SHOT DETECTION USING PRINCIPAL COORDINATE SYSTEM

Alper YILMAZ Mubarak Ali Shah
University of Central Florida, USA

Abstract

In this study, a new framework for parsing a video sequence into shots is presented. The new algorithm uses eigenspace decomposition of the RGB color space to describe the frames in a more descriptive coordinate system. The new basis eliminates the stumbling blocks of current shot detection approaches. The algorithm executes in real time and gives robust boundaries for shots even when gradual transitions, illumination changes or camera motion occur. The idea is extended to characterization of the scenes without using extra effort. The experiments demonstrate the speed and accuracy in shot detection and characterization.

Keywords: Shot Detection, Eigenspace Decomposition

1. Introduction

The availability of the processing power boosts the research in video based applications, including content analysis, surveillance, story extraction and etc. All these applications require re-organization of the video data into a hierarchically structured data. This hierarchy is defined from top to bottom by story, scene and shot, where each level has one or more entries in lower level. The three levels of the hierarchy from bottom to top can be given as follows: a shot is an unbroken sequence of frames from one camera, a scene is a collection of shots where the object of interest is visible [1] and a story is collection of scenes that defines an unbroken event. Composing all hierarchy form a video stream is sometimes very hard. Hence detection of fundamental units, shots, from the stream becomes more important.

During the last decade, there have been many algorithms in the published literature to solve the shot detection problem given a video sequence [2, 3, 5, 10, 12]. All these algorithms have their own merits and demerits. However, on the average, an acceptable shot detection algorithm should satisfy the following properties: reasonable computational complexity, ease of implementation and robustness to gradual transitions and applicability to real time problems. A gradual transition can either be a fading out which is defined by dissolving of frame into a black frame or a dissolving effect which is defined by transition from one image to another image [1].

In this paper, we propose a new shot detection algorithm satisfying the properties defined above. The method is based on calculation of principal coordinate system from spatial information determined from RGB color space by applying eigenspace decomposition.

The organization of the paper is as follows. In the next section we will give a brief review of the existing methods. In section 3, we intuitively describe our approach to shot detection. In the following section, we will discuss the advantages of the proposed algorithm. In section 4, the experiments and the test data is described. Finally, conclusions are drawn in section 5.

2. Review of Existing Techniques

There are two ways to categorize the shot detection algorithms. The first categorization can be done according to the features they are using: color space, texture information, shape information or motion vector. They can also be categorized according to the method they use to solve the problem, pixel differences, statistical differences, histograms, compression differences, edge tracking or motion vector differences. A comparison between these approaches can be found in [1].

The most common approach to obtain boundaries of a shot is based on color information of the frame, [4, 8, 2, 9]. In the recent studies, no matter which color space is of interest, main advantages of using color is its ease of implementation, descriptive characteristic both in spatial and temporal space and real-time applicability due to simplicity to obtain a feature vector, such as histogram [7]. However, color based algorithms have false positives in presence of camera motion, object motion or illumination change. In most cases, post processing is accompanied to compensate for the stumbling blocks of color-based shot detection, such as motion compensation [6] or illumination reduction [4]. However, using post processing introduces the well-known trade off between speed and accuracy, the more you improve accuracy the longer it takes.

Shot detection algorithms based on statistical differences extract statistical characteristics of the spatial information and look for changes in temporal domain. Most commonly used features are the mean and the standard deviation [10], however these algorithms are slow due to high computational cost of statistical formulae. These algorithms also generate false positives on illumination change [1].

Compression difference based shot detection algorithms utilize the availability of the features stored in video and image compression algorithms, such as DCT coefficients of JPEG [11], or motion vectors in mpeg-2 [5]. Though these algorithms are fast due to the presence of features in the compressed data, they can't be applied in real time environments.

Edge map based approaches for shot detection use the assumption that the changes in spatial domain will result in appearing or fading of edges. The percentage of change in edges decides the shot boundary. Though the approach, as stated in [8], is sensitive to motion, the response time of the algorithm is not real time, because of the motion compensation step, thus can't be applied to surveillance systems.

To sum up, all algorithms in the printed literature suffer the trade of between speed and accuracy. On the other hand, it should also be noted that introduction of complex algorithms do not further improve the detection of the shot boundaries compared to simple algorithms [1].

In the next section, we will give intuitive description of our algorithm based on eigenspace decomposition. We will also discuss the advantages and disadvantages of this new approach.

3. Principal Coordinate System

Eigenspace decomposition, which is also called principal component analysis (PCA), was first defined by statisticians, and it was basically used for reducing the dimensions of the working space. The reduction is obtained through selecting the axes that maximize the variation in the data, which corresponds to maximum eigenvalued eigenvectors. These axes define the new coordinate system.

3.1 PCA in Literature

Vision community has long been using PCA for reconstruction, recognition and pattern classification purposes due to its descriptive nature. Images are converted from matrix notations to vectors by concatenating each row, which is followed by calculation of covariance. The system is trained by decomposing the working space through covariance matrix, which results in orthonormal system of eigenvectors.

In contrast to this classical approach, we use eigenspace decomposition for obtaining the principal coordinate system which best characterizes the working space. In our approach, no training step is required. The algorithm uses spatial information, which is defined in RGB color space.

Temporal aligning, which is necessary to guarantee the correlation in temporal domain, is not required for this algorithm since the color space for every image is already aligned. This fact diverts the framework from similar subspace decomposition approaches available in the literature [2]. For increasing correlation, Emile et. al. [2] used a feature vector of motion parameters and color histograms obtained from the MPEG-1 compression. Their approach suffers the trade off between speed and accuracy, in real time, you should obtain motion vectors without MPEG-1 compression or you should compress the input video prior to shot boundary detection.

In the next subsection, we will give the theory behind our approach.

3.2 Algorithm

According to the statistical description of eigenspace decomposition, eigenvectors give the most descriptive coordinate system for the working space by maximizing the variation of the data. We can define the similarity measure between two consecutive frames in the same shot by calculating the transformation between principal coordinate systems. Below we define the steps of our algorithm:

1. Find 3x3-covariance matrix of the color space given by

$$C_{f_{i,j}} = [R_{f_{i,j}} \quad G_{f_{i,j}} \quad B_{f_{i,j}}]^{\mathrm{T}} [R^T_{f_{i,j}} \quad G^T_{f_{i,j}} \quad B^T_{f_{i,j}}] \qquad (1)$$

where R, G and B row vectors denote mean normalized red, green and blue components of ith frame of jth shot.

2. Determine eigenvectors of the covariance matrix. For determining the shot boundaries, we use maximum eigenvalued eigenvector, principal axis, because of its uniqueness and descriptive nature of the data.
3. For finding the shots given a video stream, we use a “cluster seeking” approach commonly used among pattern recognition society,

$$S(X,Z) = (X^T \cdot Z) / (\|X\| \cdot \|Z\|) \qquad (2)$$

where X and Z are principal axes of succeeding frames’ color spaces and $S(X,Z)$ denotes the angle of rotation between these axes.

4. Shots boundaries are defined by thresholding the rotation changes for the whole video stream.

4. Experiments

To demonstrate the efficiency of our approach, we compared our results with three very commonly used approaches. Two of these algorithms are based on histogram differences, where the first one is on gray level and the second one is in HSV color space, and the other algorithm utilizes a simple change detection approach.

We implement and run all algorithms in the same development environment to have a fair timing comparison. The comparisons between these four algorithms are based on computational complexity, ease of implementation, robust shot boundary detection and applicability to real time problems, such as surveillance.

We tested the algorithms on two 20 minutes long video streams, sampled by 10 frames per second. The first test stream is from CNN Money-Line and the other test stream is from Channel 6 Primetime News. The test streams can be categorized into commercials and anchor news. Both categories have high motion and gradual transitions.

4.1 Accuracy

We defined the accuracy of an algorithm by measuring its distance from manually generated ground truth by,

$$\theta = \sum_{i=1}^{N_s} \min_{j=1..N_g} \left| x_i - Z_j \right| . \qquad (3)$$

In equation 3, Ns is the number of shots found by the algorithm, Ng is the number of shots in the ground truth, Z_j is jth shot boundary belonging to the ground truth and x_i is the ith shot boundary found by the algorithm. In addition to the θ value, we penalize the algorithms if they fail to detect a shot boundary by the distance between missing boundary and the nearest boundary to it. In table 1, we summarize the accuracy of the four approaches for the measure defined in equation 3. In table 2 gives the shot detection timings for these approaches.

In figure 1, we plot the shot boundaries for four different approaches and manually generated ground truth, for the CNN Money-line program. In figure 4, we give the shot boundaries obtained using four approaches for the CH 6 primetime news program.

Table 1: Accuracy comparison of four algorithms.

	Gray Histogram	Change detection	HSV color space	Principal coor. sys.
Video stream 1	3927	4016	2218	2556
Video stream 2	6864	7002	6474	6517

Table 2: Timing comparison of four algorithms (10 fps).

	Gray Histogram	Change detection	HSV color space	Principal coor. sys.
Video stream 1	7 min.	7 min.	123 min.	22 min.
Video stream 2	7 min.	7 min.	123 min.	22 min.

4.2 Complexity and Applicability to Real Time Environments

Complexity is an important issue. The complexity of an algorithm is defined by the number instructions executed. For comparing the algorithms, we defined the complexity of them by the number of additions and multiplications they execute. We assume each frame is mxn. Note that for gray level based approach there are only mxn data points, however other two algorithms have 3xmxn data points.

Gray level histogram based approach has total of 512+mxn additions, which results in a very fast algorithm, thus it is applicable to real time environments. The other gray level based algorithm, which calculates the change difference between images has mxn subtractions only. In literature there are extensions of these algorithms, where statistical information is also incorporated, however introduction of additional overhead does not drastically improve the performance.

The HSV color space based shot detection algorithm we implemented used 16-bin histogram to determine the shots. The algorithm requires 30xmxn compares and mxn additions. Like in the previous algorithm we didn't apply any statistical constraints on the data. However, because of using better bases for data, this algorithm gives better detection then the gray level based approach. However, one drawback of this algorithm is the timing constraints, which is important for the real time situations.

Our approach has no compares like the gray level approach, where as it has more multiplication and additions than the other approaches. The algorithm has 30xmxn multiplications to determine the covariance matrix and constant amount of time for obtaining eigenvectors of 3x3 matrix. The approach responds six times faster then the HSV based algorithm. However, it is three times slower than the gray-level based approaches, the reason is basically our approach uses three times more data then the gray-level based approach.

4.3 Ease of Implementation

Gray level based approaches are the simplest algorithms to be applied to the data set. The only steps to be considered are to obtain the gray level histogram from the image. On the other hand, HSV color-space based algorithm and the principal coordinate system based approach are equally likely to implement.

4.4 Clustering Video Stream into News and Commercials

News streams are generally composed of anchor news, advertisements where advertisements may be either commercial or non-commercial. The advertisements are identified by the presence of high motion, many gradual transitions and changes in color space. On the other hand, anchor news has low motion and less change in color space.

To cluster news video streams into news and advertisements, based on the shots boundaries detected by principal coordinate system approach, we used the minimum eigenvalued eigenvector, v_3. Minimum eigenvalued eigenvectors have been used in the dimensionality reduction of the feature space in pattern clustering due to its characteristic of increasing in-class correlation.

In contrast to v1, any small change in color space results in high rotation of succeeding v3's. In figure 2, we plot v1 and v3 for the video stream composed of anchor news and advertisements. Since the variation of the colors in advertisements is more than the news clusters, one can notice the regions of advertisements.

To define a shot if it is anchor news or advertisement, we calculated the mean of v3's in a shot and if it is below a threshold, it is labeled as anchor news; otherwise it is labeled as advertisement. We assume that the variation in color space for the anchor news is very low where as the variation in color space for the advertisement is very high. This assumption does not necessarily hold for every kind of advertisement, there may be shots where the color space does not change. In figure 3, we give the clustering result for the money-line program along with the ground truth, the black regions depict the news and the white regions show the advertisements. The decision results for CH6 news stream is given in figure 5. As seen the clustering has false positives due to the not changing color space in the advertisement regions. These false positives can be overcome by introducing another constraint imposed on the decision boundaries obtained from our approach. This constraint can be the audio information for each decision regions.

5. Conclusion

The results for shot detection are satisfactory and the speed of the proposed approach is quite fast to be applied to real-time surveillance systems. The only problem we encountered in the shot detection was the instable response of eigenspace decomposition for frames where no all colors are concentrated on one point such as black or blue frames between some shots. The algorithm gave several false positives for these one-colored frames.

The story segmentation is a useful real-time extension to the proposed approach. As can be seen in figure 3 and 5, the anchor news are identified with an acceptable error, where as the system has more false positives in the regions of advertisements. If an audio analysis is also employed along with the image data, the false positives in the advertisement regions can be solved, since the advertisements generally accompany music.

References

[1] S. Boreczky, L.A. Rowe, “Comparison of video shot boundary detection techniques,” *Storage & Retrieval for Image and Video Databases IV*, I.K. Sethi, and R.C. Jain, Editors, Proc. SPIE 2670, pp. 170-179 (1996).

[2] E. Sahouria, A. Zakhor, “Content Analysis of Video Using Principal Components,” *IEEE transactions on circuits and systems for video technology*, Vol. 9, No. 8, December 1999, pp. 1290-1298.

[3] E. Stringa, C. S. Regazzoni, “Real-time Video-shot Detection for Scene Surveillance applications,” *IEEE Trans. On Image Processing*, Vol. 9, No. 1, January 2000, pp. 69-79.

[4] W. Zhao, J. Wang, D. Bhat, K. Sakiewicz, N. Nandhakumar and W. Chang, “Improving color based video shot detection,” *IEEE Int. Conf. On Multimedia Computing and Systems*, Vol. 2, 1999, pp. 752-756.

[5] I. Koprinska, S. Carrato, “Video segmentation of MPEG compressed data,” *IEEE Int. Conf. On Electronics, Circuits and Systems*, Vol. 2, 1998, pp. 243-246.

[6] H.J. Zhang, A. Kankanhalli and S.W. Smoliar. “Automatic Partitioning of full-motion video,” *Multimedia Systems*, Vol. 1, No. 1, 1993, pp. 10-28.

[7] M.J. Swain, D.H. Ballard, “Color indexing,” Int. Journal of Computer Vision, Vol. 7. 1991, pp. 11-32.

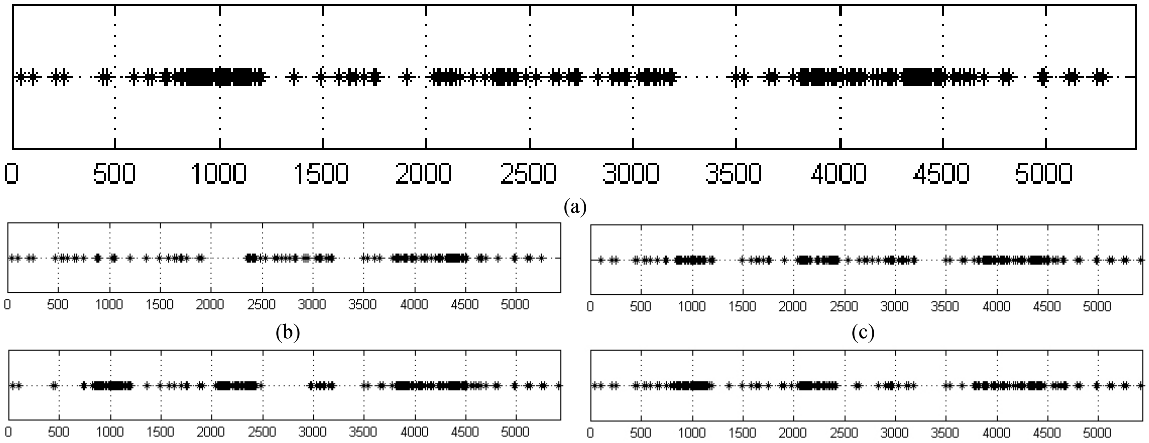
[8] R. Zabih, J. Miller, and K. Mai, “A feature-based algorithm for detecting and classifying scene breaks,” *Proc. ACM Multimedia 95*, San Francisco, CA, November, 1993, pp. 189-200.

[9] N. Haering, “A framework for the design of event detectors,” *Ph.D. Dissertation University of Central Florida*, 1999.

[10] R. Kasturi, R. Jain, “Dynamic vision,” in *Computer Vision: Principles*, R. Kasturi, R. Jain, Editors, IEEE Computer Society Press, Washington, 1991.

[11] F. Arman, A. Hsu, and M-Y Chiu, “Image processing on encoded video sequences,” Multimedia Systems, Vol. 1, No. 5, 1994, pp. 211-219.

[12] V. Kobla, D. Doermann, C. Faloutsos, “Developing high level representations of video clips using videotrails,” Prc. SPIE, vol. 3312, 1998, pp. 81-92.



(a)

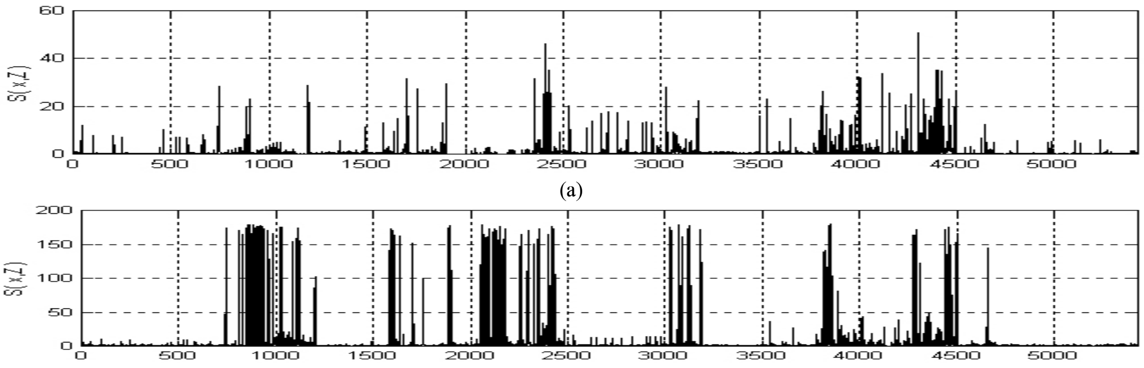
(b)

(c)

(d)

(e)

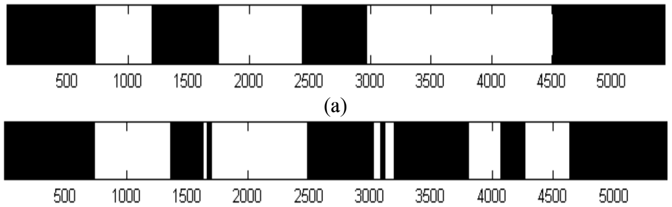
Figure 1: Shot boundaries for CNN Money-Line program; (a) ground truth determined manually, (b) principal coordinate system approach, (c) HSV color space approach, (d) gray color space histogram approach, (e) change detection.



(a)

(b)

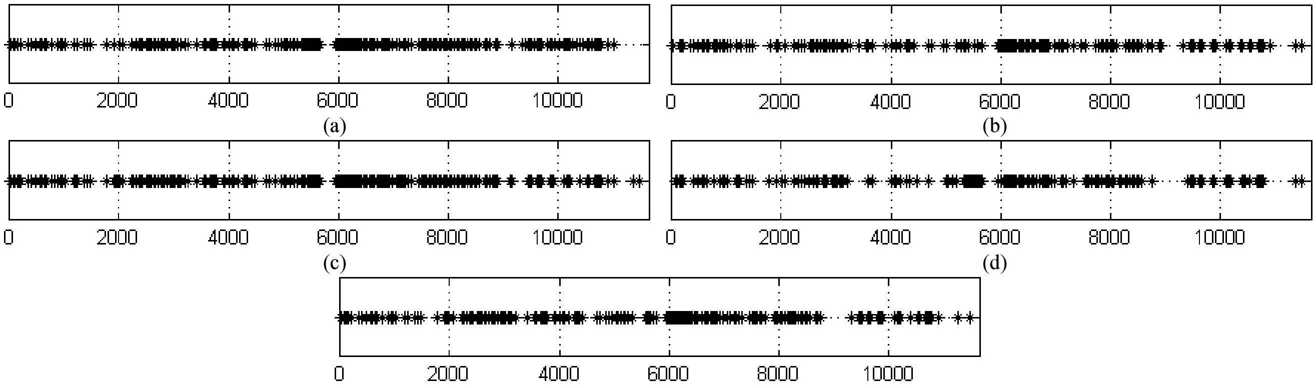
Figure 2: Rotation angles for (a) v1, maximum eigenvalued eigenvector, (b) v3, minimum eigenvalued eigenvector.



(a)

(b)

Figure 3: Clustering video into news and advertisement, black regions depict news and white regions are advertisements; (a) ground truth manually extracted from CNN stream, (b) clustered video stories by v3 axis of principal coordinate system, we allowed 4.5 degrees of rotation to differentiate between news and advertisements.



(a)

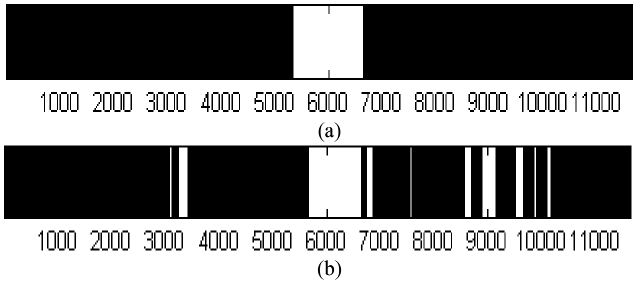
(b)

(c)

(d)

(e)

Figure 4: Shot boundaries for CH6 News program; (a) ground truth determined manually, (b) principal coordinate system approach, (c) HSV color space approach, (d) gray color space histogram approach, (e) change detection.



(a)

(b)

Figure 5: Clustering video into news and advertisement, black regions depict news and white regions are advertisements; (a) ground truth manually extracted from CH6 stream, (b) clustered video stories by v3 axis of principal coordinate system, we allowed 4.5 degrees of rotation to differentiate between news and advertisements.