A Formal Study of Video Segmentation

Ananya SenGupta
Department of Computer Science and Engineering
NIT Silchar, Assam, India

Abstract- Video scene segmentation is the first step towards automatic video annotation. For efficient video indexing and retrieval first step is to divide the video into shots and then classify the similar shots into scenes. In this paper we have discussed some existing algorithms for shot boundary detection, key frame extraction and classification of similar shots into scenes. The task of video segmentation emerges in many application areas, such as image interpretation, video analysis and understanding, video summarization and indexing, and digital entertainment.

Keywords – Cuts, fades,dissolves, shot boundary detection, key frame

I. INTRODUCTION

Video is a sequence of frames that have a high degree of temporal correlation among them. Now a days, due to the huge growth of publicly available digital data it is difficult to index and retrieve videos from large databases. The most commonly used method is to divide the video into shots, and classifying the semantically related shots into scenes. Segmentation is one of the important computer vision processes that is used in many applications such as medical imaging, machine vision, object recognition, surveillance, content-based browsing, etc. A shot is defined as unbroken sequence of frames taken from a camera. The first step towards dividing the video into shots is shot boundary detection. The types of shot transitions are gradual transition and abrupt transition.

Gradual transitions are also known as edit effects. The different types of gradual transition are fades, dissolves, wipes. A fade in is a slow decrease in brightness resulting in a black frame and a fade out is a gradual increase in brightness starting from a black frame to a bright frame. When one frame gets superimposed on another, the effect is called dissolve. In case of wipes one frame replaces another with some special shape.

Abrupt transitions are also known as cuts or hard cuts. It occurs within a single frame when stopping and restarting of camera.[1]

A shot is represented by a single frame, known as the key frame. Typically the key frame is the first frame of a shot, or can be any frame of shot. The set of all key frames are classified and shots with semantically related key frames are combined into scenes.

Fig 1: Example of abrupt shot boundary in a video

Fig 2: Example of wipe effect in a video

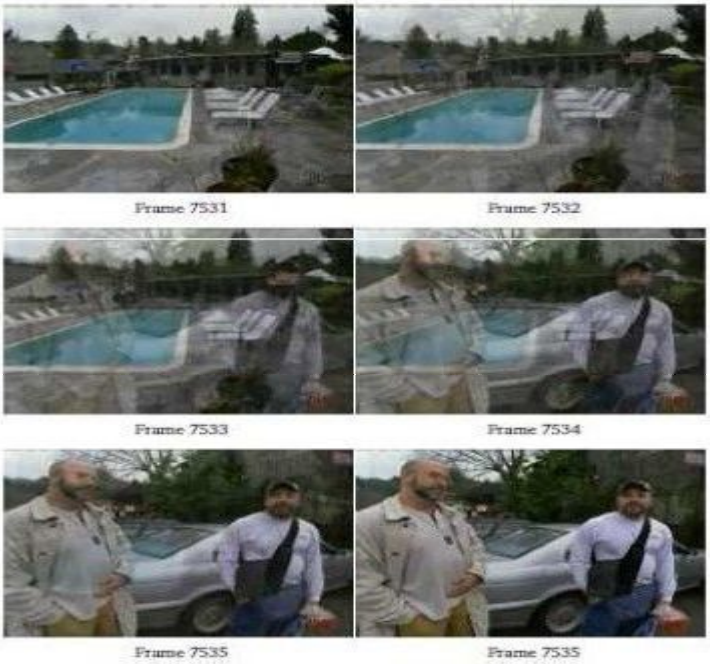

Fig 3: Example of dissolve effect

Fig 4: Example of fade in and fade out effect

II. SHOT BOUNDARY DETECTION

The different approaches for shot boundary detection are:

A. Pixel comparison–

Pixel comparison between two consecutive frames results in difference in intensity value between two corresponding pixels. The simplest way is to calculate the absolute sum of pixel differences and compare it against a threshold set experimentally.

B. Histogram comparison–

Computing the histogram differences between two consecutive frames shot boundary is detected. Color histogram of all the frames are formed and histogram difference between two successive frame is calculated, a transition is declared if the difference is greater than a threshold. It is found that YIQ, ¸L*a*b and Munsell color coordinate spaces perform well, followed by HSV, L*u*v and RGB. [2]. Histograms are invariant to image rotation and changes slowly under the variations of viewing angle and scale [3]. The frames with different content may have similar histograms, this limitation is resolved in [4]. Histogram comparison can be global and local. In global comparison histogram of the entire frame is computed whereas in local histogram comparison is similar to block based comparison. The frame is divided into blocks and histogram of each block is compared with corresponding block of next frame. Local histogram comparison uses spatial information.

C. Feature based comparison:

Global features of the frame such as color, texture, shape are computed and compared with the consecutive frames.

Local features with in a frame can also be evaluated using SIFT [5], SURF [6], MSER [7] and compared with the next frame. A transition is declared if the number of matched features is less than threshold set experimentally.

SIFT detector consists of four steps: scale-space extrema detection, keypoint localization, orientation assignment and keypoint descriptor. In scale space extrema detection difference-of-Gaussian function is used to identify the interest points, which were invariant to scale and orientation [8]. In key point localization low contrast points are rejected to eliminate the edge responses. The third step involves generating hessian matrix to compute the principal curvatures and eliminate the keypoints. For orientation assignment, an orientation histogram was formed from the gradient orientations of sample points within a region around the keypoint.

SIFT builds an image pyramids, filtering each layer with Gaussians of increasing sigma values and taking the difference whereas SURF creates a stack without 2:1 down sampling for higher levels in the pyramid resulting in images of the same resolution [9]. SURF filters the stack using a box filter approximation of second-order Gaussian partial derivatives, due to the use of integral images.

MSER(Maximally stable extremal regions) are set of pixels within frame such that the set is closed under continuous transformation of image coordinates, the set of regions are closed under monotonic transformation of image intensities. MSER are regions whose intensity is strictly greater than or strictly less than the image boundary.

III. SURVEY ON SHOT BOUNDARY DETECTION

Liuhong Liang et al[10] , presented a Shot Boundary Detection technique using text Information where a number of edge-based techniques have been proposed for detecting abrupt shot boundaries and to eliminate the effect of flashlights in many video types, such as sports, news, entertainment videos. The similar approach is also used in[13].

[12] Uses edge detection technique, by counting the number of entering and exiting edges in two consecutive frames, shot transition is detected.

Daniel De Menthon et al [11] proposed shot boundary detection technique based on image correlation features. The cut detection is based on 2max ratio criterion in a sequential image buffer. The dissolve detection is based on the skipping image difference and linearity error in a sequential image buffer. This paper is an implementation of global features in an image.

The color and texture features are extracted by wavelet transform, then the dissimilarity between two successive frames are evaluated which colligates the mutual information of color feature and texture feature. The threshold is selected adaptively based on image entropy/intensity.[14]

Cernekova et al [15] used mutual information to detect abrupt transition whereas they used joint entropy between frames to detect the gradual transition. The amount of information transported to next frame is called the mutual information.

Guillermo Cisneros et al [16] proposed a unified model on video shot transition detection. The approach presented here is based on mapping the space of inter-frame distances into a new space of decision better suited to achieving a sequence independent thresholding.

Joyce and Liu [17] presented two algorithms for detecting dissolve and wipes. The first is a dissolve detection algorithm which is implemented both as a simple threshold-based detector and as a parametric detector by modeling the error properties of the extracted statistics. The second is an algorithm to detect wipes based on image histogram characteristics during transitions.

Jinchang Ren et al [18] proposed a paper on Shot Boundary Detection where DC images are extracted from MPEG video and features are extracted and selected through AdaBoost for cut detection. Several local indicators are extracted from MPEG macro blocks, and Ada Boost is employed for feature selection and fusion. The selected features are then used in classifying candidate cuts into five sub-spaces via pre-filtering and rule based decision making, then the global indicators of frame similarity between boundary frames of cut candidates are examined using phase correlation of dc images. Fade transition is detected by considering the change in luminance, where the intensity values shows a V-shaped graph, detection of dissolve is identification of downward parabola / U- shaped pattern.

In [19], both local and global features are used for shot boundary detection. A frame is considered as abrupt boundary if it has less/ no matched SURF features [9] with its successive frame. Entropy of all the frames are computed and compared with adjacent frames for fade detection. In case of fade out entropy of frames increases while in case of fade in entropy decreases.

The video frames are divided into several different groups through performing graph-theoretical algorithm. In [20], shot boundary detection based on graph theory is proposed.

Using HSV color histrogram difference and adaptive threshold hard cuts are detected in [21]. They also calculated the local histogram difference and local adaptive threshold for gradual shot transition detection.

Pablo Toharia et al [22] proposed shot boundary detection technique using Zernike moments in multi-GPU and multi-CPU architectures along with the different possible hybrid combinations based on Zernike moments.

In [23] the positions of the shot boundaries and lengths of gradual transitions are predicted using adaptive threshold and most of the non-boundary frames are discarded. This method is based on segment selection and singular value decomposition (SVD).

X. Gao, J. Li, and Y. Shi [24] proposed an algorithm that extracts a set of corner-points from the first frame of a shot. Using Kalman Filtering these features are matched with the features of the subsequent frames, accordingly with the changing pattern of pixel intensity shot boundary is detected.

IV. KEY FRAME EXTRACTION

Key-frame selection techniques extract one or several frames from each video that represents the entire video sequence. A frame is selected from every shot that represents the entire shot is key frame. Key frame selection techniques are based on density measurement metrics, such as motion, color or edge density [25]. Key-frames are often a random selection between abrupt cuts or fades [26]. In case of sport videos, as there are no clear scenes, key-frames are a group of frames with similar motion [27]. The features used for similarity measures are color, shape, and motion as well as entropy value of frames [28]. Key-frame selection involves (i) motion feature analysis (ii) image quality measurement and (iii) frame based similarity. The selection process is based on the sum of the magnitudes from components of optical flow at each pixel of frame at time t.

Gresle and Huang [29] computed the intra and reference histograms and an activity indicator is generated. Based on the activity curve, the local minima are selected as the key frames.

Wolf [30] proposed an algorithm for key frame extraction based on motion analysis. The algorithm computes the optical flow for each frame and changes in the optical flow along the sequence using a simple motion metric. Key frames are then found at places where the metric has its local minima.

Key frames are extracted using the characteristics of I frames, P frames and B frames in the MPEG video stream after shot segmentation. If a scene cut occurs, the first I frame is chosen as a key frame. Fidelity and compression ratio are used to measure the validity of the method. [31]

In [32] key-frames are extracted by maximizing the AIKLD (Average Interclass Kullback Leibler Distance) of major video objects in the unified feature space. The algorithm provides an integrated platform where the inherent and explicit relationship between key-frames and video objects is found.

Key-frame selection methods are based on the measurement of frame content, and the key-frames were selected while comparing their similarity pixel by pixel. [33]

IV. SCENE CLASSIFICATION

Key frames are classified depending on various categories such as color, texture, shape, text etc. Similar key frames are classified into same categories. The shots belonging to similar key frames are combined into scenes.

Y. Gong et al [34] proposed a technique using Singular Value decomposition. Based on the visual contented feature metrics are computed for key-frames and compared. This type of classification is called hierarchical classification. Similarly in KKN (k nearest neighbour classification), the system finds the k-nearest neighbours among the training features vectors, and uses the categories of the k neighbours to determine the category of the test vector whereas in NB (Naïve Bayes) probabilistic classification, features joint probabilities is used to eliminate the probabilities of a given data input. [35]

[36] Presents methods based on motion feature. Two methods of extracting motion from a video sequence are foreground object motion and background camera motion. These dynamics are extracted, processed and applied to classify three broad classes: sports, cartoons and news.

Visual features such as color and motion are extracted for the key frame and shots. The rules between these features and shots generes are discovered by applying decision tree and the rules are finally exploited to classify the scenes. [37]

For scene detection key frames are extracted from abrupt boundaries and K size window is used for similarity of key frames in temporal order. They proposed a new method for shot boundary and scene detection based on frame entropy and SURF features. [38]

V. EXPERIMENT AND RESULT

The test set for this evaluation experiment is taken from OPEN VIDEO PROJECT database.

Table -1 Experiment Result of Shot boundary detection

video	cuts	[11]	[12]	[14]	[15]	[16]
ani002	**7**	**6**	**5**	**6**	**4**	**5**
indi004	**8**	**7**	**7**	**6**	**7**	**7**
indi009	**5**	**5**	**5**	**5**	**4**	**4**
indi008	**14**	**13**	**12**	**13**	**11**	**12**
ani003	**24**	**21**	**23**	**21**	**22**	**23**
indi002	**6**	**6**	**6**	**6**	**5**	**6**

Table 1 shows the number of abrupt boundary detection for the respective algorithms. Cut column shows the number of cuts actually present in the video sequence. The results are the comparisons of some existing algorithms.

VI.CONCLUSION

Now a days, segmentation plays huge role in computer vision process as it very important to retrieve videos from large databases like YouTube, daily motion etc. In this paper we have presented different approaches for shot boundary detection and scene classification. The challenges in shot boundary detection are gradual transition detection in video. Early work focused on cut detection, while more recent techniques deal with gradual transition detection. In case of gradual transition detection, an important evaluation criterion is the algorithm's ability to determine exactly between which frames the transition occurs and to classify the type of the transition.

REFERENCES

[1] Koprinska, Irena, and Sergio Carrato. "Temporal video segmentation: A survey." *Signal processing: Image communication* 16.5 (2001): 477-500. A. A. Reddy and B. N. Chatterji, "A new wavelet based logo-watermarking scheme," Pattern Recognition Letters, vol. 26, pp. 1019-1027, 2005.

[2] U. Gargi, S. Oswald, D. Kosiba, S. Devadiga, R. Kasturi, Evaluation of video sequence indexing and hierarchical video indexing, in: Proceedings of SPIE Conference on Storage and Retrieval in Image and Video Databases, 1995, pp. 1522}1530..

[3] M.J. Swain, Interactive indexing into image databases, in: Proceedings of SPIE Conference on Storage and Retrieval in Image and Video Databases, 1993, pp. 173}18.

[4] G. Pass, R. Zabih, Comparing images using joint histograms, Multimedia Systems (1999) in press.

[5] Juan, Luo, and Oubong Gwun. "A comparison of sift, pca-sift and surf."*International Journal of Image Processing (IJIP)* 3.4 (2009): 143-152.

[6] Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. "Surf: Speeded up robust features." *Computer Vision–ECCV 2006*. Springer Berlin Heidelberg, 2006. 404-417.

[7] Donoser, M.; Bischof, H., "Efficient Maximally Stable Extremal Region (MSER) Tracking," *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* , vol.1, no., pp.553,560, 17-22 June 2006

[8] D. Lowe."Distinctive Image Features from Scale-Invariant Keypoints", IJCV, 60(2):91–110, 2004.

[9] Yang zhan-long and Guo bao-long. "Image Mosaic Based On SIFT", International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp:1422-1425,2008.

[10] Liuhong Liang Liuhong Liang, Yang Liu, Hong Lu, Member, IEEE, Xiangyang Xue, and Yap-Peng Tan, Senior Member, IEEE [2005],"A Enhanced Shot Boundary Detection Using Video Text Information", IEEE Transactions on Consumer Electronics, Vol. 51, No. 2, MAY 2005

[11] Daniel DeMenthon [2006] "Shot boundary detection based on Image correlation features in video"

[12] Zabih, R., Miller, J., & Mai, K. (1999). A feature-based algorithm for detecting and classifying production effects. *Multimedia systems*, *7*(2), 119-128.

[13] Hauptmann, A., and M. Smith. "Text, Speech, and Vision for Video Segmentation: The Informedia TM Project." *AAAI fall symposium, computational models for integrating language and vision*. 1995.

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

[15] Cernekova, Z., Pitas, I., & Nikou, C. (2006). Information theory-based shot cut/fade detection and video summarization. *Circuits and Systems for Video Technology, IEEE Transactions on*, *16*(1), 82-91.

[16] Jesús Bescós, Guillermo Cisneros, José M. Martínez, José M. Menéndez, and Julián Cabrera[2005] "A Unified Model for Techniques onVideo-Shot Transition Detection" IEEE transactions on multimedia, vol. 7, no. 2, april 2005

[17] Joyce, R.A., Liu, B.: Temporal segmentation of video using frame and histogram space. IEEE Trans. Multimedia 8(1) 130–140 (2006)

[18] Jinchang Ren; Jianmin Jiang; Juan Chen, "Shot Boundary Detection in MPEG Videos Using Local and Global Indicators," Circuits and Systems for Video Technology, IEEE Transactions on , vol.19, no.8, pp.1234,1238, Aug. 200

[19] Baber, J., Afzulpurkar, N., & Satoh, S. I. (2013). A framework for video segmentation using global and local features. *International Journal of Pattern Recognition and Artificial Intelligence*, *27*(05).

[20] Wenzhu Xu & Lihong Xu[2010] "A Novel Shot Detection Algorithm Based on Graph Theory".

[21] Hua Z., Ruimin Hu, and Lin Song. "A shot boundary detection method based on color feature." *Computer Science and Network Technology (ICCSNT), 2011 International Conference on*. Vol. 4. IEEE, 2011

[22] Pablo Toharia et al "on Shot boundary detection using Zernike moments in multi-GPU multi-CPU architectures"- Journal of Parallel and Distributed Computing Volume 72 Issue 9, September, 2012

[23] Zhe Ming Lu and Yong Shi "Fast Video Shot Boundary Detection Based on SVD and Pattern Matching-"Image processing IEEE Transactions (Volume:22 , Issue: 12), Dec. 2013

[24] Gao, Xinbo, Jie Li, and Yang Shi. "A video shot boundary detection algorithm based on feature tracking." *Rough Sets and Knowledge Technology*. Springer Berlin Heidelberg, 2006. 651-658

[25] J. Luo, C. Papin, and K. Costello, "Towards extracting semantically meaningful key frames from personal video clips: from humans to computers," IEEE Transactions on Circuits and Systems for Video Technology, vol. 19, no. 2, pp. 289–301, 2009.

[26] Z. Cernekov´a, C. Nikou, and I. Pitas, "Entropy metrics used for video summarization," in Proceedings of the 18th Spring Conference on Computer Graphics, 2002, pp. 73–82.

[27] L. Li, X. Zhang, Y.-G. Wang, W. Hu, and P. Zhu, "Nonparametric motion feature for key frame extraction in sports video," in Chinese Conference on Pattern Recognition, 22-24 2008, pp. 1–5.

[28] M. Mentzelopoulos and A. Psarrou, "Key-frame extraction algorithm using entropy difference," in Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, 2004, pp. 39–45.

[29]]P. Gresle, T. S. Huang, "Gisting of video documents: a key frames selection algorithm using relative activity measure," The 2nd International Conference On Visual Information System, 1997

[30] W. Wolf, "Key frame selection by motion analysis," Proc. IEEE Int. Conf. Acoust., Speech Signal Proc., vol. 2, pp. 1228–1231, 1996.

[31] Liu, Guozhu, and Junming Zhao. "Key frame extraction from MPEG video stream." *Information Processing (ISIP), 2010 Third International Symposium on*. IEEE, 2010.

[32] Song, Xiaomu, and Guoliang Fan. "Joint key-frame extraction and object-based video segmentation." *Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on*. Vol. 2. IEEE, 2005.

[33] J. Rong, W. Jin, and L. Wu, "Key frame extraction using inter-shot information," in IEEE International Conference on Multimedia and Expo, vol. 1, 2004, pp. 571–574 Vol.1.

[34] Gong, Yihong, and Xin Liu. "Video shot segmentation and classification." Pattern Recognition, 2000. Proceedings. 15th International Conference on. Vol. 1. IEEE, 2000.

[35] Qi, Yanjun, Alexander Hauptmann, and Ting Liu. "Supervised classification for video shot segmentation." Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on. Vol. 2. IEEE, 2003.

[36] Roach, Matthew J., John SD Mason, and Mark Pawlewski. "Video genre classification using dynamics." Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on. Vol. 3. IEEE, 2001.

[37] Zhao, Shiwei, et al. "A Data-Mining Based Video Shot Classification Method." Image and Signal Processing, 2009. CISP'09. 2nd International Congress on. IEEE, 2009.

[38] Baber, J., Afzulpurkar, N., & Satoh, S. I. (2013). A framework for video segmentation using global and local features. International Journal of Pattern Recognition and Artificial Intelligence, 27(05).