Video Shot Segmentation and Classification

Yihong Gong, and Xin Liu
C&C Research Laboratories, NEC USA, Inc.
110 Rio Robles, San Jose, CA 95134, U.S.A.
Email: {ygong,xliu}@ccrl.sj.nec.com

Abstract

In this paper, we propose a novel technique for video shot segmentation and classification based on the Singular Value Decomposition (SVD). For the input video sequence, we create a feature-frame matrix $\mathbf{A}$*, and perform the SVD on it. From this SVD, we are able to not only derive the refined feature space to better segment the video sequence along time axis, but also define metrics to enable classifications of the detected video shots. Using these SVD properties, we achieve the two goals of accurate video shot segmentation, and visual content-based shot classification at the same time.*

1 Introduction

The wide spread distribution of video in computer systems and networks has presented both excitements and challenges. Video is exciting because it conveys real-world scenes most vividly and faithfully. Handling video is challenging because video images are voluminous, redundant, and unstructured. With a large video data collection, it is always a painful task to find either the appropriate video sequence, or the desired portions of the video. Traditional text indexing and retrieval techniques have turned out to be powerless in managing video images. To tap into the rich and valuable video resources, video images must be transformed into a medium that is structured, manageable and searchable.

The initial steps toward the above goal include the segmentation of video sequences into shots for indexing and access, and the extraction of features/metadata from the shots to enable their classifications and retrieval. For video shot segmentation, a great number of methods have been proposed in past years. Typical methods include shot segmentation using pixel values [1, 2], global or local histograms [3], motion vectors [3], DCT coefficients from MPEG files [4], etc. While many methods in the literature use the simple approach of frame-pair comparisons and can detect only the abrupt shot boundaries, some methods involve more frames in the comparison to accommodate the detection of gradual scene changes [5]. For video shot retrieval and classification, the most common approach todate is to first carry on the video shot segmentation, perform additional operations to extract features from each detected shot, and then create indexes and metrics using these features to accomplish shot retrieval and classifications. As several processing steps must be performed in tandem, high computational cost and long processing time are usually required by the systems based on this approach.

In this paper, we propose a novel technique for video shot segmentation and classification based on the Singular Value Decomposition (SVD). The system is able to detect shot boundaries with a high accuracy, and to classify the detected shots at the same time. The details of the proposed system and its performance are presented in the following part of this paper.

2 The Proposed System

Our video shot segmentation and classification system uses the SVD as an important basis. The SVD is known for its capabilities of deriving the low dimensional refined feature space from a high dimensional raw feature space, and of capturing the essential structure of a data set in the refined feature space [6]. To reduce the number of frames to be processed by the SVD, we sample the input video sequence with a fixed rate of 10 fps. For each frame i in this sampling set, we create an m-dimensional feature vector A_i. Using A_i as a column, we obtain the feature-frame matrix $\mathbf{A} = [A_1\ A_2\ \cdots\ A_n]$. Performing SVD on this matrix $\mathbf{A}$ will project each frame i from the m-dimensional raw feature space into the κ-dimensional refined feature space (usually $\kappa \ll m$). In this new space, noise and trivial variations in video frames will be ignored, and frames with similar color distribution patterns will be mapped near to each other. Therefore, the κ-dimensional vectors representing each of the frames in the refined feature space can be used not only for accurate shot segmentation, but also for similarity matching among the detected video shots.

Besides the above unique SVD feature, our mathematical analysis has further revealed that in the same refined feature space, there are strong correlations between the degree of visual changes, the evenness of color distributions, in a shot, and the positions at

which its constituent frames are projected, respectively. The degree of visual changes depicts the dynamic level, while the evenness of color distributions reflects the color appearance, of the video shot. Combining these SVD properties, our system is able to efficiently achieve the two goals of accurate video shot segmentation, and visual content-based shot classification at the same time.

2.1 Construction of Feature Vector

From a variety of image features, we selected color histograms to represent video frames. Histograms are very good for detecting overall differences in images, and are cost-effective for computing. In our system, we create three-dimensional histograms in the RGB color space with 5 bins for R,G, and B, respectively, resulting in a total of 125 bins. To incorporate spatial information of the color distribution, we divide each frame into 3×3 blocks, and create a 3D-histogram for each of the blocks. These nine histograms are then concatenated together to form a 1125-dimensional feature vector for the frame. Using the feature vector of frame i as i'th column, we create the feature-frame matrix $\mathbf{A}$ for the video sequence. Since a small image block does not normally contain all kinds of colors, matrix $\mathbf{A}$ is usually sparse. Therefore, SVD algorithms for sparse matrix can be applied here, which is must faster and memory efficient compared to regular SVD algorithms.

2.2 SVD Properties for Shot Segmentation and Matching

Given an $m \times n$ matrix $\mathbf{A} = [a_{ij}]$, where $m \geq n$, the SVD of $\mathbf{A}$ is defined as [7]:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{1}$$

where $\mathbf{U} = [u_{ij}]$ is an $m \times n$ column-orthonormal matrix whose columns are called left singular vectors; $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ is an $n \times n$ diagonal matrix whose diagonal elements are the singular values, and $\mathbf{V} = [v_{ij}]$ is an $n \times n$ orthonormal matrix whose columns are called right singular vectors. If rank($\mathbf{A}$)=r, then $\mathbf{\Sigma}$ satisfies

$$\sigma_1 \geq \sigma_2 \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0. \tag{2}$$

In our system, applying SVD to the feature-frame matrix $\mathbf{A}$ can be interpreted as follows. The SVD derives a mapping between the m-dimensional raw feature space spanned by the concatenated color histograms and the r-dimensional refined feature space with all of its axes linearly-independent. This mapping maps each column vector A_i in matrix $\mathbf{A}$, which represents the concatenated histogram of frame i, to column vector $\psi_i = [v_{i1}\ v_{i2}\ \cdots\ v_{ir}]^T$ of matrix $\mathbf{V}^T$, and maps each row vector j in matrix $\mathbf{A}$, which tells the occurrence count of the concatenated histogram entry j in each of the video frames, to row vector $[u_{j1}\ u_{j2}\ \cdots\ u_{jr}]$ of matrix $\mathbf{U}$.

The SVD has the following property that has been widely utilized for text indexing and retrieval ([7]).

Theorem 1 *Let the SVD of matrix* $\mathbf{A}$ *be given by Eq.(1),* $\mathbf{U} = [U_1 U_2 \cdots U_n]$, $\mathbf{V} = [V_1 V_2 \cdots V_n]$, *and rank(*$\mathbf{A}$*)=*$r$. *Matrix* $\mathbf{A}_\kappa$ *(*$\kappa \leq r$*) defined below is the closest rank-*κ *matrix to* $\mathbf{A}$ *for the Euclidean and Frobenius norms.*

$$\mathbf{A}_\kappa = \sum_{i=1}^{\kappa} U_i \cdot \sigma_i \cdot V_i^T \tag{3}$$

The use of κ-largest singular values to approximate the original matrix with Eq.(3) has significant implications. Discarding small singular values is equivalent to discarding linearly semi-dependent or practically non-essential axes of the feature space. The truncated feature space removes the noise or trivial variations in video frames. Minor differences between histograms will be ignored, and video frames with similar color distribution patterns will be mapped near to each other. From analogy with the SVD-based text clustering and retrieval [6], clustering visually similar frames in this refined feature space will certainly yield better results than in the raw feature space. The value of κ is a design parameter. Our experiments show that $\kappa = 150$ gives satisfactory video segmentation results. The above discussions lead us to define the following similarity metric between frame i and j for shot segmentation and matching:

$$\text{SIM}(i,j) = D(\psi_i, \psi_j) = \sqrt{\sum_{l=1}^{\kappa} \sigma_l (v_{il} - v_{jl})^2} \tag{4}$$

where ψ_i, ψ_j are the vectors representing frames i, j in the refined feature space, respectively, and σ_l's are the singular values from the SVD.

2.3 SVD Properties for Video Classification

Besides the above SVD features, we have further discovered the following SVD properties which constitutes the basis of our video classification (The proof is omitted due to the page limit).

Theorem 2 *Let the SVD of* $\mathbf{A}$ *be given by Eq.(1),* $\mathbf{A} = [A_1 \cdots A_i \cdots A_n]$, $\mathbf{V}^T = [\psi_1 \cdots \psi_i \cdots \psi_n]$. *Using the notation associated with Eq.(1), we have* $A_i = [a_{1i}\ a_{2i}\ \cdots\ a_{mi}]^T$, *and* $\psi_i = [v_{i1}\ v_{i2}\ \cdots\ v_{in}]^T$.

(1) *Define the length of* ψ_i *as:*

$$||\psi_i|| = \sqrt{\sum_{j=1}^{rank(\mathbf{A})} v_{ij}^2}\ . \tag{5}$$

If rank($\mathbf{A}$*)=*n, *from the orthonormal property of matrix* $\mathbf{V}$, *we have* $||\psi_i||^2 = 1$, *where* $i = 1, 2, \ldots, n$. *If* A_i *has* k *duplicates in matrix* $\mathbf{A}$, *Then,* $||\psi_i||^2 = 1/k$.

(2) *Define the singular value weighted length of ψ_i as:*

$$||\psi_i||_\Sigma = \sqrt{\sum_{j=1}^{rank(\mathbf{A})} \sigma_j^2 v_{ij}^2}\ . \qquad (6)$$

Then, $||\psi_i||_\Sigma^2 = A_i \cdot A_i = \sum_{j=1}^{m} a_{ji}^2$.

Translating Property (1) in Theorem 2 into the video domain, it can be inferred that, in the refined feature space, frames in a static video segment (e.g., shots of anchor persons, weather maps) will be projected into the points with shorter length, while frames in a video segment containing a lot of changes (e.g., shots containing moving objects, camera pan and zoom) will be projected into the points with larger length. In other words, by looking at the location at which a shot is projected, we can roughly tell the degree of visual changes of the shot.

On the other hand, Property (2) in Theorem 2 can be used as an indicator of the evenness of color distributions in frames and shots. Because A_i is the concatenated histogram of frame i, the sum of its elements a_{ji} equals a constant value $\sum_{j=1}^{m} a_{ji} = C$ (=the number of pixels in the frame). Hence, $||\psi_i||_\Sigma^2$ reaches the minimum when $a_{1i} = a_{2i} = \cdots = a_{mi}$, while it reaches the maximum when one element $a_{ki} = C$ and the remaining elements all equal zero. In summary, the singular value weighted length $||\psi_i||_\Sigma^2$ is proportional to the evenness of the color distribution of the corresponding frame i. This length becomes the shortest when frame i has a complete even color distribution, and it becomes the longest when frame i consists of only one color.

2.4 Operational Details

Our video segmentation and classification system consists of the following major steps:

Step 1. Sample the input video sequence with a fixed rate of 10 fps, and create the feature-frame matrix $\mathbf{A}$ as described in Section 2.1.

Step 2. Perform the SVD on $\mathbf{A}$ to obtain matrices $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ and $\mathbf{V}^T = [\psi_1 \cdots \psi_i \cdots \psi_n]$.

Step 3. Compute the similarity $\text{SIM}(i, i+1)$ defined by Eq.(4) for all the frames in the sample set, and segment the video sequence into video shots along the time axis (see the following segmentation algorithm for the detail).

Step 4. For each video shot Θ_S, compute the following two average lengths:

$$\overline{||\Theta_S||^2} = \frac{1}{P(\Theta_S)} \cdot \sum_{\psi_i \in \Theta_S} ||\psi_i||^2 \qquad (7)$$

$$\overline{||\Theta_S||_\Sigma^2} = \frac{1}{P(\Theta_S)} \cdot \sum_{\psi_i \in \Theta_S} ||\psi_i||_\Sigma^2 \qquad (8)$$

where $P(\Theta_S)$ is the number of frames included in shot Θ_S. The two values indicate the degree of visual changes, and the evenness of color distributions in shot Θ_S, respectively.

Step 5. Compute the average feature vector $\bar{\Psi}_S$ for each shot Θ_S. Distance $D(\bar{\Psi}_X, \bar{\Psi}_Y)$ defines the visual similarity between shots Θ_X and Θ_Y.

In the above operation, Step 1 and 2 perform the SVD; Step 3 conducts the shot segmentation; and Step 4 and 5 compute the metrics for each detected shot to measure their color distributions, dynamic levels, and visual similarities.

The step of shot segmentation (Step 3) involves two thresholds, T_{low} and T_{high}. If the distance between two consecutive frames is below T_{low}, the two frames will be grouped into the same shot without further examination. If this distance is above T_{high}, shot boundary will be declared. If this distance is between T_{low} and T_{high}, further examination involving more frames will be performed to determine if the large distance is due to the temporary variation, or the gradual scene transition. The operation detail is given as follows:

1. Set shot counter $S = 1$, and frame index $I = 1$.
2. Create shot Θ_S with frame I as its first element.
3. if $D(\psi_I, \psi_{I+1}) \leq T_{low}$, insert frame $I+1$ into shot Θ_S; increment I by one. Repeat this step if I is not the last frame; otherwise, go to 6.
4. if $D(\psi_I, \psi_{I+1}) > T_{high}$, mark the location between frames I and $I+1$ as a shot boundary; increment S and I by one. Go to 2.
5. if $T_{low} < D(\psi_I, \psi_{I+1}) \leq T_{high}$, do the following:
   (a) Find the frame $X > I+1$ that satisfies $D(\psi_X, \psi_{X+1}) \leq T_{low}$.
   (b) If $D(\psi_X, \psi_I) > T_{high}$, mark the frames between $I+1$ and X as a gradual transition between the two scene shots; set $I = X+1$, and increment the shot counter S by one. Go to 6.
   (c) Otherwise, group frames from $I+1$ to X into shot Θ_S; and set $I = X+1$. Go to 3.
6. If the last frame has been reached, terminate the entire operation; otherwise, go to 2.

3 Evaluation and Summary

The proposed shot segmentation and classification system is evaluated using a total of two hour CNN news video programs. The video programs contain almost all possible video edit effects such as abrupt scene changes,

Table 1: Experimental Evaluations

	Abrupt Shot Cut		Gradual Transition		Shot Classification	
	Recall	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**
Local Hist. Method	92.6%	72.1%	-	-	-	-
Proposed Method	97.3%	92.7%	94.3%	87.0%	90.2%	85.1%

fades, wipes, dissolves, etc, and have a great variety of scene categories such as portraits, landscapes, interviews, crowds, moving camera/objects, etc. For comparison, the local histogram-based shot segmentation method was also implemented and tested using the same set of video programs. We chose to compare with the local histogram method because its performance was reported to be one of the best among the existing methods [8]. The experimental results are listed in Table 1. It is observed that for abrupt shot cut detection, the proposed system improves on the recall remarkably, and on the precision dramatically. These improvements are achieved by the frame comparison in the truncated feature space derived from the SVD, and the use of the two thresholds T_{high} and T_{low} that divide the entire domain of the frame distance into the low, gray, and high zones. If the distance between two consecutive frames falls into the gray zone, more frames will be examined to determine if this large distance is due to the presence of noise, jitters from camera/object motions, or the genuine scene change (See Section 2.4). This approach greatly reduces outliers and hence leads to a high recall, high precision rates of the shot boundary detection.

In addition, the proposed system can further detect the gradual scene transitions, and classify the detected shots into the four categories such as identical shots, shots with high degree variations, static shots without remarkable changes, and shots with a uniform color (e.g., black/white frames). In many video programs, it is quite often that the same person or the same scene appears repeatedly (e.g. the anchor persons, the interviewers/interviewees) for a certain time period. Finding these identical scenes serves to detect duplicates and redundancies, which will be critical for generating concise video content summaries. On the other hand, dynamic shots with lots of variations may contain either camera pan and zoom that aim to capture the entire event, or dramatic object motions that come from violent scenes. The ability of identifying dynamic shots contributes to the ultimate goal of detecting visually important scenes. Finally, as the black or white frames often appear around scene shot boundaries, right before or after the commercials, detecting these kinds of frames is useful for many applications. The recall and precision for shot classification are obtained by averaging the recall and precision values of the four shot categories (to save the space). From the table, it is clear that our proposed system has achieved competitive performance for the gradual scene transition detection as well as the shot classification. In contrast to many traditional shot classification systems which rely heavily on heuristic rules and bunch of thresholds, we classify shots based on the metrics derived from the SVD properties. This has led to simple, robust, and accurate classifications of the video shots.

In summary, our SVD-based shot segmentation and classification system has achieved the two goals of accurate video shot segmentation, and visual content-based shot classification at the same time.

References

[1] K. Otsuji, Y. Tonomura, and Y. Ohba, "Video browsing using brightness data," in *SPIE Proc. Visual Communications and Image Processing*, (Boston), pp. 980–989, 1991.

[2] A. Hampapur, R. Jain, and T. Weymouth, "Digital video segmentation," in *Proceedings of ACM Multimedia 94*, (San Francisco), Oct. 1994.

[3] H. Ueda, T. Miyatake, and S. Yoshizawa, "Impact: An interactive natural-motion-picture dedicated multimedia authoring system," in *Proc. ACM SIGCHI'91*, (New Orleans), Apr. 1991.

[4] F. Arman, A. Hsu, and M.-Y. Chiu, "Image processing on encoded video sequences," *Multimedia Systems*, vol. 1, no. 5, pp. 211–219, 1994.

[5] H. Zhang, A. kankanhalli, and S. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Systems*, vol. 1, pp. 10–28, 1993.

[6] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391–407, 1990.

[7] G. Golub and C. Loan, *Matrix Computations*. Baltimore: Johns-Hopkins, 2 ed., 1989.

[8] J. Boreczky and L. Rowe, "Comparison of video shot boundary detection techniques," in *Proceedings of SPIE: Storage and Retrieval for Image and Video Databases IV*, vol. 2670, 1996.