2008 Seventh International Conference on Machine Learning and Applications

Video Shot Segmentation using Late Fusion Technique

C. Krishna Mohan[1], N. Dhananjaya[2], and B. Yegnanarayana[3]

[1] National Institute of Technology Karnataka Suratkal
Srinivasnagar - 575 025, Karnataka, India
[2] Indian Institute of Technology Madras, Chennai - 600 036, India
[3] Internation Institute of Information Technology, Hyderabad - 500 032, India

email: chalavadi_km@yahoo.com, dhanu@cs.iitm.ernet.in, yegna@iiit.ac.in

Abstract

In this paper, a new method for detecting shot boundaries in video sequences using a late fusion technique is proposed. The method uses color histogram as the feature, and processes each bin separately for detecting shot boundaries. The decisions from individual bins are combined later for hypothesizing the presence of shot boundaries. The method provides a certain degree of robustness against illumination and camera/object motion, as it ignores small changes in the bins. While the early fusion techniques rely on the extent of change in color information, the proposed technique relies on the number of significant changes. Experimental results successfully validate the new method and show that it can effectively detect both abrupt and gradual transitions.

Keywords: video segmentation, shot boundary detection, early fusion, late fusion, video content analysis.

1. Introduction

Content-based indexing and retrieval of digital video is an active research area. Video segmentation is the first preprocessing step to further analyze the video content for indexing and browsing. Video segmentation or shot boundary detection involves temporal segmentation of video sequences into elementary units, called shots. A shot in a video is a contiguous sequence of video frames recorded from a single camera operation, representing a continuous action in time and space [4]. Shot transitions can be abrupt (cuts) or gradual (fades, dissolves and wipes).

Many techniques have been developed to detect the video shot boundaries. A good review and comparison of existing methods is presented in [6, 5]. Gradual transitions are generally more difficult to detect, as they can often be confused with camera/object motion. Detection of gradual transitions, such as fades and dissolves, is examined in [3, 7]. Color histogram is one of the most widely used feature for detecting the shot transitions [3, 8]. The feature vectors representing color information are generally large and sparse, prompting a reduction in dimensionality. Singular value decomposition [1] and independent component analysis [9] have been examined for this purpose. A wide variety of dissimilarity measures have been used in the literature [9, 1]. Some of the commonly used measures are: Euclidean distance, cosine dissimilarity, Mahalanobis distance and log-likelihood ratio. Information theoretic measures like mutual information and joint entropy between consecutive frames have also been proposed for detecting cuts and fades [2].

In this paper, we propose a method for shot boundary detection which involves late fusion of decisions obtained by processing individual bins separately. The early fusion technique computes the net changes in all the bins or dimensions, which can be significantly large although the change in the individual bins is small. This can lead to false hypotheses of shot boundaries thereby bringing down the performance. The late fusion technique provides robustness against such cases which are typically caused by illumination changes and camera/object motion. Experimental results indicate that the proposed method can effectively detect both abrupt transitions and gradual transitions.

The remainder of this paper is organized as follows: In Section 2, we propose two modifications to early fusion algorithm. In Section 3, the proposed late fusion technique for shot boundary detection is described. Experimental results are discussed in Section 4. Section 5 gives conclusion of this work.

978-0-7695-3495-4/08 $25.00 © 2008 IEEE
DOI 10.1109/ICMLA.2008.88

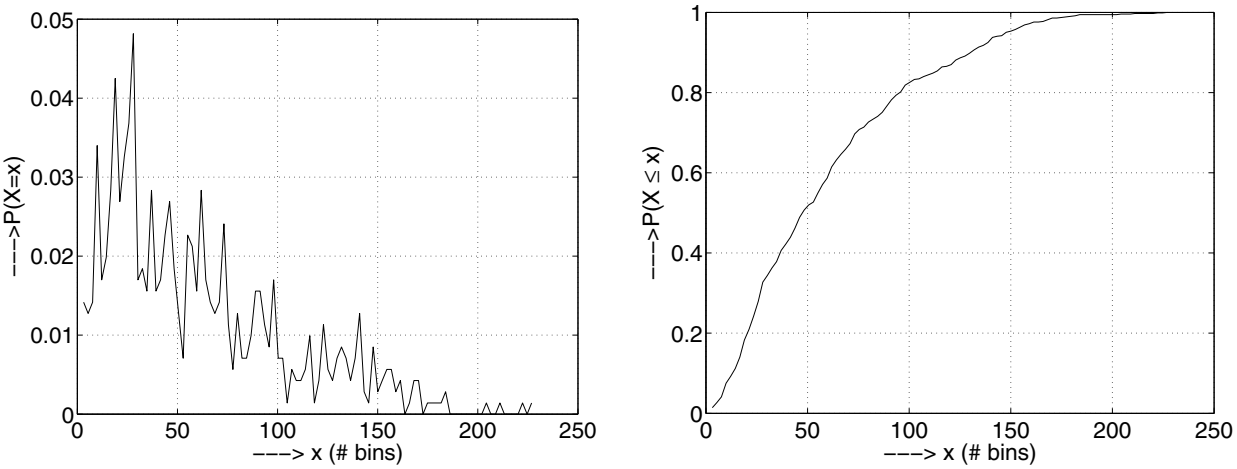

Figure 1. (a) Probability distribution of shot boundaries in terms of the number of color bins changing significantly. (b) Cumulative distribution of (a).

2 Shot Boundary Detection by Early Fusion

Shot boundary detection involves testing, at every frame index n of a given video of length N_v frames, the following two hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \text{A shot boundary exists at frame index } n. \\ \mathcal{H}_1: &\quad \text{No shot boundary exists at frame index } n. \end{aligned} \quad (1)$$

Let $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n, \ldots, \boldsymbol{x}_{N_v}\}$ be the sequence of feature vectors of dimension p representing the N_v frames in the video. Testing of the hypotheses at the frame index n involves computation of a dissimilarity value, $d[n] = d(\mathcal{X}_L, \mathcal{X}_R)$, between two sequences of N feature vectors $\mathcal{X}_L = \{\boldsymbol{x}_{n-1}, \boldsymbol{x}_{n-2}, \ldots, \boldsymbol{x}_{n-N}\}$ and $\mathcal{X}_R = \{\boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+N-1}\}$ to the left and right of n, respectively. The value of N can vary from one frame to a few frames (corresponding to less than one or two secs). If the dissimilarity value is greater than a threshold $\tau[n]$, the hypothesis $\mathcal{H}_0$, that a shot boundary exists, is chosen.

In this work, Euclidean distance is used as the dissimilarity measure. An adaptive threshold is computed using the variance of a few frames before the frame at which the hypotheses is tested. Let $d[n] = d_{euc}(\boldsymbol{\mu}_L, \boldsymbol{\mu}_R)$ denote the Euclidean distance between the means of $\mathcal{X}_L$ and $\mathcal{X}_R$. If $\sigma_L[n] = \sqrt{||\Sigma_L||}$ represents the amount of variability within a block of N frames to the left of n, then the dynamic threshold is computed as $\tau_L[n] = \beta * \sigma_L[n]$, where Σ_L is the covariance matrix of $\mathcal{X}_L$, and β is a scaling parameter.

The first set of shot boundaries are hypothesized using the dynamic threshold based technique described above with a window size of N =1. In order to reduce the number of misses, the video is also processed in the reverse (or backward) direction, which is equivalent to comparing the dissimilarity value with $\tau_R[n]$, the amount of variability to the right of n. The condition for hypothesizing a shot boundary thus becomes

$$d_{euc}(\boldsymbol{\mu}_L, \boldsymbol{\mu}_R) > \tau_L[n] \quad | \quad d_{euc}(\boldsymbol{\mu}_L, \boldsymbol{\mu}_R) > \tau_R[n] \quad (2)$$

The use of 'OR' (|) logic in the bidirectional processing of the video increases the number of false hypotheses, which are reduced by validating the hypothesized boundaries using the same condition as in Eq. 2, but with a larger window size, say N =10.

3 Shot Boundary Detection by Late Fusion

The early fusion technique described in the previous section was based on the overall change in color histogram between adjacent frames in a video sequence. The dimension of color histogram(512 in this case) was chosen to provide adequate representation to each color component. However, not all components of the color histogram feature vectors are populated for a given frame of video. Secondly, not all components of the color histogram change significantly in the neighbourhood of a shot boundary. It is observed that in general, a small number of color bins undergo a significant change when there is shot boundary. Figs. 1 (a) and (b) show the probability and cumulative distributions of the number of bins changing significantly at the actual shot boundaries, for a threshold factor of β =5. It can be seen from Fig. 1 (b) that around 50% of the shot boundaries have 50 or less number of bins changing significantly (approximately 10% of the total number of bins) and around 82% of the shot boundaries have 100 or

Table 1. Video data used for shot boundary detection experiments.

Clip ID	Duration (min)	# cuts	# graduals
BBC-1	22	125	18
BBC-2	23	154	29
CNN-1	24	76	72
CNN-2	24	43	47
NDTV-1	27	135	7
Overall	120	533	173

less number of bins (approximately 20%) changing significantly. Hence we see that a shot boundary can be detected by observing a significant change in a small number of bins.

At the same time, if all components of the color histogram are considered for the computation of dissimilarity, as is the case in early fusion, even a small contribution from each component results in a large value of the overall dissimilarity. This is typically the case when frames in a video sequence change gradually due to object/camera motion and intensity variations, even when there is no shot boundary. To overcome the problem of false hypothesis due to small changes accumulated over a large number of bins, we propose to use the number of bins changing significantly as a measure to hypothesize a shot boundary. We call this as late fusion technique, since the components of color histogram are first observed for significant change, and only then included in the process of decision making. The condition for hypothesizing a shot boundary is exactly same as the early fusion technique outlined in the previous section, except that it is applied on individual bins separately. If the number of bins voting for a shot boundary exceeds a threshold $M = 20$, a shot boundary is hypothesized. The optimal threshold factor M_{opt} corresponds to best F_1 measure. Our observation from the experiments is that less than 20 components of the color histogram are sufficient for detection of shot boundaries, provided that these components change significantly in the vicinity of a shot boundary.

4 Experimental Results

The performance of the shot boundary detection is evaluated on a database of approximately 2 hours of news video. For each video sequence, a human observer has determined the precise location and duration of the edits to be used as ground truth. The database contains a total of 533 cuts and 173 gradual transitions, the details of which are shown in Table 1. The video clips were captured at a rate of 25 frames per second, at 320×240 pixel resolution, and stored in AVI format. A 512-dimension RGB color histogram, obtained by quantizing the 3-D color space into a 8 x 8 x 8 grid, is used as the feature vector.

The performance of the shot boundary detection task is measured in terms of recall $R = N_c/N_m$ and precision $P = N_c/(N_c+N_f)$, where N_m is the total number of actual (or manually marked) shot boundaries, N_c is the number of shot boundaries detected correctly, and N_f is the number of false alarms. A good performance requires both the recall and the precision to be high. The choice of the threshold factor β is crucial. A small value of β improves the recall, while at the same time reducing the precision. A large value of β has the reverse effect on the recall and precision. A compromise between recall and precision is obtained by using a measure combining the recall and precision, given by $F_1 = 2RP/(R+P)$. The performance of shot boundary detection using forward and backward processing of video by early fusion of evidence is given in Table 2. Performance after combining the evidence obtained using forward and backward processing is given in Table 3.

The optimal threshold factor β_{opt} corresponds to best F_1 measure. It is to be noted here that the optimal threshold factor is different for different clips, and also for forward and backward directions of the same clip. It can also be seen that the OR logic improves the performance, while the AND logic reduces the optimal F_1 value. This is mainly due to a high probability that one of the two sides of a shot boundary has a large variance among the feature vectors. The performance of shot boundary detection by late fusion of decisions along individual dimensions is given Table 4. The comparison of Tables 3 and 4 indicates that the performance of late fusion algorithm is comparable to that of early fusion (when OR logic is used for combination).

5 Conclusions

In this paper, we have presented a new method for video shot boundary detection based on the late fusion of evidences. The early fusion technique computes the net changes in all the bins or dimensions, which can be significantly large although the change in the individual bins is small. This can lead to false hypotheses of shot boundaries thereby bringing down the performance. The late fusion technique provides robustness against such cases which are typically caused by illumination changes and camera/object motion. The simulations show that the method achieved good performance for detecting both abrupt transitions and gradual transitions. To further improve the performance of the shot boundary detection, we can combine the evidence due to late fusion with the evidence due to early fusion to exploit the advantages of both the methods which will be our future work.

Table 2. Performance of shot boundary detection using forward and backward processing of video by early fusion of evidence.

Clip ID	Forward Processing				Backward Processing			
	β_{opt}	R	P	F_1	β_{opt}	R	P	F_1
BBC-1	10.0	0.881	0.926	0.903	8.0	0.902	0.942	0.921
BBC-2	9.5	0.852	0.912	0.881	6.5	0.858	0.844	0.851
CNN-1	8.0	0.878	0.909	0.893	6.5	0.892	0.904	0.898
CNN-2	11.0	0.867	0.897	0.881	9.0	0.844	0.854	0.849
NDTV-1	12.5	0.768	0.908	0.832	6.5	0.873	0.780	0.824
Overall	9.5	0.854	0.889	0.871	6.5	0.890	0.820	0.853

Table 3. Performance of shot boundary detection by combining the evidence obtained using forward and backward processing of video.

Clip ID	Combined (OR)				Combined (AND)			
	β_{opt}	R	P	F_1	β_{opt}	R	P	F_1
BBC-1	10.0	0.937	0.918	0.927	3.0	0.881	0.592	0.708
BBC-2	9.5	0.902	0.878	0.889	3.0	0.863	0.583	0.696
CNN-1	8.0	0.953	0.898	0.925	3.0	0.736	0.407	0.524
CNN-2	13.0	0.878	0.940	0.908	3.0	0.700	0.240	0.358
NDTV-1	9.5	0.901	0.837	0.868	3.0	0.852	0.531	0.654
Overall	10.0	0.908	0.882	0.895	3.0	0.806	0.470	0.588

Table 4. Performance of shot boundary detection by late fusion of decisions along individual dimensions.

Clip ID	M_{opt}	R	P	F_1
BBC-1	19.000	0.930	0.911	0.920
BBC-2	17.000	0.880	0.880	0.880
CNN-1	12.000	0.865	0.914	0.889
CNN-2	12.000	0.900	0.976	0.936
NDTV-1	17.000	0.930	0.841	0.883
Overall	15.000	0.888	0.877	0.882

References

[1] Z. Cernekova, C. Kotropoulos, and I. Pitas. Video shot segmentation using sigular value decomposition. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Hong Kong, pp. 181-184, Apr. 6-10, 2003.

[2] Z. Cernekova, C. Nikou, and I. Pitas. Shot detection in video sequences using entropy-based metrics. In *Proc. IEEE Int. Conf. Image Processing*, pp. 421-424, 2002.

[3] M. Drew, Z.-N. Li, and X. Zhong. Video dissolve and wipe detection via spatio-temporal images of chromatic histogram diffrences. In *Proc. IEEE Int. Conf. Image Processing*, pp. 929-932, 2000.

[4] U. Gargi, R. Kasturi, and S. H. Strayer. Performance characterization of video-shot-change detection methods. *IEEE Trans. Circuits, Systems, Video Technology*, 10(1):1–13, Feb. 2000.

[5] A. Hanjalic. Shot-boundary detection: unraveled and resolved. *IEEE Trans. Circuits, Systems, Video Technology*, 12(2):90–104, Feb. 2002.

[6] R. Lienhart. Reliable transition detection in videos: A survey and practitioners guide. In *Int. Journal of Image and Graphics*, pp. 469-486, 2001.

[7] B. T. Truong, C. Dorai, and S. Venkatesh. New enhancements to cut, fade and dissolve detection processes in video segmentation. In *ACM Int. Conf. on Multimedia*, pp. 219-227, Nov. 2000.

[8] S. Tsekeridou and I. Pitas. Content-based video parsing and indexing based on audio-visual interaction. *IEEE Trans. Circuits, Systems, Video Technology*, 11(4):522–535, 2001.

[9] J. Zhou and X.-P. Zhang. Video shot boundary detection using independent component analysis. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Philadelphia, USA, pp. 541-544, Mar. 18-23, 2005.