A Multi-Level Framework for Video Shot Structuring

Yun Zhai and Mubarak Shah

School of Computer Science
University of Central Florida
Orlando, Florida 32816

Abstract. Video shots provide the most basic meaningful segments for video analysis and understanding. In this paper, we present a detection and classification framework for the video shot segmentation in a coarse-to-fine fashion. The initial transitions are detected from a sub-sampled video space. These coarse segments are later refined in the original video space with the technique of illumination artifacts removal and transition finalization. The transition type (abrupt or gradual) are finally determined by exploiting the histogram intersection plot. The proposed method has been tested on a large amount of videos, which contain a variety of types of shot transitions. Accurate and competitive results have been obtained.

1 Introduction

The increasing amount of video data available to us poses challenges to develop tools for video indexing and searching, so that users can efficiently navigate through it. As the most based semantic meaningful segments of the video, detecting shots becomes an important and interesting problem in video processing and analysis. A video shot is defined as a sequence of frames taken by a single camera with no major changes in the visual content. The transitions between shots can be categorized into two types: abrupt and gradual. The abrupt transitions is generated by directly appending one shot after another. On the other hand, the gradual transitions are generated to present production effects, e.g., wipes, fade-in, fade-out, dissolve, etc. The goal in the shot detection usually contains two parts: transition localization and transition type determination.

Many efforts have been devoted into this area for the past years, and they can be differentiated by their underlying mechanisms. Boreczky *et al.* [2] has proposed a shot detection method based on the direct comparison of the pixels in the consecutive frames. One modification of this approach [8] is to count the number of pixels that are significantly changed. They work well for sequences taken by still camera, but is highly sensitive to the camera and scene object motion. Another group of approaches use the similarity measures between global feature vectors. One popular feature used is the color histogram. Yeo *et al.* [9] proposed an difference measure by computing the sum of absolute differences between corresponding bins in the histogram. Furht *et al.* [3] has used the color histograms in HSV space to make the system less sensitive to the lighting conditions. An alternative to the global comparison is the block-wise histogram comparison proposed by Nagasaka *et al.* [7]. There methods are robust against the motion. However, they do not perform well on slow gradual transitions. Another important

feature have been used is the edge information. Zabih *et al.* [10] has proposed an edge-based similarity measure between frames. After motion compensation, the percentage of edge pixels exiting from one frame to its following frames is computed. Even though it is robust against global motion, the computational complexity is high.

In this paper, we present a coarse-to-fine approach for not only detecting the transitions between video shots, but also classifying the transitions into one of the two types: abrupt transition and gradual transition. The video is first segmented into coarse segments by analyzing in the sub-sampled video space. The shots transitions then are refined by illumination artifacts removal technique and are finalized in the original video space. The types of the transitions are determined by looking at the neighborhoods of the initial transition boundaries. The rest of this paper is organized as follows: Section 2 discusses the overall algorithm, including boundary initialization, illumination artifact removal, transition type determination and transition boundary finalization; Section 3 presents the system evaluation results; finally, Section 4 concludes our work.

2 Proposed Framework

2.1 Transition Boundary Initialization

During a shot transition, the visual similarity of the consecutive frames changes. This can be detected by observing the color histograms of the frames. We use a 3-D color histogram in RGB space, allocating 8 bins for each dimension. Let $D(i)$ represents the histogram intersection between frames f^{i-1} and f^i, which is computed as,

$$D(i) = \sum_{allbin\ b} min(H_{i-1}(b), H_i(b)), \tag{1}$$

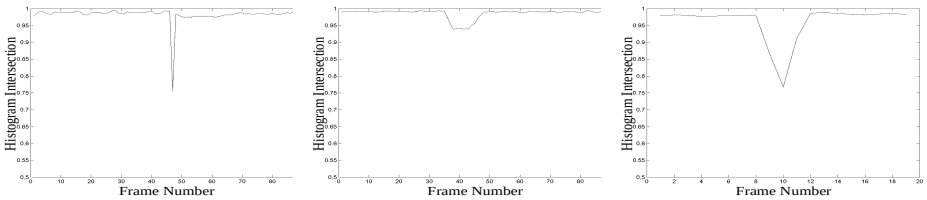
where H_{i-1} and H_i are histograms of frames f^{i-1} and f^i respectively, and b is the individual bins. A transition boundary at f^i is found if:

$$\begin{aligned} D(i-1) - D(i) > T_{color}, \\ D(i+1) - D(i) > T_{color}, \end{aligned} \tag{2}$$

where T_{color} is the threshold that captures the significant difference between the color statistics of two frames.

For abrupt transitions (Fig.1(a)), where the difference in color distribution is large enough, the above frame-to-frame histogram intersection performs well. However, for the gradual transitions (Fig.1(b)), the color differences between consecutive frames are not sufficiently significant to be captured, miss detection occur. To solve this problem, we first temporally sub-sample the original video sequences (every fifth frame in the experiments). The histogram intersection is then applied to the sub-sampled sequences, thus amplifying the frame-to-frame visual differences (Fig.1(c)). This initial estimate of shot boundary, which may not be accurate, is refined in the next step.

Once the approximate location of a transition boundary is obtained, we localize it at the highest sampling rate. This is achieved by finding the frame corresponding to the



(a) (b) (c)

Fig. 1. (a). Histogram intersection plot for a short video containing an abrupt transition; (b). Histogram intersection plot for a sequence containing a gradual transition; (c). Histogram intersection plot for the sub-sampled video in (b) with sampling at every fifth frame.

local minimum of the color histogram intersection plot. Let P be the initial estimate of the transition boundary and a be the search range, the localized transition boundary at frame M is determined as,

$$M = arg\ min(\{D(P-a), ..., D(P+a)\}). \tag{3}$$

2.2 Illumination Artifact Removal

The detection of the shot transitions is followed by the removal of outliers. We have observed that in many videos that relate to meetings, briefings, celebrities, politicians, the most common outliers are caused by the camera flashes. In such cases, the illumination of consecutive frames abruptly changes and results in over detection of the transition boundaries. To remove such outliers, we compute the average color statistics, K_L and K_R, of the immediate left and right neighborhoods of a candidate transition boundary. The visual similarity of these two neighbors is computed in terms of the Bhattacharya distance d_B between K_L and K_R,

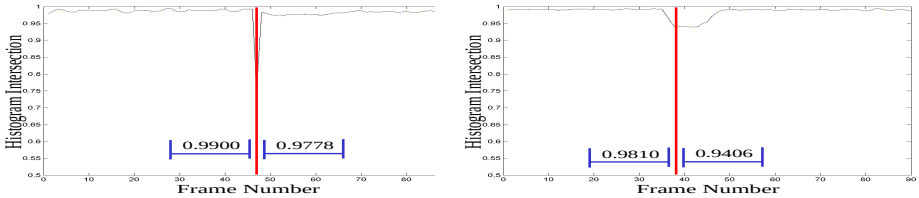
$$Sim(K_L, K_R) = exp\Big(- d_B^2(K_L, K_R)\Big), \tag{4}$$

where $d_B = -ln(\sum_{b=1}^{k} \sqrt{K_L^b, K_R^b})$. The candidate boundary is removed if both of the neighborhoods present high similarity.

2.3 Determining Transition Type

Once the transition boundaries are localized, they are then classified into one of two categories: abrupt and gradual. Examples of gradual transitions are dissolves, fade-ins, fade-outs, wipes, and etc, and they often last for longer temporal periods. The length of transition, however, may differ for different types of transitions. Since the estimated initial boundaries from previous steps are represented by single frames, the gradual transition could take place before the initial boundary, after the initial boundary, or across the initial boundary (Fig.2(b)).

To determine the transition type, we consider frames in a neighborhood of size b, on each side of the detected transition boundary P. If the transition is a gradual



(a) (b)

Fig. 2. (a). Histogram intersection plot for a sequence with an abrupt transition; (b). Histogram intersection plot for a sequence with a gradual transition. The average histogram intersection of the neighbors around the initial transitions are shown in the figure. The windows size is 30 frames in the experiment.

transition, either one of the neighborhoods or both of them higher visual activity. The visual activity of each neighborhood is computed as the average histogram intersections D_L and D_R in each neighbor,

$$D_L = \frac{1}{b}\sum_{i=1}^{b} D(P-i),$$
$$D_R = \frac{1}{b}\sum_{i=1}^{b} D(P+i). \tag{5}$$

If both D_L and D_R are high (visually smooth), the transition is categorized as abrupt (Fig.2(a)). Otherwise, the transition is classified as gradual. Examples for both of the types are shown in Fig.2(b) with visual activities in their neighborhoods.

2.4 Gradual Transition Boundary Determination

Once the transition type is determined, the exact starting and ending locations of the boundaries need to be determined. The determination for the abrupt transitions is straightforward, since the transition only takes place in two frames. It is more important to locate the beginning and ending frames of the gradual transitions, such the accurate shot representation can be found and used in future video analysis and understanding.

We assume that the transition length is not infinite long. If we pick a point that is far away from the initial boundary, and that point is inside the shot instead of on the transition, then, the neighborhood around that point should be visually smooth. As we move this point towards the initial boundary, the visual activity level in its neighborhood starts raising up at the places where the transition starts or ends. If the point comes from the left side of the initial boundary, the raising up point is the starting frame of the transition. If it comes from the right side of the boundary, it is the ending frame of that transition. This process is demonstrated in Fig.3 .

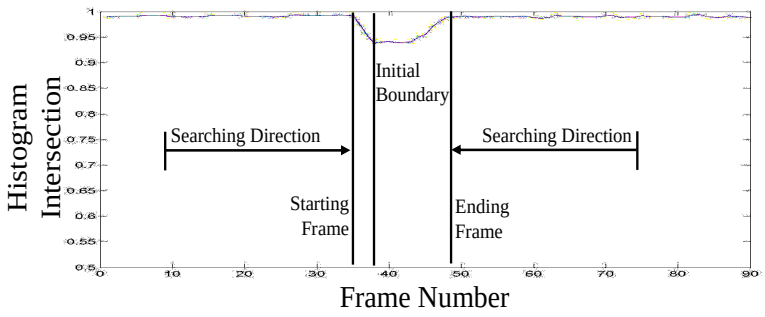


Fig. 3. Locating the starting and ending frames in a gradual transition.

3 System Evaluation Results

The data set we have experimented our framework on is provided by the US National Institute of Technologies and Standards (NIST). The data set is an open benchmark which contains 13 MPEG-1 news videos from CNN, ABC and C-SPAN news networks. It has been used for the TRECVID forum in 2003 [1]. Each video is around thirty minutes long. The news videos from CNN and ABC contain both news program and commercials, while the videos from C-SPAN only contain news programs. The reason of testing on news videos is that there are a rich amount of shots transitions covering both abrupt and gradual types. Furthermore, since it is a Television program, many variation of the gradual transitions are present. Therefore, it is a good testing bed for shot transition detection and type determination.

We have applied our framework on all 13 news videos. There are three types of accuracy measurements:

– Precision/Recall for Abrupt Transitions:

$$Precision = \frac{A_{abrupt}}{X_{abrupt}}, \quad Recall = \frac{A_{abrupt}}{Y_{abrupt}}, \tag{6}$$

where A_{abrupt} is the number of matched abrupt transitions, X_{abrupt} is the number of detected abrupt transitions, and Y_{abrupt} is the number of reference abrupt transitions.

– Precision/Recall for Gradual Transitions:

$$Precision = \frac{A_{gradual}}{X_{gradual}}, \quad Recall = \frac{A_{gradual}}{Y_{gradual}}, \tag{7}$$

where $A_{gradual}$ is the number of matched gradual transitions, $X_{gradual}$ is the number of detected gradual transitions, and $Y_{gradual}$ is the number of reference gradual transitions.

– Frame Based Precision/Recall for Gradual Transitions:

$$Precision = \frac{A'_{gradual}}{X'_{gradual}}, \quad Recall = \frac{A'_{gradual}}{Y'_{gradual}}, \tag{8}$$

where $A'_{gradual}$ is the total number of frames in the overlapping regions in matched gradual transitions, $X'_{gradual}$ is the total number of frames in the detected gradual transitions, and $Y'_{gradual}$ is the total number of frames in the reference gradual transitions.

Filename	Type	Cut Recall	Cut Precision	Grad Recall	Grad Precision	Frame Recall	Frame Precision
19990303.121216	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
19980619_ABC	ABC	0.890	0.934	0.811	0.673	1150/1711	1150/1314
19980224_ABC	ABC	0.820	0.874	0.717	0.648	970/1733	970/1074
19980425_ABC	ABC	0.732	0.896	0.872	0.570	1522/2287	1522/1815
19980222_CNN	CNN	0.737	0.904	0.712	0.365	735/1439	735/1043
19980515_CNN	CNN	0.770	0.923	0.824	0.545	1252/2063	1252/1747
19980531_CNN	CNN	0.757	0.897	0.824	0.494	749/1255	749/937
19980412_ABC	ABC	0.800	0.907	0.861	0.598	1105/2118	1105/1301
2001614.1647460	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
19980308.1216980	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
20010628.1649460	CSPAN	0.962	0.916	0.000	0.000	0/0	0/0
20010702.1650112	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
19980203_CNN	CNN	0.732	0.843	0.853	0.626	1478/2221	1478/1697
Mean		2097/2644	2097/2285	887/1090	887/1616	8961/14827	8961/10928
		0.793	0.918	0.814	0.550	0.604	0.820

Fig. 4. System evaluation results for the first run of the proposed method. Since there is no gradual transition in the C-SPAN videos, the corresponding precision and recall measures are set to zeros.

Filename	Type	Cut Recall	Cut Precision	Grad Recall	Grad Precision	Frame Recall	Frame Precision
19990303.121216	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
19980619_ABC	ABC	0.890	0.940	0.826	0.677	1163/1734	1163/1326
19980224_ABC	ABC	0.820	0.886	0.725	0.641	980/1745	980/1084
19980425_ABC	ABC	0.718	0.898	0.866	0.561	1514/2279	1514/1818
19980222_CNN	CNN	0.728	0.914	0.722	0.361	738/1475	738/1039
19980515_CNN	CNN	0.766	0.923	0.824	0.545	1252/2063	1252/1747
19980531_CNN	CNN	0.740	0.898	0.833	0.491	753/1266	753/964
19980412_ABC	ABC	0.794	0.907	0.861	0.595	1101/2114	1101/1292
2001614.1647460	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
19980308.1216980	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
20010628.1649460	CSPAN	0.962	0.916	0.000	0.000	0/0	0/0
20010702.1650112	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
19980203_CNN	CNN	0.675	0.887	0.894	0.607	1544/2320	1544/1774
Mean		2065/2644	2065/2193	898/1090	898/1687	9045/14996	9045/11044
		0.781	0.942	0.824	0.532	0.603	0.819

Fig. 5. System evaluation results of the second run.

After the completion of the processing pipeline, transitions that are less than 5 frames long are declared as abrupt, and shots shorter than 20 frames are merged with its previous one.

We have experimented for two runs with different processing parameters and the evaluation results generated from the matching program provided by NIST are shown in Fig.4 and Fig.5. Since there is no gradual transition in the C-SPAN videos, the corresponding precision and recall measures are set to zero.

4 Conclusions

In this paper, we have presented a framework for the detection transitions between video shots and the determination of their corresponding types. The method utilizes the visual features in the video frames and performs in a coarse-to-fine fashion. The framework contains four steps: Transition Boundary Initialization, Illumination Artifact Removal, Transition Type Determination and Gradual Transition Boundary Localization. The process is straightforward and easy for implementation. It has been tested on an open benchmark data set provided by NIST, and competitive results have been obtained.

References

1. http://www-nlpir.nist.gov/projects/tv2003/ tv2003.html
2. J.S. Boreczky and L.A. Rowe, “A Comparison of Video Shot Boundary Detection Techniques”, *Journal of Electronic Imaging*, 5(2), April 1996.
3. B. Furht, S. W. Smoliar and H. Zhang, “Video and Image Processing in Multimedia Systems”. Norwell, MA: Kluwer, 1995.
4. A. Hanjalic, “Shot Boundary Detection: Unraveled and Resolved?”, *IEEE Transactions on Circuits and Systems for Video Technology*, 2002.
5. R. Lienhart, “Reliable Transition Detection In Videos: A Survey and Practitioner’s Guide”, *International Journal of Image and Graphics*, 2001.
6. X. Liu and T. Chen, “Shot Boundary Detection Using Temporal Statistics Modeling”, ICASSP, 2002.
7. A. Nagasaka and Y. Tanaka, “Automatic Video Indexing and Full-Video Search For Object Appearances”, *Visual Database Systems II*, 1992.
8. K. Otsuji, Y. Tonomura and Y. Ohba, “Video Browsing Using Brightness Data”, SPIE/IST VCIP, 1991.
9. B.-L. Yeo and B. Liu, “Rapid Scene Analysis On Compressed Video,” *IEEE Trans. Circuits and Systems for Video Technology*, 1995.
10. R. Zabih, J. Miller, and K. Mai, “A Feature-Based Algorithm For Detecting Cuts and Classifying Scene Breaks”, *ACM Multimedia*, 1995.
11. D. Zhang, W. Qi and H.J. Zhang, “A New Shot Detection Algorithm”, *IEEE Pacific-Rim Conf on Multimedia*, 2001.