AIIA shot boundary detection at TRECVID 2006

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

***Abstract*— In this paper, we describe the Artificial Intelligence and Information Analysis (AIIA) laboratory approach for shot boundary detection as applied to the TRECVID 2006 video retrieval benchmark. The paper describes the approach as well as the performance analysis. The method relies on evaluating mutual information between multiple pairs of frames within a certain temporal window. The performance of the method on the benchmark data was in general very satisfactory.**

I. Introduction

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 [1], [2]. In [3] mutual information was also used for the detection of gradual transitions. Our approach for automated shot boundary detection applied in TRECVID 2006 is based on the utilization of information from multiple pairs of frames within a temporal window W which ensures effective detection of gradual transitions in addition to abrupt cut detection. This method is described in detail in [4]. The experimental results indeed prove that the method is capable of detecting both abrupt cuts and gradual transitions with good precision and recall rates. AIIA has submitted 10 runs to the Shot Boundary Detection task. All runs use the same algorithm with different threshold settings.

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

II. Shot detection algorithm

When developing the method we focused on the detection of gradual transitions such as dissolves and wipes, which are the most difficult to be detected. Unlike the abrupt cuts, a gradual transition spreads across a number of frames, therefore in order to capture 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 can be 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 between two frames we chose mutual information (MI) since, as shown in our previous research, it provides very good results for abrupt cut detection [1] 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 [1]:

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

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 shots, we do not have to calculate the mutual information between all pairs of frames in a video sequence, because relations between frames, which are far apart are not important for the shot cut detection task. Thus, in our method we use only mutual information calculated between frames in a sliding temporal window W.

More specifically, 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 calculate the mutual information between $(N_W/2)^2$ pairs of frames shown in Figure 1, namely, between all pairs of frames $\mathbf{f}_k$, $\mathbf{f}_l$ where $k \leq i$, $l > i$. Then a cumulative measure which combines information from all these frame pairs 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. The terms of the sum provide information on whether frames

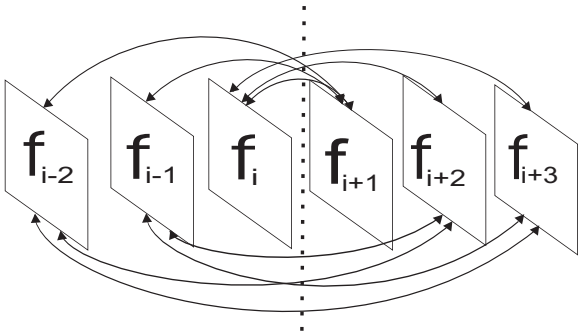

Fig. 1. Diagram showing pairs of frames, which contribute to $I_{cumm}(i)$ for a window size of $N_W = 6$.

$\mathbf{f}_i$, $\mathbf{f}_{i+1}$ are within a shot, belong to a transition or are the last and first frame of two shots separated by abrupt cut. The procedure is repeated for the whole video sequence by sliding the window over all frames.

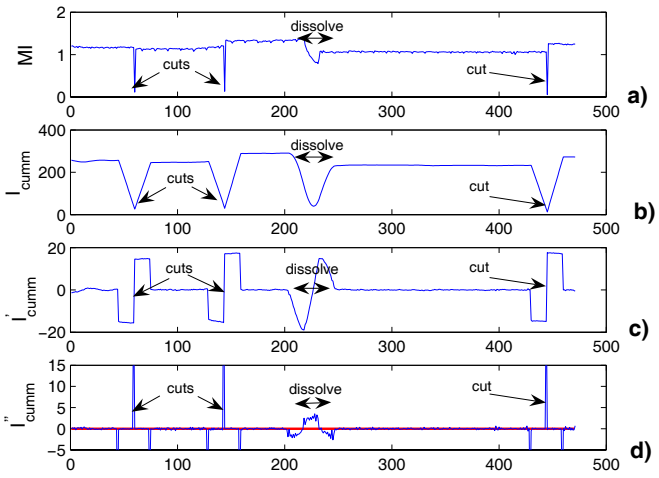

Fig. 2. Plot of mutual information patterns for a part of video sequence that correspond to a dissolve: a) mutual information between two consecutive frames, b) cumulative mutual information I_{cumm} calculated within a window, c) first derivative of I_{cumm} and d) second derivative of I_{cumm}.

The use of multiple pairs of frames aims at overcoming the difficulties caused by the use of pairs of consecutive frames when trying to detect gradual transition. Indeed, as can be seen in Figures 2(a) and 3(a) mutual information calculated between two consecutive frames provides in the case of abrupt cuts easily detectable peaks, whereas for gradual transitions like dissolves does 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. 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 [3] 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 3(a) one can see that a dissolve does not always produce this pattern. Since I_{cumm} contain 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 Figures 2(b) and 3(b) 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. This is also obvious from Figures 2(d) and 3(d), where the second derivative of the same part of video sequence is drawn.

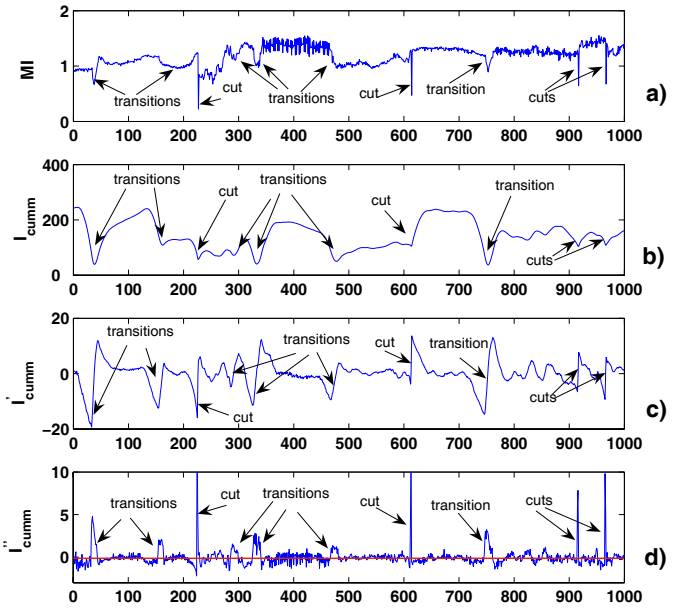

Fig. 3. Plot of mutual information patterns for a part of video sequence: a) mutual information between two consecutive frames, b) cumulative mutual information I_{cumm} calculated within a window, c) first derivative of I_{cumm} and d) second derivative of I_{cumm}.

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. But 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 a threshold T (Figure 4).

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

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 if the corresponding points have positive values of the second derivative, in order to locate the local minima of I_{cumm}.

After identifying the minima of the I_{cumm}, we search around the minimum for the start t_s and end t_e time instant of the transition (transition boundaries). 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.

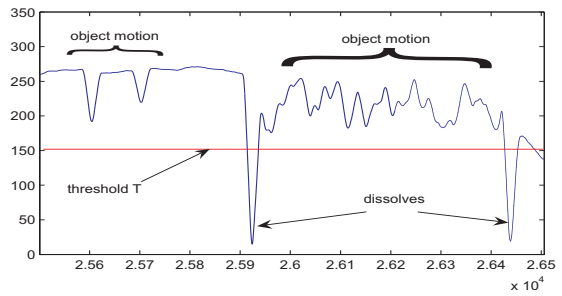

Fig. 4. Plot of I_{cumm} pattern for a part of a video sequence with the threshold T used to separate the local minima caused by the motion within the shot and the local minima caused by the transitions.

Since, a gradual transition has a certain duration, at least three frames should be involved to declare a gradual transition:

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

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

As was shown in [1] mutual information is much less sensitive to camera flashes comparing to histogram comparisons. Therefore, our approach has a big advantage in cases where camera flashes are present in a video sequence. Moreover, since I_{cumm} is evaluated on the basis 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).

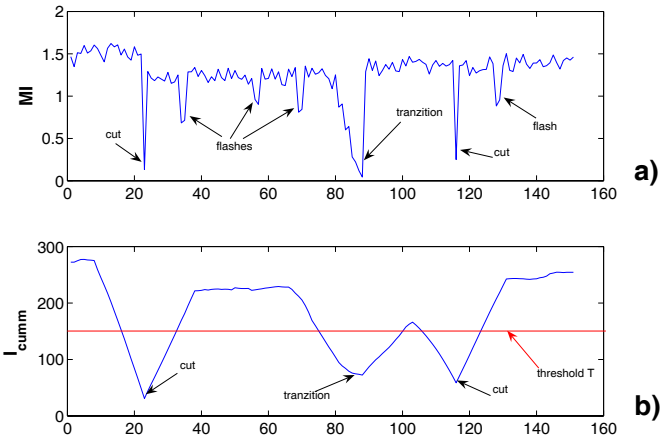

Fig. 5. *A part of video sequence containing many camera flashes: a) mutual information between two consecutive frames and b) the pattern of cumulative mutual information I_{cumm}.*

III. Experimental results

The method was tested on the reference video test set TRECVID 2006 [5] containing newscasts having many commercials in-between. Both news and commercials are characterized by significant camera effects like zoom-ins/outs, pans, and significant object and camera motion inside one shots. Video sequences 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 detection performance for cuts and gradual transitions was measured by precision and recall. Accuracy for reference gradual transitions successfully detected are measured using frame-based precision and recall.

A temporal window W of size $N_W = 30$ has been used in the tests. Previous research showed that for small values of N_W some gradual transitions might not be captured very well whereas for big values of N_W the computation of I_{cumm} becomes more time consuming without adding any information. Moreover, if the window size is very big, it might enclose two transitions.

In the 10 rounds we have submitted we were experimenting with the threshold T in (4). We tested values in the range $[100 \ldots 220]$. For high values of T we have got high false acceptance whereas for the low values of T we have a lot of missdetections. The best performance seems to be for the threshold $T = 150$.

In Figure 6 the recall and the precision for all transitions obtained in TRECVID 2006 for all 26 participants is shown. The Figure is zoomed to values $0.6 \ldots 1.0$ for the recall and the precision. The results shows that the performance of our method is threshold dependent. The best performance of our method is among the best 30% of all submissions.

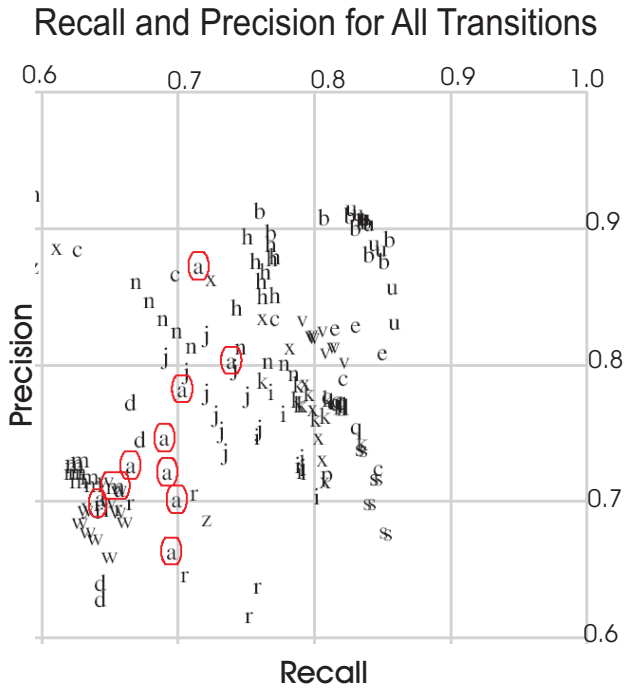

Fig. 6. *Recall and precision obtained by our system (encircled points denoted by a) for all transitions on the TRECVID 2006 shot boundary detection task.*

In Figure 7 the recall and the precision for cuts obtained by our method in TRECVID 2006 is shown along with the results of the other groups submissions. The Figure is zoomed

TABLE I

AIIA LABORATORY SHOT DETECTION RESULTS ON TRECVID 2006.

		All		CUTS		GRADUAL			
								Frame	
Tail of sysid	Plot char.	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
AI10	**a**	**0.715**	**0.872**	**0.764**	**0.882**	**0.584**	**0.836**	0.766	0.688
AI1	**a**	0.699	0.701	0.749	0.704	0.563	0.693	0.674	0.833
AI2	**a**	0.695	0.663	0.704	0.685	0.673	0.608	0.724	0.744
AI3	**a**	0.657	0.709	0.690	0.716	0.568	0.685	0.688	0.763
AI4	**a**	0.692	0.721	0.712	0.735	0.637	0.681	0.714	0.732
AI5	**a**	0.665	0.726	0.665	0.755	0.665	0.659	0.752	0.673
AI6	**a**	0.703	0.782	0.720	0.862	0.658	0.614	**0.732**	**0.878**
AI7	**a**	**0.739**	**0.803**	**0.765**	**0.863**	**0.669**	**0.662**	**0.734**	**0.858**
AI8	**a**	0.690	0.746	0.684	0.831	0.704	0.588	0.722	0.860
AI9	**a**	0.643	0.702	0.612	0.835	0.725	0.515	0.737	0.821

to values $0.6 \ldots 1.0$ for the recall and precision. Results are not as good as the ones we obtained in TRECVID 2004 where the mutual information between two frames was used for detecting cuts [1].

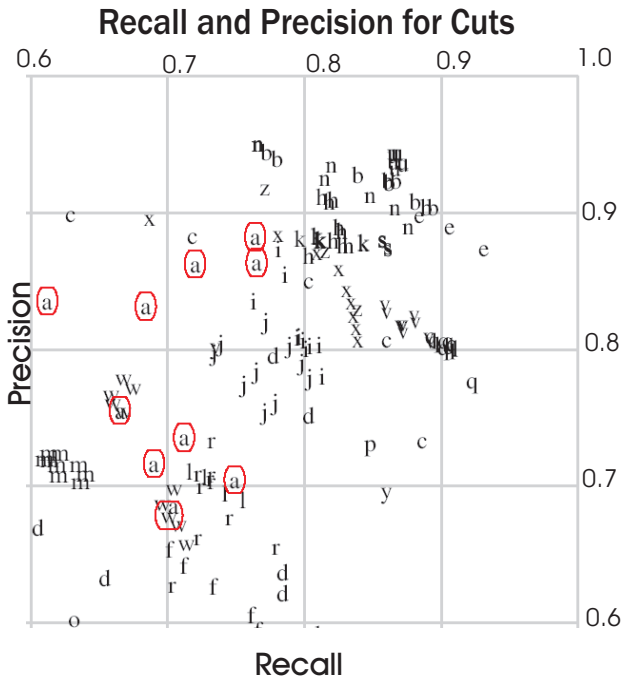

Fig. 7. *Recall and precision obtained by our system (encircled points denoted by a) for cuts on the TRECVID 2006 shot boundary detection task.*

In Figure 8 the recall and the precision for gradual transitions for all 26 participants is shown. The Figure is zoomed to values $0.5 \ldots 0.9$ for the recall and precision.

The FrameRecall and FramePrecision for the correctly detected gradual transitions is shown in Figure 9. The Figure is zoomed to values $0.6 \ldots 1.0$ for the recall and precision. The accuracy for the successfully detected gradual transitions is very good. It outperformed the results which we obtained in TRECVID 2004, where a combination of SVD and clustering method was used for detecting gradual transitions [6]. This is due to the fact that while developing the algorithm presented

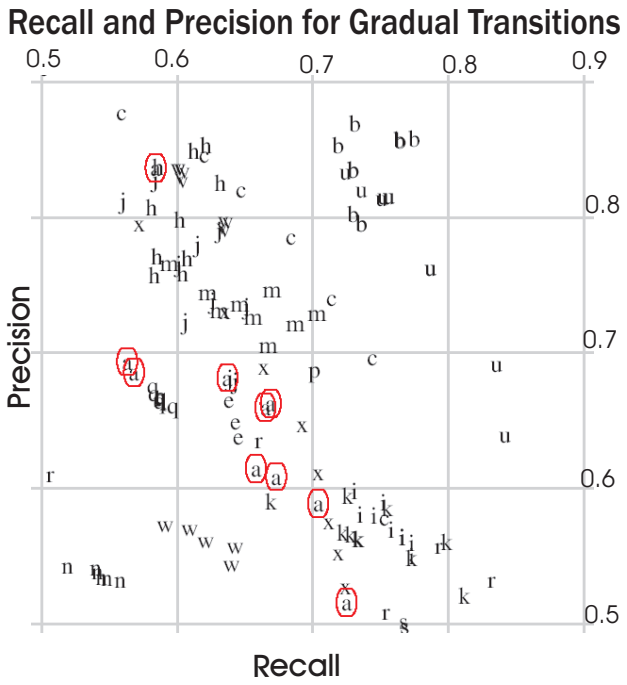

Fig. 8. *Recall and precision obtained by our system (encircled points denoted by a) for gradual transitions on the TRECVID 2006 shot boundary detection task.*

in this paper we focused on better capturing the duration of the transition by taking into account all frames within a certain temporal window W.

In Table I one can see in detail the recall and precision rates obtained by our method for all transitions, cuts, gradual transitions, as well as the FrameRecall and FramePrecision for the gradual transitions. The best two performance values obtained on each category are marked in bold.

IV. CONCLUSIONS AND DISCUSSION

We experimented with a new technique for automated shot boundary detection in video sequences in the context of TRECVID 2006. The method relies on evaluating mutual information between multiple pair of frames within a certain

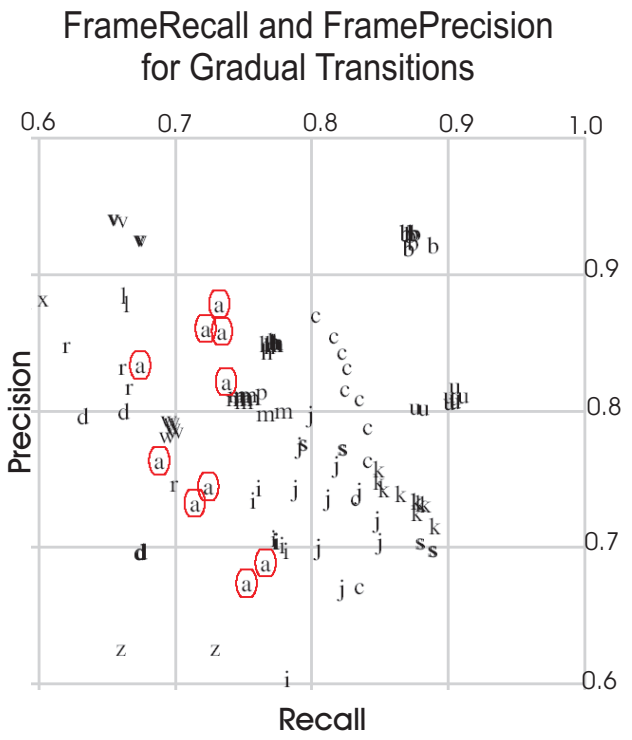

Fig. 9. *FrameRecall and FramePrecision obtained by our system (encircled points denoted by a) for correctly detected gradual transitions on the TRECVID 2006 shot boundary detection task.*

temporal frame window. The method is able to detect efficiently abrupt cuts and all types of gradual transitions, such as dissolves, fades and wipes with good accuracy. We have observed very good accuracy of the gradual transitions which were successfully detected. The new method outperformed the results which we obtained in TRECVID 2004 for the gradual transitions using the method described in [6]. On the contrary the results obtained for the cuts were not as good as the ones we obtained in TRECVID 2004.

ACKNOWLEDGEMENT

This work has been supported by the FP6 European Union Network of Excellence MUSCLE "Multimedia Understanding Through Semantic Computation and Learning" (FP6-507752)

REFERENCES

[1] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 1, pp. 82– 91, January 2006.

[2] 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.

[3] 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.

[4] Z. Cernekova, N. Nikolaidis, and I. Pitas, "Temporal video segmentation by graph partitioning," in *Proc. 2006 IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 14-19, 2006.

[5] "Trec video retrieval evaluation," 2006. [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[6] Z. Cernekova, C. Kotropoulos, and I. Pitas, "Video shot segmentation using singular value decomposition," in *Proc. 2003 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 3, Hong Kong, 6-10 April 2003, pp. 181–184.