Video shot segmentation using fusion of SVD and mutual information features

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

Abstract— **A new method for detecting shot boundaries in video sequences by fusing features obtained by singular value decomposition (SVD) and mutual information (MI) is proposed. The first method relies on performing singular value decomposition on a matrix created from 3D color histograms of single frames. The method can detect cuts and gradual transitions, such as dissolves, fades and wipes. The second method relies on evaluating mutual information between two consecutive frames. It can detect abrupt cuts, fade-ins and fade-outs with very high accuracy. Combination of features derived from these methods and subsequent processing through a clustering procedure results in very efficient detection of abrupt cuts and gradual transitions, as demonstrated by experiments on TRECVID2004 video test set containing different types of shots with significant object and camera motion inside the shots.**

I. Introduction

Indexing and retrieval of digital video is a very active research area. Shot boundary detection is the first step towards further analysis of the video content for indexing, shot classification, browsing, searching and summarization [1].

Early work on shot detection mainly focused on abrupt cuts. A comparison of existing methods is presented in [2], [3]. In some early approaches a cut is 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. Noticing the weakness (hight sensitivity to the object and camera motions) of pixel differencing methods, many researchers suggested the use of different measures based on global information such as intensity histograms or color histograms [3], [4], [5]. The use of more complex features, such as image edges or motion vectors [6], improves the situation, but does not solve the problem completely [7].

Detection of gradual transitions, such as dissolves, fade-ins, fade-outs, and wipes is examined in [8], [9], [10]. 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. Gradual transitions are generally more difficult to be detected than abrupt cuts. Existing techniques for detecting transitions rely on twin thresholding [1] or gray level statistics [2] and have a relatively high false detection rate. Transitions between shots are widely used in TV.

In this paper, we develop a method for automated shot boundary detection using fusion of singular value decomposition and mutual information features. The first method relies on performing singular value decomposition on a matrix created by the 3D color histograms of single frames [11]. The second method relies on the mutual information between two consecutive frames [12]. Fused (concatenated) features from these two methods are next processed by a clustering method in order to detect the shot boundaries.

The remainder of the paper is organized as follows: In Section 2, a description of the features as well as their fusion are presented. The description of the shot detection method is addressed in Section 3. Experimental results are presented and commented in Section 4 and conclusions are drawn in Section 5.

II. Feature level fusion

The developed method is based on feature level fusion. Features are obtained from two shot boundary detection methods. The first method is based on singular value decomposition and can detect abrupt cuts and all types of gradual transitions. We have used SVD for its capabilities to derive a low dimensional refined feature space from a high dimensional raw feature space, where pattern similarity can be easily detected.

The SVD of an $M \times N$ matrix $\mathbf{A}$ is any factorization of the form $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ is an $M \times R$ column-orthogonal matrix, $\mathbf{V}$ is an $N \times R$ column orthogonal matrix, and $\mathbf{\Sigma} = diag(\sigma_1, ..., \sigma_R)$ is a diagonal matrix with non-negative elements, $\sigma_1 \geq ... \geq \sigma_R \geq 0$ for $R = \min(M, N)$.

In our case the column $\mathbf{a}_i$ of the matrix $\mathbf{A}$ corresponds to the three-dimensional normalized histogram in the RGB color space of frame f_i. We have selected 16 bins, for each of the R, G, B color components. Thus, each frame is described by a column vector of $\mathbf{A}$ having dimensions $M \times 1$, where $M = 16^3 = 4096$. This is the raw feature vector corresponding to this frame. More details can be found in [11].

After performing SVD we preserved only the 10 largest singular values of Σ. Let us denote the resulting matrix Σ_{10}. Then, we calculated for every frame a 10-dimensional feature vector as follows

$$\widetilde{\mathbf{v}}_i^T = \mathbf{v}_i^T \mathbf{\Sigma}_{10}, \quad i = 1, 2, \ldots, N \qquad (1)$$

where $\mathbf{v}_i^T$ is the i-th row vector of $\mathbf{V}$. Therefore, each column vector $\mathbf{a}_i$ in $\mathbf{A}$ is mapped to a row vector $\widetilde{\mathbf{v}}_i^T$.

The truncated feature space removes noise and trivial variations in the video sequence. Frames with similar color distribution patterns will be mapped close to each other. Thus, clustering of visually similar frames in the refined feature space will yield better results than in the raw feature space.

The second method uses the mutual information measure. Mutual information is a measure of information transported from one frame to another. It has been shown to provide very good results, because it exploits the inter-frame information flow in a more compact way than frame subtraction. It can be used for detecting abrupt cuts, where the image intensity or color changes abruptly. The mutual information value I_i for each two consecutive frames was calculated (as presented in [12]) separately for R, G and B color components. For example, I_i^R for the R component is expressed by:

$$I_i^R = -\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} \mathbf{C}_i^R(x,y)\log\frac{\mathbf{C}_i^R(x,y)}{\mathbf{C}_i^R(x)\mathbf{C}_i^R(y)}. \qquad (2)$$

The element $\mathbf{C}_i^R(x,y)$, corresponds to the probability that a pixel with gray level x in frame f_i has gray level y in frame f_{i+1}. The total mutual information has been defined as $I_i = I_i^R + I_i^G + I_i^B$.

For every frame f_i, a normalized 11-dimensional feature vector $\overline{\mathbf{w}}_i$ was created, by adding to the 10-dimensional feature vector $\tilde{\mathbf{v}}_i^T$ obtained by applying SVD on frame f_i one additional dimension that corresponds to the mutual information value I_i between consecutive frames f_i and f_{i+1}:

$$\overline{\mathbf{w}}_i = [\tilde{\mathbf{v}}_i I_i]^T \quad i = 1, 2, \ldots, N-1 \qquad (3)$$

III. Shot Detection

In order to detect the video shots, the feature vectors are processed using a dynamic clustering method. The frames are grouped into clusters. Then, consecutive clusters are tested for a possible merging.

At fist, frames are clustered into L clusters, $\{\mathcal{C}_i\}_{i=1}^{L}$, by comparing the following cosine similarity measure between consecutive frames to a threshold δ as is described in [11].

$$\Phi(\overline{\mathbf{w}}_i, \overline{\mathbf{w}}_j) = \cos(\overline{\mathbf{w}}_i, \overline{\mathbf{w}}_j) = \frac{(\overline{\mathbf{w}}_i^T \cdot \overline{\mathbf{w}}_j)}{\|\overline{\mathbf{w}}_i\|\|\overline{\mathbf{w}}_j\|} \qquad (4)$$

Shot detection using clustering on the feature space is justified by the fact that static shots with small camera and object movements are projected to clusters with small dispersion, while shots with some action inside the shot are projected to clusters with a large dispersion. Frames corresponding to transitions between two shots form paths between two dense clusters of points in the feature space. Therefore, we can easily distinguish between transitions and shots with high camera and object movements inside them. Accordingly, from the obtained clusters, the dense ones are identified and associated to shots. An example of a dissolve pattern in the feature space is shown in Figure 1.

Due to the fixed threshold used for clustering, it happens that several shots are split into different clusters. False detections occur for shots with significant motion, because the corresponding clusters are more spread and a fixed threshold cannot preserve all frames in the same cluster. To avoid false detections, the clusters obtained by the above procedure are tested for a possible merger. Merging is performed in two steps.

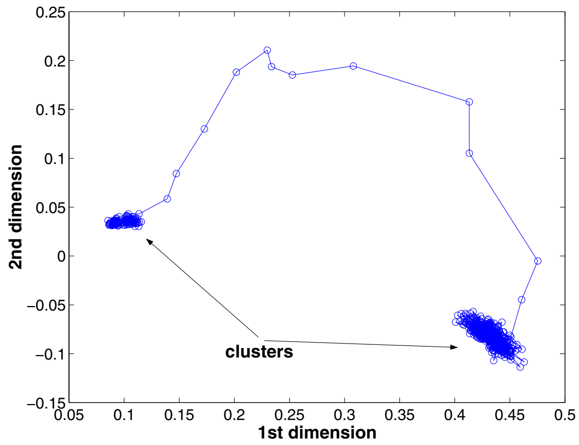

Fig. 1. Plot of the first and second dimension of the feature vector depicting the dissolve pattern between two shots (the dense clusters) in a video sequence.

A. Heuristic cluster merging

The first cluster merging step is based on the fact that if a cluster was erroneously split in two ($\mathcal{C}_k$ and $\mathcal{C}_{k+1}$), the cosine measure between the last frame in cluster $\mathcal{C}_k$ and the first frame in cluster $\mathcal{C}_{k+1}$ is comparable to the cosine measures within the clusters.

Let us denote by f_j^i the j-th frame of the i-th cluster and by $\overline{\mathbf{w}}_j^i$ the corresponding feature vector. For each cluster $\mathcal{C}_k$, we calculate the mean cosine measure $\overline{\phi}_k$ as follows

$$\overline{\phi}_k = \frac{1}{n_k - 1}\sum_{i=1}^{n_k-1} \Phi(\overline{\mathbf{w}}_i^k, \overline{\mathbf{w}}_{i+1}^k) \qquad (5)$$

where n_k is number of frames assigned to the cluster $\mathcal{C}_k$. Then we evaluate the validity of the following condition that involves the mean cosine measures $\overline{\phi}_k$ and $\overline{\phi}_{k+1}$ of two consecutive clusters $\mathcal{C}_k$ and $\mathcal{C}_{k+1}$ and the cosine measure between the last frame $f_{n_k}^k$ in cluster $\mathcal{C}_k$ and the first frame f_1^{k+1} in cluster $\mathcal{C}_{k+1}$:

$$\Phi(\overline{\mathbf{w}}_{n_k}^k, \overline{\mathbf{w}}_1^{k+1}) < \alpha \cdot \overline{\phi}_k \quad \& \quad \Phi(\overline{\mathbf{w}}_{n_k}^k, \overline{\mathbf{w}}_1^{k+1}) < \alpha \cdot \overline{\phi}_{k+1} \quad (6)$$

where α is a constant.

If (6) is fulfilled the clusters $\mathcal{C}_k$ and $\mathcal{C}_{k+1}$ are merged together. Otherwise, the cluster $\mathcal{C}_{k+1}$ is tested for a possible merging with $\mathcal{C}_{k+2}$ and so on.

B. Cluster merging based on statistical hypothesis testing

Let us consider a random sample $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_n$ that is composed of normalized 11-dimensional feature vectors. These normalized vectors can be considered as random directions in K dimensions and can be viewed as a points on the surface of a K-dimensional sphere S_K of unit radius around the origin. The *sample mean vector* is defined by [13], [14]

$$\bar{\boldsymbol{l}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{l}_i \tag{7}$$

and its *direction* is given by

$$\bar{\bar{\boldsymbol{l}}} = \frac{\bar{\boldsymbol{l}}}{\bar{R}} \tag{8}$$

where

$$\bar{R} = \sqrt{\bar{\boldsymbol{l}}^T\bar{\boldsymbol{l}}}. \tag{9}$$

$\bar{\bar{\boldsymbol{l}}}$ can be regarded as the mean direction of the sample. The parameter $\bar{R}$ is closely related to the notion of the *spherical variance*. A value of $\bar{R}$ close to 0 implies that the points $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_n$ are uniformly distributed, whereas a value of $\bar{R}$ close to 1 implies that the points are heavily concentrated near $\bar{\bar{\boldsymbol{l}}}$.

Two other terms of interest in our analysis are, the quantity $R = n\bar{R}$ called the *resultant length* and the vector $\boldsymbol{r} = n\bar{\bar{\boldsymbol{l}}}$ known as the *resultant vector*.

We assume that the sample consisting of feature vectors assigned to a cluster $\mathcal{C}_k$ is a random sample from a K-variate von Mises-Fischer distribution with mean direction $\boldsymbol{\mu}$ and concentration parameter κ. The von Mises-Fisher distribution can be considered as the equivalent of the Gaussian distribution for directional data.

Let $\boldsymbol{r}_k$ be the resultant vector for the feature vectors of the k-th cluster and $\boldsymbol{r}_{k+1}$ be the resultant vector for the feature vectors of the (k+1)-th cluster. Let $\boldsymbol{\mu}_0$ be the mean direction after a possible merging of the two clusters

$$\boldsymbol{\mu}_0 = \frac{\bar{\boldsymbol{\mu}}_0}{\left(\bar{\boldsymbol{\mu}}_0^T\bar{\boldsymbol{\mu}}_0\right)^{1/2}}, \quad \text{where} \quad \bar{\boldsymbol{\mu}}_0 = \frac{\mathbf{r}_k + \mathbf{r}_{k+1}}{n_k + n_{k+1}} \tag{10}$$

We propose the following cluster merging approach: we compare the sample mean direction $\bar{\bar{\boldsymbol{l}}}_k$ and $\bar{\bar{\boldsymbol{l}}}_{k+1}$ of two consecutive shot clusters $\mathcal{C}_k$, $\mathcal{C}_{k+1}$ with the mean direction $\boldsymbol{\mu}_0$ of the cluster after merging and we decide to merge them if neither of $\bar{\bar{\boldsymbol{l}}}_k$, $\bar{\bar{\boldsymbol{l}}}_{k+1}$ is significantly different from $\boldsymbol{\mu}_0$. Comparison is formulated as a hypothesis testing problem. More specifically we consider the following hypothesis test for each of the clusters $\mathcal{C}_k$, $\mathcal{C}_{k+1}$.

$$\begin{aligned} H_0 &: \boldsymbol{\mu} = \boldsymbol{\mu}_0 \\ H_1 &: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0. \end{aligned} \tag{11}$$

Let δ be the angle between $\mathbf{r}_k$ and $\boldsymbol{\mu}_0$, then

$$r_k^T\boldsymbol{\mu}_0 = R_k\cos\theta, \tag{12}$$

where R_k is the resultant length of the resultant vector $\mathbf{r}_k$ that corresponds to the k-th cluster. The null hypothesis is accepted if [13], [14]

$$\cos\delta \geq 1 - \frac{(n_k - R_k)F_{K-1,(n_k-1)(K-1);\alpha}}{(n_k-1)R_k}, \tag{13}$$

where $F_{K-1,(n_k-1)(K-1);\alpha}$ is the upper α percentage point of the F-distribution with degrees of freedom $K-1$ and $(n_k - 1)(K-1)$. Merging is performed only if the null hypothesis is accepted for both $\mathcal{C}_k$ and $\mathcal{C}_{k+1}$.

IV. Experimental results

The proposed method was tested on newscasts from the reference video test set TRECVID 2004 [15] having many commercials in-between. Both news and commercials are characterized by significant camera effects like zoom-ins/outs and pans, abrupt camera movement and significant object and camera motion inside single shots. Video sequences of more than 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	1378
fades	117
dissolves	758
others	126

Let GT denote the ground truth, Seg be the segmented (correct and erroneusly) 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 III, the following measures, inspired by receiver operating characteristics in statistical detection theory [2], [16] 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:
$$Recall = \frac{|Seg \bigcap GT|}{|GT|}. \tag{14}$$
- 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:
$$Precision = \frac{|Seg \bigcap GT|}{|Seg|}. \tag{15}$$

At first, we applied the SVD and MI methods separately on the video test set and we performed result-level fusion of the results. An OR operation was used to fuse the results. Results were superior than those obtained when using each method separately. The recall-precision curve is shown on Figure 2.

The second set of experiments involved the feature-level fusion method described in Sections II and III. By adding the mutual information and increasing the dimension of the feature vector the clusters became better separable. Results verify that the proposed feature-level fusion method outperforms both the decision level fusion and the SVD method whose recall-precision curve is also depicted in the same figure (Figure 2).

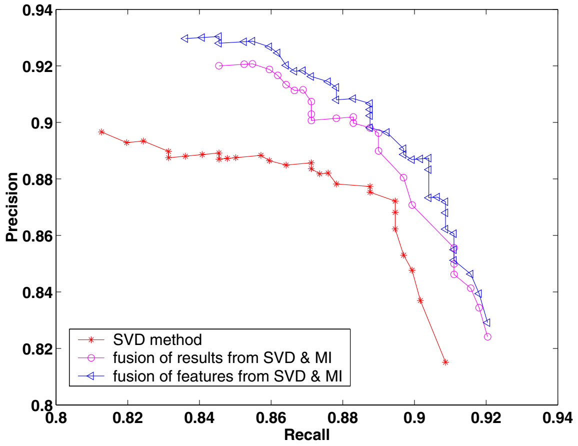

Fig. 2. Recall-precision curves for the three methods tested in the experiments.

Table II summarizes the recall and precision rates obtained by the proposed method for cuts, gradual transitions, as well as for both of them using a threshold $\delta = 0.98$.

TABLE II

SHOT DETECTION RESULTS.

CNN & ABC news	Recall	Precision
cuts	0.95	0.93
gradual transitions	0.86	0.85
overall	**0.91**	**0.89**

A final set of experiments demonstrates the improvements obtained by applying the statistical hypothesis testing cluster merging approach presented in Section III-B. Results are presented in Figure 3.

V. CONCLUSIONS AND DISCUSSION

A new technique for automated shot boundary detection using fusion of singular value decomposition and mutual information features was presented. The method is able to detect efficiently abrupt cuts and all types of gradual transitions. Performance improvements have been achieved by introducing criteria for the statistical merging of clusters. The reported results are very promising.

ACKNOWLEDGMENT

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

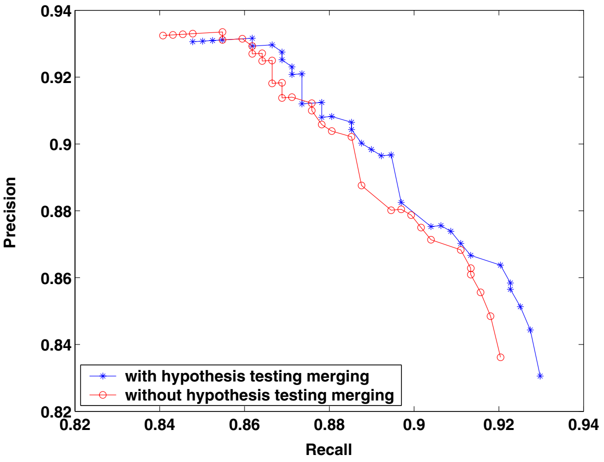

Fig. 3. Recall-precision curves obtained with and without using the statistical hypothesis testing cluster merging step.

REFERENCES

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

[2] 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.

[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] 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.

[6] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] Z. Cernekova, C. Nikou, and I. Pitas, “Shot detection in video sequences using entropy-based metrics,” in *Proc. 2002 IEEE Int. Conf. Image Processing*, Rochester, N.Y., USA, 22-25 September 2002.

[13] K. V. Mardia, *Statistics of Directional Data*. London and New York: Academic Press, 1972.

[14] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis, 2/e*. London, UK: Academic Press, 1980.

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

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