Temporal Video Segmentation by Graph Partitioning

Z. Černeková, N. Nikolaidis and I. Pitas
Department of Informatics
Aristotle University of Thessaloniki
Box 451, Thessaloniki 541 24, GREECE
E-mail: (zuzana, nikolaid, pitas)@zeus.csd.auth.gr

***Abstract*— A novel temporal video segmentation method that, in addition to abrupt cuts, can detect with very high accuracy gradual transitions such as dissolves, fades and wipes is proposed. The method relies on evaluating mutual information between multiple pairs of frames within a certain temporal frame window. This way we create a graph where the frames are nodes and the measures of similarity correspond to the weights of the edges. By finding and disconnecting the weak connections between nodes we separate the graph to subgraphs ideally corresponding to the shots. Experiments on TRECVID2004 video test set containing different types of shot transitions and significant object and camera motion inside the shots prove that the method is very efficient.**

I. Introduction

The growing amount of digital video is driving the need for more effective methods for indexing, shot classification, browsing, searching, summarization and retrieval of video based on its content. The detection of shot boundaries provides a base for nearly all video abstraction and high-level video analysis approaches [1]. A video shot can be defined as a sequence of frames captured by *one camera in a single continuous action in time and space* [2].

Early work on the shot boundary detection focused on the abrupt cut detection, while more recent techniques deal with the more difficult problem of gradual transitions detection. An abrupt cut is usually detected when a certain difference measure between consecutive frames exceeds a threshold. The difference measure is computed either at a pixel level or at a block level. Intensity histograms or color histograms have been used for shot boundary detection in [3], [4], [5], [6]. In some works more complex features, such as image edges [7] or motion vectors [8] have been examined. In general, a lot of different approaches have been applied in shot boundary detection. However, it is widely recognized that histogram based [3], [4], [6] and feature based approaches offer some of the best results. A comparison of existing methods for shot boundary detection is presented in [3], [9].

Detection of gradual transitions, such as dissolves, fade-ins, fade-outs, and wipes is examined in [10], [11], [12]. A *fade* is a transition involving gradual diminishing (fade-out) or increasing (fade-in) visual intensity. A *dissolve* can be viewed as a fade-out and fade-in with some temporal overlap. Transitions between shots are widely used in TV. Existing techniques for detecting transitions rely on twin thresholding [1] or gray level statistics [9] and have a relatively high false detection rate. Hampapur et al [13] suggested a shot detection scheme based on modeling video edits. Lienhart [11] casts the problem of automatic dissolve detection as a pattern recognition and learning problem. In [14] Li and Wei proposed a dissolve detection method based on the analysis of Joint probability Images.

The use of mutual information as a similarity metric in shot detection has been proven very successful. Mutual information was mainly used for the detection of abrupt video shot cuts [15], [16]. In [17] mutual information was used also for the detection of gradual transitions. In this paper a method for automated shot boundary detection based on the comparison of more than two consecutive frames is proposed. Within a temporal window W we calculate the mutual information for multiple pairs of frames. This way we create a graph for the video sequence where the frames are nodes and the measures of similarity correspond to the weights of the edges. By finding and disconnecting the weak connections between nodes we separate the graph to subgraphs ideally corresponding to the shots. The major contribution of the proposed algorithm is the utilization of information from multiple frames within a temporal window, which ensures effective detection of gradual transitions in addition to abrupt cut detection. The experimental results indeed prove that the method is capable of detecting both abrupt cuts and gradual transitions with very good precision and recall rates.

The remainder of the paper is organized as follows: In Section 2, a description of the proposed shot detection method is provided. Experimental results are presented and commented in Section 3. Finally, concluding remarks are drawn in Section 4.

II. Shot detection

In this paper we focus on the detection of gradual transitions such as dissolves and wipes, which are the most difficult to be detected. Unlike the abrupt cuts, the gradual transition spreads across a number of frames. In order to capture better the duration of the transition, we consider not only two consecutive frames but take into account all frames within a certain temporal window W.

One can see the problem of shot cut detection as a problem of graph partitioning. The video frames are represented as nodes in a graph whose edge weights correspond to the pairwise similarities of the frames. In order to detect the video shots, one has to discover and disconnect the weak connections

between the nodes, thus partitioning the graph to subgraphs ideally corresponding to the shots.

As a measure of similarity we have chosen mutual information (MI) since it has been shown to provide very good results for abrupt cut detection [15], because it exploits the inter-frame information flow in a more compact way than frame subtraction. Difference in content between two frames, leads to low values of mutual information.

In our case, the mutual information between two frames is calculated separately for each of the RGB color components. In the case of the R component, the element $\mathbf{C}^R_{t,t+1}(i,j)$, $0 \leq i,j \leq N-1$, N being the number of gray levels in the image, corresponds to the probability that a pixel with gray level i in frame $\mathbf{f}_t$ has gray level j in frame $\mathbf{f}_{t+1}$. In other words, $\mathbf{C}^R_{t,t+1}(i,j)$ equals to the number of pixels which change from gray level i in frame $\mathbf{f}_t$ to gray level j in frame $\mathbf{f}_{t+1}$, divided by the total number of pixels in the video frame. The mutual information $I^R_{k,l}$ of frames $\mathbf{f}_k$, $\mathbf{f}_l$ for the R component is expressed as [15]:

$$I^R_{k,l} = -\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \mathbf{C}^R_{k,l}(i,j) \log \frac{\mathbf{C}^R_{k,l}(i,j)}{\mathbf{C}^R_k(i)\mathbf{C}^R_l(j)}. \quad (1)$$

The total mutual information (MI) calculated between frames $\mathbf{f}_k$ and $\mathbf{f}_l$ is defined as:

$$I(f_k, f_l) = I^R_{k,l} + I^G_{k,l} + I^B_{k,l}. \quad (2)$$

Since our aim is to cluster the sequence into shot, we do not have to calculate the complete graph for the whole video sequence by connecting all possible pairs of frames, because relations between frames, which are far apart are not important for the shot cut detection task. Thus, we create the graph edges only between nodes in a sliding temporal window W.

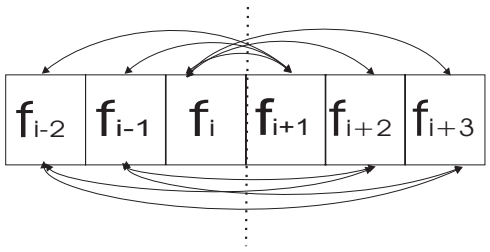

Fig. 1. Diagram showing pairs of frames, which contribute to $I_{cumm}(i)$ for a window size of $N_W = 6$.

Within the one-dimensional temporal window W of size N_W, which is centered around frames $\mathbf{f}_i$ and $\mathbf{f}_{i+1}$ we connect (i.e. calculate the measure of similarity) $(N_W/2)^2$ pairs of frames as shown in Figure 1. More specifically, we connect all pairs of frames $\mathbf{f}_k$, $\mathbf{f}_l$ where $k \leq i$, $l > i$. All these edges provide information on whether frames $\mathbf{f}_i$, $\mathbf{f}_{i+1}$ belong to a transition or are the last and first frame of two shots separated by abrupt cut. By sliding the window over all frames we create a graph for all the video sequence.

Then for each position of the window a cumulative measure which combines information from all connected frame pairs within the window is calculated as follows:

$$I_{cumm}(i) = \sum_{k=i-\delta}^{i} \sum_{l=i+1}^{i+\delta} I(f_k, f_l) \quad (3)$$

where $\delta = N_W/2$ is half the size of the temporal window W.

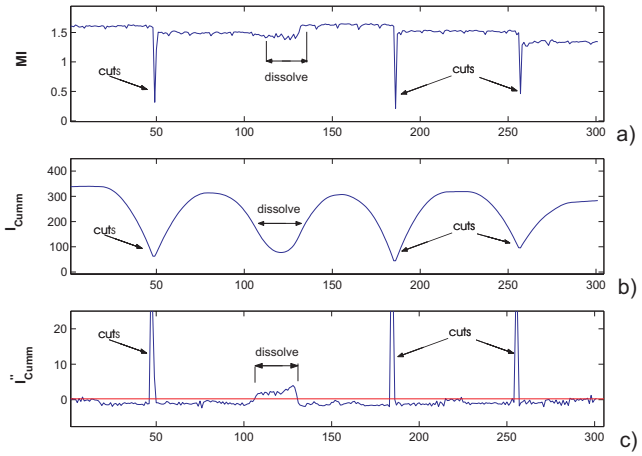

Fig. 2. Plot of patterns for a part of video sequence that correspond to a dissolve: a) mutual information between two consecutive frames, b) cumulative mutual information I_{cumm} and c) second derivative of I_{cumm}.

In case of mutual information calculated between two consecutive frames, the abrupt cuts are well detectable as single peaks but gradual transitions like dissolves do not always show any characteristic pattern. During a gradual transition, the content of the first shot diminishes whereas the content of the second one appears gradually. An example of a dissolve pattern is shown in Figure 2. In the first part of the gradual transition the amount of information shared between two frames and therefore their mutual information decreases while in the second part it increases. In [17] the authors identify a dissolve when in a sequence of mutual information values calculated between two frames a "V" pattern is formed. However, on Figure 2a) one can see that a dissolve does not always produce this pattern. Since I_{cumm} encompasses the mutual information between pairs of frames, it is reasonable to expect that the value I_{cumm} will also decrease and reach a local minimum in case of gradual transition. By observing Figure 2 one can easily conclude that a dissolve is much more prominent in the I_{cumm} curve than in the MI curve. The abrupt cuts are also clearly distinguished in this curve. Therefore, in order to detect the video shot transitions, we have to locate local minima of I_{cumm}. However, local minima in I_{cumm} can be also caused by a significant motion within a shot. In such cases, the values of I_{cumm} around the local minimum will not be very low. Therefore, we keep for further examination only I_{cumm} values which are below an experimentally chosen threshold T, as shown in Figure 3.

$$f_{grad} = \{i; I_{cumm}(i) < T\} \quad (4)$$

In the next step, we find all frames where the first derivative of $I_{cumm}(i), i \in f_{grad}$ changes sign (crosses zero) and check

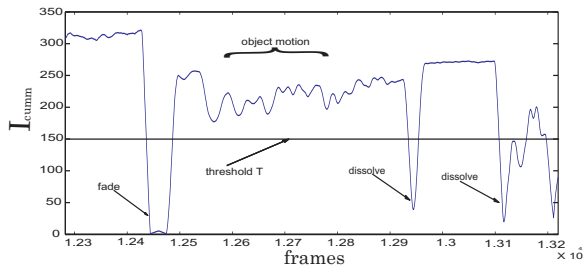

Fig. 3. Plot of I_{cumm} pattern for a part of video sequence with the threshold T used to separate the local minima caused by the motion within the shot and the local minima caused by the transitions.

if the corresponding points have positive values of the second derivative.

Once minima of the I_{cumm} are identified, we search for the start t_s and end t_e time instant (transition boundaries) of the transition around the minimum. Boundaries are detected as the points left/right of the minimum where the second derivative I''_{cumm} crosses zero in the so called inflection points. Since, a gradual transition has some duration, to declare a gradual transition, at least three frames should be involved:

$$t_e - t_s \geq 3 \tag{5}$$

If this condition does not hold, i.e. if $t_e - t_s \leq 2$ then an abrupt cut is declared.

Figure 4 shows the signal of mutual information calculated between two consecutive frames (4a), the cumulative MI I_{cumm} calculated for a window of size $N_W = 30$ (Figure 4b) and the second derivative of I_{cumm} (Figure 4c). All signals are calculated for the same part of a video sequence.

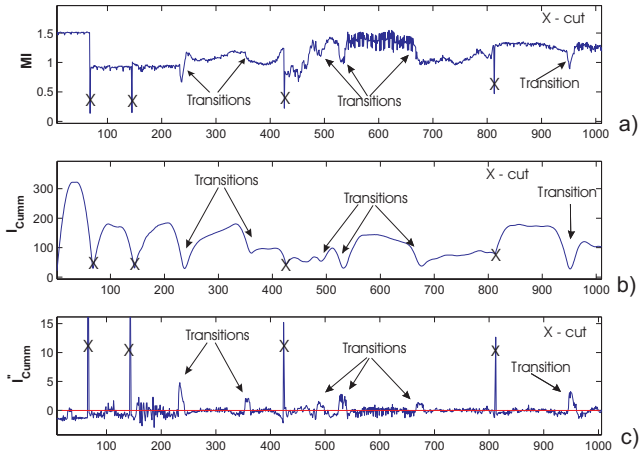

Fig. 4. a) Plot of mutual information between two consecutive frames, b) cumulative MI I_{cumm} calculated within a window and c) second derivative of I_{cumm} for the same part of video sequence.

The proposed approach has a big advantage in cases where camera flashes are present in a video sequence. This is due to the fact that mutual information is known to be much less sensitive to camera flashes comparing to histogram comparisons. Moreover, since I_{cumm} is evaluated on the bases of multiple pairs of frames whereas a flash affects only 2-3 frames, its effect on I_{cumm} is minimal. In Figure 5 one can see that even if peaks appear in the mutual information pattern due to flashes (Figure 5a), no such peaks appear in I_{cumm} (Figure 5b).

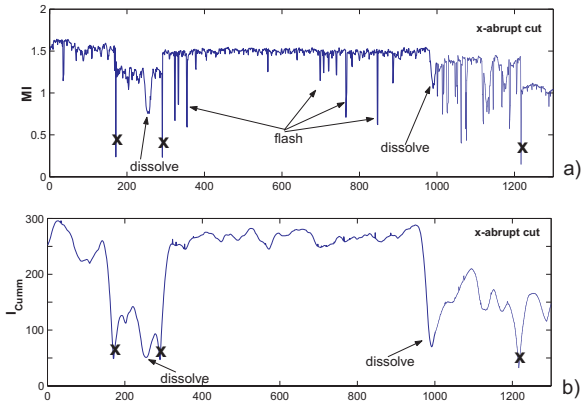

Fig. 5. *A part of video sequence containing many camera flashes: a) mutual information and b) the pattern of* I_{cumm}.

III. Experimental results

The proposed method was tested on newscasts from the reference video test set TRECVID 2004 [18] having many commercials in-between. Both news and commercials are characterized by significant camera effects like zoom-ins/outs and pans, abrupt camera movements and significant object and camera motion inside single shots. Video sequences of almost 3 hours duration have been digitized with a frame rate of 29.97fps at a resolution of 352×240. Downsampled videos with resolution 176×120 were used in our experiments to speed up the calculations. The ground truth provided by TRECVID was used for evaluating the results. The corresponding data are depicted in Table I.

TABLE I
THE VIDEO TEST SET.

video	CNN & ABC news
frames	307204
cuts	1377
fades	111
dissolves	763
others	136

Let GT denote the ground truth, Seg be the segmented (correct and erroneously) shots using our method and $|E|$ be the number of elements of a set E. In order to evaluate the performance of the segmentation method presented in Section II, the following measures, inspired by receiver operating characteristics in statistical detection theory [9], [19] were used:

- The *recall* measure, also known as the true positive function or sensitivity, corresponds to the ratio of correct experimental detections over the number of all true

detections:

$$Recall = \frac{|Seg \bigcap GT|}{|GT|}. \quad (6)$$

- The *precision* measure corresponds to the accuracy of the method considering false detections and it is defined as the number of correct experimental detections over the number of all experimental detections:

$$Precision = \frac{|Seg \bigcap GT|}{|Seg|}. \quad (7)$$

We have tested our method for different sizes $N_W \in \{5, \ldots, 40\}$ of the temporal window W. For small values of N_W the gradual transition is not captured well whereas for big values of N_W the computation of I_{cumm} becomes more is time consuming without adding any information. Moreover, if the window size is very big, it might cover two transitions. The average length of a gradual transition was found to be about 25 frames. In order to take into account all the above we decided to use for our experiments window of size $N_W = 30$ as it provided the best results.

Table II summarizes the recall and precision rates obtained by the proposed method for cuts, gradual transitions, as well as for both of them. The obtained results are very good and promising. The boundaries of gradual transitions were detected within a precision of ± 2 frames. In most cases, the boundaries were recognized without any error. There were no false detections due to camera flashes. In some cases, false detections appeared in the case of commercials where artistic camera edits were used.

TABLE II
SHOT DETECTION RESULTS.

CNN & ABC news	Recall	Precision
cuts	0.95	0.93
gradual transitions	0.80	0.83
overall	**0.87**	**0.89**

IV. CONCLUSIONS AND DISCUSSION

A new technique for automated shot boundary detection in video sequences was presented. The method relies on evaluating mutual information within a certain temporal frame window. The video frames are represented as nodes in a graph, whose edge weights signify the pairwise similarities of data points. Clustering is realized by partitioning the graph into disjoint sub-graphs. The method is able to detect efficiently abrupt cuts and all types of gradual transitions, such as dissolves, fades and wipes with very high accuracy. Results are demonstrated by experiments on TRECVID2004 video test set containing different types of shots with significant object and camera motion inside the shots.

Future research includes automatic setting of the threshold by using adaptive thresholding methods.

ACKNOWLEDGMENT

The presented work was partially developed within VISNET, a European Network of Excellence (http://www.visnet-noe.org), funded under the European Commission IST FP6 program.

REFERENCES

[1] A. D. Bimbo, *Visual Information Retrieval*. Morgan Kaufmann Publishers, Inc, San Francisco, California, 1999.

[2] X. U. Cabedo and S. K. Bhattacharjee, "Shot detection tools in digital video," in *Proc. Non-linear Model Based Image Analysis 1998*, Springer Verlag, Glasgow, July 1998, pp. 121–126.

[3] A. Dailianas, R. B. Allen, and P. England, "Comparison of automatic video segmentation algorithms," in *Proc., SPIE Photonics East'95: Integration Issues in Large Commercial Media Delivery Systems*, vol. 2615, Philadelphia 1995, Oct. 1995, pp. 2–16.

[4] G. Ahanger and T. Little, "A survey of technologies for parsing and indexing digital video," *Journal, Visual Communication and Image Representation*, vol. 7, no. 1, pp. 28–43, 1996.

[5] N. V. Patel and I. K. Sethi, "Video shot detection and characterization for video databases," *Pattern Recognition*, vol. 30, no. 4, pp. 583–592, April 1997.

[6] S. Tsekeridou and I. Pitas, "Content-based video parsing and indexing based on audio-visual interaction," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 11, no. 4, pp. 522–535, 2001.

[7] A. Hanjalic, "Shot-boundary detection: Unraveled and resolved?" *IEEE Trans. Circuits and Systems for Video Technology*, vol. 12 no.2, pp. 90–105, 2002.

[8] C.-L. Huang and B.-Y. Liao, "A robust scene-change detection method for video segmentation," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 11 no.12, pp. 1281–1288, 2001.

[9] R. Lienhart, "Comparison of automatic shot boundary detection algorithms," in *Proc. SPIE Storage and Retrieval for Image and Video Databases VII*, vol. 3656, San Jose, CA, U.S.A. January 1999, pp. 290–301.

[10] M. S. Drew, Z.-N. Li, and X. Zhong, "Video dissolve and wipe detection via spatio-temporal images of chromatic histogram differences," in *Proc. 2000 IEEE Int. Conf. Image Processing*, vol. 3, 2000, pp. 929–932.

[11] R. Lienhart, "Reliable dissolve detection," in *Proc. SPIE Storage and Retrieval for Media Databases 2001*, vol. 4315, January 2001, pp. 219–230.

[12] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis using both audio and visual clues," *IEEE Signal Processing Magazine*, vol. 17, no. 6, pp. 12–36, November 2000.

[13] A. Hampapur, R. Jain, and T. Weymouth, "Digital video segmentation," in *Proc. ACM Multimedia 94, San Francisco, CA*, October 1994, pp. 357–364.

[14] Z.-N. Li and J. Wei, "Spatio-temporal joint probability images for video segmentation," in *Proc. 2000 IEEE Int. Conf. on Image Processing*, vol. 2, 10-13 Sept. 2000, pp. 295 – 29.

[15] Z. Cernekova, C. Nikou, and I. Pitas, "Shot detection in video sequences using entropy-based metrics," in *Proc. 2002 IEEE Int. Conf. Image Processing*, vol. 3, Rochester, N.Y., USA, 22-25 September 2002, pp. III–421 – III–424.

[16] T. Butz and J. Thiran, "Shot boundary detection with mutual information," in *Proc. 2001 IEEE Int. Conf. Image Processing, Greece*, vol. 3, October 2001, pp. 422–425.

[17] W. Cheng, Y. Liu, and D. Xu, "Shot boundary detection based on the knowledge of information theory," in *Proc. 2003 IEEE Int. Conf. Neural Networks and Signal Processing*, vol. 2, 14-17 Dec. 2003, pp. 1237 – 1241.

[18] "Trec video retrieval evaluation," 2004. [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[19] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. 8, pp. 283–298, 1978.