VIDEO SHOT BOUNDARY DETECTION USING SINGULAR VALUE DECOMPOSITION*

Z. ČERNEKOVÁ, C. KOTROPOULOS AND I. PITAS
Aristotle University of Thessaloniki
Box 451, Thessaloniki 541 24, GREECE
E-mail: (zuzana, costas, pitas)@zeus.csd.auth.gr

In this paper we propose a new method for detecting boundaries between single shots in video sequences using singular value decomposition (SVD). The method relies on performing singular value decomposition on the matrix **A**, which columns are reshaped 3D color histograms of the frames. We have used SVD for its capabilities to derive a low dimensional refined feature space from a high dimensional raw feature space, where the similar patterns are placed together and can be easily detected. The detection technique was tested on TV video sequences having different types of shots and significant object and camera motion inside the shots. The method can detect cuts and gradual transitions, such as dissolves and fades, which cannot be detected easily using other methods.

1. Introduction

Shot boundary detection is the first preprocessing step to further analyze the video content for indexing, browsing, searching, summarization, etc. [1].

Early work on shot detection mainly focused on abrupt cuts. A comparison of existing methods is presented in Lienhart [2]. Gradual transitions, such as dissolves, fade-ins, fade-outs, and wipes are examined in Drew *et al.*[3] and Wang *et al.*[4]. These transitions are generally more difficult to be detected, due to camera and object motions within a shot. Their detection is a very powerful tool for shot classification and story summarization.

In a previous work, we used entropy measures for detecting abrupt cuts and fades [5]. In this paper, we develop a method for automated shot boundary detection using singular value decomposition. The method relies on performing singular value decomposition on the matrix **A** created by the

*Whis work has been supported by the European Union Research Training Network "Methods for Unified Multimedia Information Retrieval" (MOUMIR)

3D color histograms of single frames. By using SVD we are able to detect dissolves which were not addressed in our previous work.

2. Singular value decomposition

The singular value decomposition (SVD) is a powerful linear algebra technique. The SVD of an $M \times N$ matrix $\mathbf{A}$ whose number of rows M is greater than or equal to its number of columns N, is any factorization of the form $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ is an $M \times N$ column-orthogonal matrix, $\mathbf{V}$ is an $N \times N$ column orthogonal matrix, and $\mathbf{\Sigma} = diag(\sigma_1, ..., \sigma_R)$ is a diagonal matrix with non-negative elements, with $\sigma_1 \geq ... \geq \sigma_R \geq 0$ and $R = \min(M, N)$. The values σ_i are the **singular values**, whereas the first R columns of $\mathbf{V}$ and $\mathbf{U}$ are called the **right singular vectors** and the **left singular vectors**, respectively.

3. Shot detection

In our approach, we calculated an M-dimensional feature vector $\mathbf{a}_i$ for each frame f_i, $i = 1, 2, \ldots, N$. Using $\mathbf{a}_i$ as a column, we obtained the matrix $\mathbf{A} = [\mathbf{a}_1|...|\mathbf{a}_N]$. As feature vector we chose the color histogram of each frame. More specifically, we calculated the three-dimensional normalized histograms in the RGB color space with 16 bins, for each of the R, G, B color components. Thus, the dimensionality of feature vectors is $M = 16^3 = 4096$.

Using such a feature vector as a column we created the $M \times N$ feature matrix $\mathbf{A}$. Each feature is associated with a row vector of $\mathbf{A}$ of dimensions $1 \times N$ and each frame is described by a column vector of $\mathbf{A}$ of dimensions $M \times 1$. The column vectors of $\mathbf{A}$, that is, the frame color histograms, are projected onto the orthonormal basis formed by vectors of the left singular matrix $\mathbf{U}$. The coordinates of the frames in this space are given by the columns of $\mathbf{\Sigma}\mathbf{V}^T$.

3.1. *Clustering*

Performing SVD we project vectors from the M-dimensional feature space to a K-dimensional ($K << M$) refined feature space, by preserving only the K largest singular values of $\mathbf{\Sigma}$. Let us call the resulting matrix $\mathbf{\Sigma}_K$. Let us denote by $\widetilde{\mathbf{v}}_i = \mathbf{v}_i\mathbf{\Sigma}_K$ the projected frame histograms. Then each column vector $\mathbf{a}_i$ in $\mathbf{A}$ is mapped to a row vector $\widetilde{\mathbf{v}}_i$. The truncated feature space removes the noise or the trivial variations in the video sequence.

The frames with similar color distribution patterns will be mapped close to each other. From an analogy with the SVD-based document clustering and retrieval, clustering of visually similar frames in the refined feature space will certainly yield better results than in the raw feature space.

As a measure of similarity we have defined the angle between the row vectors $\widetilde{\mathbf{v}}_i$ and $\widetilde{\mathbf{v}}_j$:

$$\Phi(f_i, f_j) = \cos(\widetilde{\mathbf{v}}_i, \widetilde{\mathbf{v}}_j) = \frac{(\widetilde{\mathbf{v}}_i \cdot \widetilde{\mathbf{v}}_j^T)}{\|\widetilde{\mathbf{v}}_i\| \|\widetilde{\mathbf{v}}_j\|} \tag{1}$$

Using the similarity measure (1) we obtain values in the range $[0, 1]$, where 1 stays for identical frames. The more different the vectors are, a closer value to 0 is obtained.

To detect shots we are using a dynamic clustering method. The frames are clustered into L clusters, $\{c_i\}_{i=1}^{L}$, by comparing their similarity measure (1) to a threshold δ. The clustering algorithm works as follows.

Initialization:

- It refers to the first two frames f_1 and f_2 represented by $\widetilde{\mathbf{v}}_1$ and $\widetilde{\mathbf{v}}_2$. They form the cluster c_1 by definition. The cluster mean is simply

$$\overline{\mathbf{m}}_1 = \frac{1}{2}\{\widetilde{\mathbf{v}}_1 + \widetilde{\mathbf{v}}_2\}. \tag{2}$$

Recursion:

- Frame f_3 is tested if it should be added to c_1 or to become a seed for a new cluster. We test if

$$\Phi(\overline{\mathbf{m}}_1, \widetilde{\mathbf{v}}_3) < \delta \tag{3}$$

  If the inequality (3) is fulfilled then we create a new cluster with mean $\overline{\mathbf{m}}_1 = \widetilde{\mathbf{v}}_3$. Otherwise, we include f_3 into c_1 and update $\overline{\mathbf{m}}_1$ in the following way

$$\overline{\mathbf{upd\_m}}_1 = \overline{\mathbf{m}}_1 + \frac{1}{n_1 + 1} d \tag{4}$$

  where a quantity d is given by $d = \widetilde{\mathbf{v}}_3 - \overline{\mathbf{m}}_1$ and n_1 is number of elements in cluster c_1.
- When frame f_l is to be processed, we are interested in testing if f_l is to be included into the last cluster formed chronologically up to the l-th time instant. Let us denote by c_j this cluster. Then, we test if

$$\Phi(\overline{\mathbf{m}}_j, \widetilde{\mathbf{v}}_l) < \delta. \tag{5}$$

In case (5) is satisfied we create a new cluster c_{j+1} represented by $\overline{\mathbf{m}}_{j+1} = \widetilde{\mathbf{v}}_l$. Otherwise $\overline{\mathbf{m}}_j$ is updated after inclusion of $\widetilde{\mathbf{v}}_l$

$$\overline{\mathbf{upd\_m}}_j = \overline{\mathbf{m}}_j + \frac{1}{n_j + 1} d \tag{6}$$

where a quantity d is given by $d = \widetilde{\mathbf{v}}_l - \overline{\mathbf{m}}_j$ and n_j is number of elements in cluster c_j.

The sparse clusters usually show the transition between the shots. Accordingly, from the obtained clusters, the dense ones are identified and associated to shots.

Homogenity of the obtained clusters is also examined based on properties of covariance matrix of the clusters

$$S_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (\widetilde{\mathbf{v}}_i - \overline{\mathbf{m}}_j)(\widetilde{\mathbf{v}}_i - \overline{\mathbf{m}}_j)' \tag{7}$$

After merging the frame with tested cluster the covariance matrix is updated as follows:

$$upd\_S_j = S_j + \frac{1}{n_j + 1} dd' \tag{8}$$

where d is $d = \widetilde{\mathbf{v}}_l - \overline{\mathbf{m}}_j$.

4. Experimental results

The proposed method was tested on several real TV sequences having many commercials in-between, characterized by significant camera effects.

Let GT denote the ground truth, Seg be the segmented (correct and false) shots using our method and $|E|$ be the number of elements (frames) of a set E. In order to evaluate the performance of the segmentation method presented in Section 3, the following measures, inspired by receiver operating characteristics in statistical detection theory [2,6] were used:

$$Recall = \frac{|Seg \bigcap GT|}{|GT|} \quad Precision = \frac{|Seg \bigcap GT|}{|Seg|} \tag{9}$$

We have tested the method with several choices of K. Based on experiments the best choice was shown to be $K = 10$. For this value the clusters are separated well. By increasing K the data become more messy. The static shots with small camera and object movements inside the shot are projected to data with small dispersion, while shots with some action inside the shot are projected to data with a large dispersion. Table 1 summarizes

Table 1. Shot detection results using a fixed threshold.

video	cuts		gradual trans.		global	
	Recall	Precision	Recall	Precision	Recall	Precision
basketball	0.95	0.98	0.85	1.00	0.93	0.98
news	0.93	0.90	1.00	1.00	0.96	0.94
teste	1.00	1.00	1.00	1.00	1.00	1.00
football	0.96	0.96	1.00	1.00	0.97	0.97

the recall and precision rates measured for cuts, gradual transitions, as well as for both of them using $\delta = 0.9$ and $K = 10$.

By varying $\delta \in [0.75, 0.98]$ we have obtained the recall- precision curve shown in Figure 1.

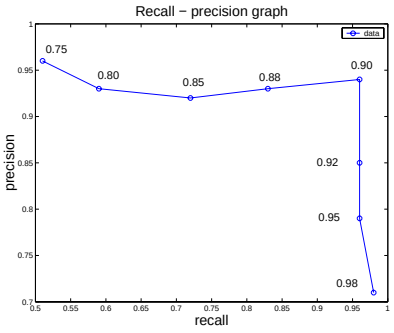

Figure 1. Recall-precision curve by varying the thresh. δ.

The transition between two shots is shown as a path between two dense clusters of points in the projected space in Figure 2. Therefore, we can easily detect the transition.

We can also distinguish between transitions and shots with high camera and object movements inside them. If we used any method based on histogram comparisons, which are the most common, we would hardly identify movements inside a shot and transitions between two shots. Most of the time, the motion inside a shot gives rise to false alarms. Using SVD, we can distinguish between the motion and transition and avoid the false alarms.

5. Conclusions and discussion

A new technique for automated shot transitions detection using singular value decomposition was presented. In the clustering phase two measures of cluster homogenity were examined, the mean value and covariance matrix

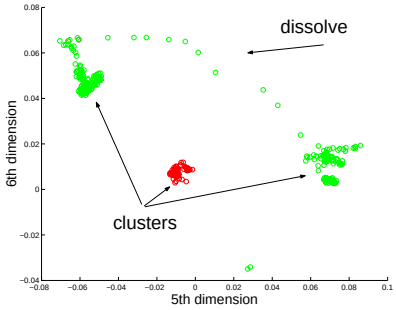

Figure 2. By using the fifth and sixth dimension of the projected frame histograms we can successfully detect the dissolve in sequence "teste" .

of the cluster. The method is able to detect well the dissolves which we did not address in the previous work with the entropy measures. The reported results are promising. However, using a fixed threshold in the dynamic clustering method yields some deficiencies. False detections occur for shots with big motion, because the clusters are more spread and a fixed threshold cannot preserve all frames in the same cluster. We intend to improve the clustering method by employing an adaptive threshold which would vary based on the density of the frames in the space and by introducing criteria for a possible merging of clusters.

References

1. A. Del Bimbo, *Visual Information Retrieval*, Morgan Kaufmann Publishers, Inc, San Francisco, California, 1999.
2. R. Lienhart, "Comparison of automatic shot boundary detection algorithms," in *Proc. of SPIE Storage and Retrieval for Image and Video Databases VII, San Jose, CA, U.S.A.*, January 1999, vol. 3656, pp. 290–301.
3. 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. on Image Processing*, 2000, vol. 3, pp. 929–932.
4. 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.
5. 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.
6. C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. 8, pp. 283–298, 1978.