 2} D_{t+i*(p/3),p}$$ -------- (2)

3) The following condition insures that there is no CT from the t'th frame to the (t + p)'th frame if the (1) and (2) are both satisfied.

$$\arg_j (D_{t,1} > T_{c1} \cap D_{t-2,1} < T_{c2} \cap D_{t-1,1} < T_{c2} \cap D_{t+1,1} < T_{c2} \cap D_{t+2,1} < T_{c2}) \notin [0, p]$$ -------- (3)

4) In the case when all the three conditions above are met if t = t', the minimum value of D t,p will be searched using the following equation:

$$D_{min} = \min_{-5 < i < 5} D_{t'+i*(p/3),p} < T_{g2}$$ ----- (4)

Wenjing Tong, Li Song, Xiaokang Yang, Hui Qu, Rong Xie made by With the explosive growth of video data, content based video analysis and management technologies such as indexing, browsing and retrieval have drawn much attention. Video shot boundary detection (SBD) is usually the first and important step for those technologies. Great efforts have been made to improve the accuracy of SBD algorithms. However, most works are based on signal rather than interpretable features of frames. In this paper, propose a novel video shot boundary detection framework based on interpretable TAGs learned by Convolutional Neural Networks (CNNs). Firstly, adopt a candidate segment selection to predict the positions of shot boundaries and discard most non-boundary frames. This preprocessing method can help to improve both accuracy and speed of the SBD algorithm. Then, cut transition and gradual transition detections which are based on the interpretable TAGs are conducted to identify the shot boundaries in the candidate segments. Afterwards, In this Paper synthesize the features of frames in a shot and get semantic labels for the shot. Experiments on TRECVID 2001 test data show that the proposed scheme can achieve a better performance compared with the state-of-the-art schemes. Besides, the semantic labels obtained by the framework can be used to depict the content of a shot.

C. Proposed Methodology

1) Feature Extraction Using CNN:

CNN model which is similar to the Image Net challenge winning model. The main architecture of the CNN model is shown in Figure. The network contains eight layers with weights: the first five ones are convolutional layers and the remaining layers are fully-connected layers. The first convolutional layer has a total of 96 convolutional kernels of size 11×11×3 with a stride of 4 pixels. The second convolutional layer has 256 kernels of size 5 × 5 × 48. The third, fourth and fifth convolutional layers have 384 kernels of size 3×3×192. The first and second convolutional layers have local response normalization layers right behind them. After the local response normalization layers are max-pooling layers. And the fifth convolutional layer directly connects with a max-pooling layer. They train the network on Image Net dataset with 50000 iterations. Taken one frame as input, the output of the network is a probability distribution among 1000 classes. The five classes with the highest probabilities are selected as the high level features of the frame and called as the TAGs of the frame for simplicity. As shown in Fig., these TAGs can describe the contents of the frame well. Therefore, they can be used for better shot detection. Because frames in the same shot tend to share similar TAGs while frames in different shots do not.

Manu Aery and Sharma Chakravarthy made by In this paper propose improved cut detection or shot detection algorithm adapted to any domain of movies with various result sets. Shot is actually the series of interrelated consecutive pictures or frames taken from a film or part of a film contiguously and representing a continuous action in time and space. Consecutive two different shots produce an important visual discontinuity in the video stream which is called a cut. Here the video shots are assumed to be fuzzy sets and the fuzzy correlation between them is defined on the same universal support. It is shown that Spearman's rank correlation coefficient can be applied if the members of the supports are ranked according to the fuzzy membership values of each set. Next a membership-value-based fuzzy correlation measure is explained with the experimental result. Results indicate encouraging avenues for detection of hard cuts with high precision. The proposed method visualizes video sequences as fuzzy sets of vagueness and resorts to computation of fuzzy correlation measures between image frames therein to achieve the objective. It is found that the proposed approach is efficient enough to detect hard cuts in complex videos with improved accuracy. A comparative study is also presented with another soft computing based method and it is established that the proposed approach is more time efficient. Methods however, remain to be investigate to apply the proposed method for the detection of fades and dissolves in video sequences. [7]

Kyoungmin Lee, and Mathias K¨olsch made by This paper presents an evaluate the best-known approach on a contemporary, publicly accessible corpus, and present a method that achieves better performance, particularly on soft transitions. This method combines color histograms with key point feature matching to extract comprehensive frame information. Two similarity metrics, one for individual frames and one for sets of frames, are defined based on graph cuts. These metrics are formed into temporal feature vectors on which a SVM is trained to perform the final segmentation. The evaluation on said "modern" corpus of relatively short videos yields a performance of 92% recall (at 89% precision) overall, compared to 69% (91%) of the best-known method. They proposed a method for shot boundary detection (SBD) that combines two SVM classifiers; one based on color histograms, and one based on appearance features. A similarity measure between two frames was defined based on key point feature matching, and the similarity between two groups relied on graph theory and on selecting member frames according to the Fibonacci sequence. Finally, the two independent classifiers were combined into one classifier, SVMOR, through a logical OR operation. The proposed method SVMOR was compared with the best-known SBD on a novel video corpus. This corpus was assembled in consideration of characteristics of recent consumer-produced video and video-editing technology. Our experiments showed that SVMOR achieved a 12% point improvement overall and over 46% point improvement for soft cut detection. Beyond the contribution of a contemporary corpus for evaluation and a novel method for performing SBD and hope to revive interest in this topic as it is a core building block for video

indexing, search and retrieval–increasingly important capabilities for dealing with the influx of videos from the current generation of video-capturing gadgets.[8]

D. Proposed Methodology

1) Key Point Features

Key point features are appearance-based descriptors calculated at image interest points and designed to be robust to brightness, scale, rotation, and other image transformations. For the proposed method, the SIFT (scale-invariant feature transform) algorithm selects the location of feature points and represents each with descriptors. Each video frame is represented with a set of 128-dimensional descriptor vectors.

2) Color Block Histograms

Color Block Histograms (CBH) are computed in RGB space. To consider spatial information, the frame is partitioned into several blocks in which separate color histograms are computed. Each video frame is represented with a set of 2000-dimensional vectors (using five bins for each color space with 4×4 blocks)

V. CONCLUSION

It concludes overview of Image and video processing and comprehensive survey on different approach of video annotation and shot boundary detection. Different approaches are proposed for video annotation and shot boundary detection into different approach is discussed in this survey. Form this survey it concludes that detect the shot boundary from the video and also video annotation done.

REFERENCES

[1] N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, "Applications of video-content analysis and retrieval,"IEEE Multimedia Magazine, vol.9, no.3, pp. 42 – 55, July 2002.
[2] Video Shot Boundary Detection and Condensed Representation: A Review Costas Cotsaces, Student Member, IEEE, Nikos Nikolaidis, Member, IEEE,and Ioannis Pitas, Senior Member, IEEE.
[3] Video Shot Boundary Detection Techniques by Ms. Sonal P. Waghmare, Prof. A. S. Bhide, IJARECE, Volume 3, Issue 11, November 2014
[4] A Survey on Video Annotations Different Techniques by Archana.V.Potnurwar Research Scholar Computer Science &Engg Department P.I.E.T Nagpur Mohammad Atique, Ph.D Associate Professor P.G.Department Of Computer Science 2S.G.B.A.U,Amravati, IJAIS 2013.
[5] Mr Ravi Mishra, Dr. S. K Singhai ,Dr. Monisha Sharma, "Real Time And Non Real Time Video Shot Boundary Detection Using Dual Tree Complex Wavelet Transform"2015 IEEE.
[6] Kyoungmin Lee, and Mathias Kolsch, "Shot Boundary Detection with Graph Theory using Keypoint Features and Color Histograms"2015 IEEE.
[7] Sharma Chakravarthy, Aravind Venkatachalam "Improving the Video Shot Boundary Detection Using the HSV Color Space and Image Subsampling" 2015 IEEE.
[8] Biswanath Chakraborty, Siddhartha Bhattacharyya, "An Unsupervised Approach to Video Shot Boundary Detection Using Fuzzy Membership Correlation Measure" 2015 IEEE.

