A Unified Framework Using Spatial Color Descriptor and Motion-based Post Refinement for Shot Boundary Detection†

Wei-Ta Chu[1], Wen-Huang Cheng[2], Sheng-Fang He[1],
Chia-Wei Wang[1], and Ja-Ling Wu[1,2]

[1] Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
{wtchu, nacci, wjl}@cmlab.csie.ntu.edu.tw,
b89075@csie.ntu.edu.tw
[2] Graduate Institute of Networking and Multimedia,
National Taiwan University, Taiwan
wisley@cmlab.csie.ntu.edu.tw

Abstract. We propose a unified framework which combines a novel color representation, i.e. spatial color descriptors, and a post-refinement process to detect various types of shot boundaries, including abrupt shot changes, flashlights, dissolves, fade-ins and fade-outs. The spatial color descriptor involving color adjacency and color vector angle histograms incorporates spatial information into color representation and provides robust performance in shot boundary detection. Moreover, a motionbased post-refinement process is developed to effectively eliminate false positives in gradual transition detection, where rapid camera motion or object movement may lead to performance degradation. Experimental results show that these two techniques are integrated seamlessly to give satisfactory performance and present the robustness of spatial color descriptors.

1 Introduction

The development of shot boundary detection algorithms has attracted a large amount of attention in the last decade. There is a rich literature of approaches for detecting video shot boundaries based on color histograms [2], edge pixels [3], motion vectors, and entropy metrics [4]. Although many approaches provide satisfactory results in general cases, few methods are robust to significant appearance changes caused by large-scale object movement or camera motion. One of the solutions to this problem is to design a representation method that takes spatial information into account.

Lee et al. [1] proposed a spatial color descriptor to effectively describe the color distributions and spatial information of video frames. In HLS color space, spatial color descriptors use the metric of color vector angle that is insensitive to variations in intensity, yet sensitive to differences in hue and saturation. When shape or

† Part of the work presented in this paper was published in the fifth Pacific-Rim Conference on Multimedia, Nov. 30 – Dec. 3, 2004.

appearance changes, the color pairs at the color edges mostly remain unchanged. Therefore, pixels in a video frame are first classified as either edge or smooth ones and then represented by two color histograms. The proposed color adjacency and color vector angle histograms convey this frame's characteristic. This technique provides robustness to substantial appearance changes and is suitable to be used in image retrieval and video segmentation.

We exploit spatial color descriptors to detect some commonly used shot boundary effects, such as flashlights, abrupt cuts, dissolves, fade-ins and fade-outs. A post-refinement process based on motion analysis is also developed and combined to the framework for eliminating the false alarms caused by rapid camera motion or object movement. This integrated approach is examined by several types of videos and demonstrates its effectiveness on shot boundary detection.

This paper is structured as follows. An overview of spatial color descriptors is stated in Section 2. In Section 3, we describe the proposed framework which hierarchically integrates spatial color descriptors and motion analysis techniques to detect various types of shot boundaries. Section 4 shows the experimental results, and the concluding remarks are given in Section 5.

2 An Overview of Spatial Color Descriptor

Two problems exist in the conventional histogram-based color descriptors. The first one is the lack of spatial information, and the second one is that similar colors are treated as dissimilar because of the uniform quantization of each color axis [1]. To solve these problems, two types of color histograms, i.e. color adjacency histogram for describing edge pixels and color vector angle histogram for describing smooth pixels, are constructed to characterize video frames effectively.

Pixels are classified as edge or smooth pixels based on color vector angle first. A 3 × 3 window is applied to every pixel of a video frame, where the center pixel and neighboring pixels making the maximum color vector angle are used to detect a color edge. If the center pixel in a window is an edge pixel, the global distribution of the color pairs around the edges is represented by a color adjacency histogram based on colors nonuniformly quantized in HLS color space. On the other hand, if the center pixel is a smooth pixel, the color distribution is represented by a color vector angle histogram. The overall distance measure of two successive video frames is represented as

$$d_i = D(i,i+1) = \alpha \times D_{adj}(i,i+1) + \beta \times D_{vec}(i,i+1) \qquad (1)$$

where $D_{adj}(i, i+1)$ and $D_{vec}(i, i+1)$ are distance values (differences of normalized bin values) of color adjacency and color vector angle histograms between frame i and $i + 1$, respectively. α and β are scalars for adjusting the weights of two histograms for different genres of videos. In the experiments present in this paper, α and β are both set as 0.5. For video cut detection, if the distance value is larger than a pre-defined threshold, a shot boundary candidate is declared.

3 The Proposed Framework

To robustly address various shot boundary detection issues, we develop a framework which integrates spatial color descriptors and post-refinement techniques, as shown in Fig. 1. After constructing two color histograms, the distance values of successive frames are calculated. To avoid the false alarms caused by flashlights, they are first detected and eliminated before conducting the abrupt cut detection. After detecting abrupt cut, the gradual transition detection process is applied within the range between two abrupt boundaries. Finally, by taking advantage of motion information, a post-refinement process is developed to eliminate some false positives caused by large-scale object movement or camera motion.

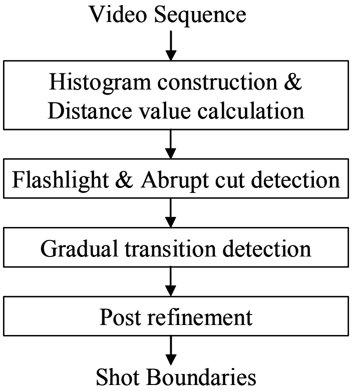
Video Sequence
Histogram construction & Distance value calculation
Flashlight & Abrupt cut detection
Gradual transition detection
Post refinement
Shot Boundaries

Fig. 1. The proposed framework for shot boundary detection

3.1 Flashlight and Abrupt Cut Detection

We examine a video sequence by applying a sliding window that spans m frames. The distance values between every two frames, i.e. d_i, d_{i+1}, …, d_{i+m-1}, are used to characterize the behavior of this video segment. If the distance value of two successive frames is larger than a threshold, the frame is declared as a shot boundary candidate. Unfortunately, this simple rule often falsely detects shot boundaries when flashlights occur, which greatly change the luminance of video frames and increase the distance value abruptly.

According to our observation, flashlights often last only one or two frames, and the frames neighboring to a flashlight would have similar color layouts. Therefore, we can detect flashlights by comparing the neighbors of the frame with exploding distance value. The detection rule is defined as follows.

```
If $d_{i+k} > \varepsilon$ for $1 \leqq k \leqq m$-2
  If there exists an $l$, $1 \leqq l \leqq 4$ such that
    $d' = D(i+k-l, i+k+l) < \varepsilon$
    then frame $i+k$ is a flash light
    otherwise frame $i+k$ is an abrupt shot boundary
```

$D(.)$ is the distance value defined in (1) between any two frames, and ε is a predetermined threshold for detecting abrupt discontinuity. In the experiments, the threshold is defined fixedly in the same type of videos without significantly changing the detection performance.

3.2 Gradual Transition Detection

The gradual transitions we considered are dissolves, fade-ins and fade-outs. Unlike abrupt cuts, comparison based on successive frames will not be useful for gradual transition detection because distance values are small during transition [5]. One alternative is to consider local edge information over a series of video frames [3] and match the change patterns of various gradual transitions. However, this method often leads to too many false positives and is not reliable when rapid camera motion or object movement occurs.

In the proposed framework, after detecting abrupt cuts, a gradual transition detection process is applied within the range between two cuts to further explore the structure of this video segment. We exploit the global edge information which is conveyed by the color adjacency histograms. For each frame i, two edge-change values, i.e. edge-increasing value ($E_{i,inc}$) and edge-decreasing value ($E_{i,dec}$), are considered as the metrics for gradual transition detection.

$$E_{i,inc} = \sum_{k=1}^{n} (H_{i,k} - H_{i-1,k}) \quad \text{if} \quad H_{i,k} > H_{i-1,k} \tag{2}$$

$$E_{i,dec} = \sum_{k=1}^{n} (H_{i-1,k} - H_{i,k}) \quad \text{otherwise} \tag{3}$$

where $H_{i,k}$ is the value of the k-th bin of the color adjacency histogram, and n is the total bin number. Note that different gradual transitions would have different edge change patterns. When a fade-in occurs, the value of edge-increasing will show a peak, while the edge-decreasing value remains smooth. In the case of a dissolve, both edge-increasing and edge-decreasing values would reflect the behavior of edge changes. Fig. 2 shows an example of the curve of edge-increasing values. By using edge change metrics, almost all abrupt cuts have sharp and great-scale peaks. Comparing to the case of abrupt shot change, gradual transitions have smaller change values but are still easily to be distinguished from the frames without shot changes.

An approach based on mean filter is applied to gradual transition detection. Assume that frames i and j ($i < j$) are declared as abrupt cuts, the rule for detecting fade-ins is defined as follows.

$$E_{mean\_inc} = mean(E_{i,inc}, E_{i+1,inc}, ..., E_{j,inc})$$
$$\text{if } (E_{k+l,inc} / E_{mean\_inc}) > \varepsilon_g \text{ for } 1 \le l \le R \text{ and } i \le k \le j - R \quad (4)$$
$$E_{k+l} \text{ are frames with fade-ins}$$

R denotes the width of the sliding window we examined for detecting fade-ins. It is set as 4 in the experiments. ε_g is a pre-defined threshold for detecting fadeins. Similarly, the rules for detecting fade-outs and dissolves are defined through considering edge-decreasing values or the combination of two edge change information.

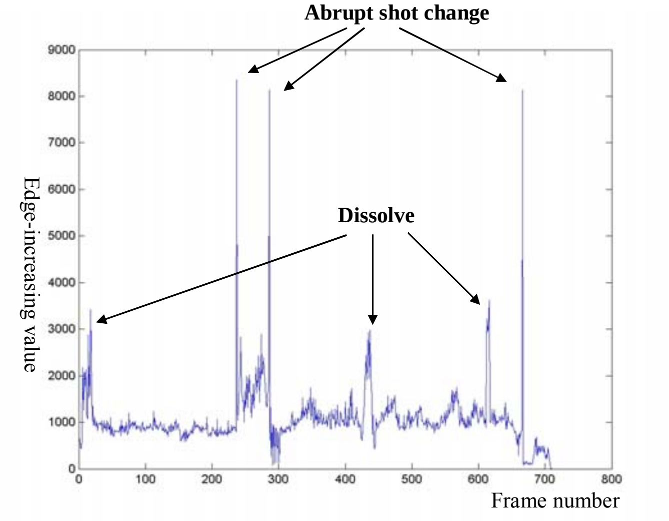

Fig. 2. The curve of edge-increasing values

3.3 Post Refinement by Motion Analysis

Although the spatial color descriptors and edge-based metrics can effectively characterize the behaviors of gradual transitions, some false alarms still exist when extreme camera motion takes place. Therefore, some post-refinement techniques based on motion, priori information, or statistical distributions [7, 8] are proposed to eliminate possible false positives.

In the proposed framework, we adopt a motion-based technique which analyzes motion through spatio-temporal slices processing [9]. The local orientations of temporal slices are estimated by the structure tensor [6]. By modeling the trajectories in tensor histograms, this technique is capable of detecting camera motion so that some false positives can be eliminated.

According to [9], the tensor histograms for horizontal slice with dimension (*x,t*) and vertical slice with dimension (*y,t*) are computed. Let $\phi(x,t)|_{y=i}$ and $c(x,t)|_{y=i}$ denote the local orientation and the associated certainty value of a pixel at a horizontal slice in which $y = i$. A 2-D tensor histogram $\mathrm{M}(\phi,t)$ of this video frame in (x,t) dimension is expressed as

$$\mathrm{M}(\hat{\phi},t) = \begin{cases} \sum_i \sum_x \sum_t c(x,t)_{|y=i} & \text{if } \phi(x,t)_{|y=i} = \hat{\phi}, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

which means that each pixel in slices votes for the bin $\phi(x,t)$ with its certainty value c(c=[0,1]). After normalizing by the frame size m $\times$ n, the resulting histogram with associated confidence value is represented as

$$C = \frac{1}{T \times m \times n} \sum_{\phi} \sum_{t} \mathrm{M}(\phi,t) \tag{6}$$

where T is the temporal duration of the video sequence. Detailed descriptions about structure tensor please refer to [9]. In the proposed framework, we detect camera pan and tilt via analyzing the motion in horizontal and vertical slices, respectively. Given a 2-D tensor histogram $\mathrm{M}(\phi,t)$, the tensor orientation ϕ is nonuniformly quantized into three bins, where

$$\Phi_1 = [-90^\circ, -5^\circ),\ \ \Phi_2 = [-5^\circ, -5^\circ],\ \ \Phi_3 = (5^\circ, 90^\circ].$$

The scheme quantifies motion information based on its intensity and direction. Φ_1 and Φ_3 represent intense motion, and Φ_2 represents no or slight motion. The normalized 1-D motion histogram N is computed by

$$\mathrm{N}(\Phi_k) = \frac{\sum_{\phi_i \in \Phi_k} \sum_t \mathrm{M}(\phi_i,t)}{\sum_{j=1}^{3} \mathrm{N}(\Phi_j)} \tag{7}$$

Finally, for every three successive frames, they are declared with a camera pan if the following criteria are satisfied in horizontal slices.

$$\begin{aligned} &(\mathrm{N}_{k,k+1,k+2}(\Phi_1) > \varepsilon_N) \wedge (\mathrm{N}_{k,k+1,k+2}(\Phi_3) < \varepsilon_N), \\ &(\mathrm{N}_{k,k+1,k+2}(\Phi_1) < \varepsilon_N) \wedge (\mathrm{N}_{k,k+1,k+2}(\Phi_3) > \varepsilon_N), \end{aligned} \tag{8}$$

where k is the frame index, and ε_N is a threshold defined empirically. Similar rules are defined for camera tilt by analyzing vertical slices. Through these processes, the video frames which are declared with both gradual transition and camera motion are discarded from the shot boundary candidates.

4 Experimental Results

We evaluate the proposed framework by using twenty test sequences that belong to four different program categories: news, movies, sports and commercials. They are recorded from TV broadcasts or extracted from MPEG-7 test corpus. Note that these sequences are carefully selected so that they contain many special effects or significant object/camera motions that often cause detection errors. For example, there are many editing effects and dazzling spotlights in selected commercials. The events of camera motion and players walking through screen occur frequently in sports games. The thresholds used in steps described in Section 3 are fixedly defined for different categories of videos without greatly degrading the detection performance. Moreover, to compare the proposed approach with conventional color- and edge-based method, the same test sequences are also applied in the algorithm presented in [3].

Table 1 shows the summary of shot detection results. In general, satisfactory performance could be achieved for different categories of videos. The detection results before and after post-refinement are listed separately to demonstrate the effectiveness of the refinement process. In the gradual transition detection, the refinement process especially shows its effectiveness in sports and commercial sequences because more rapid camera motion and object movement are detected and eliminated from the shot boundary candidates. Overall, the proposed framework provides 86.56% recall rate and 97.64% precision rate in abrupt cut detection and almost 60% recall and 50% precision rate in gradual transition detection.

Meanwhile, we found that detection accuracy degrades in some cases. Because the spatial color descriptors are based on HLS color space, the color vector angle between two colors with very low or very high intensity would vary significantly even if the Euclidean distance between them is small [1]. That's why the performance of gradual transition in some news and movies sequences is lower than others. This problem can be solved by considering Euclidean distance and color vector angle integrally or slightly modifying the representation of HLS color space.

Table 2 shows the performance comparison between the proposed framework and conventional approach [3]. Only the results of abrupt cut detection are listed because gradual-transition detection was not completely implemented in [3]. Although the conventional approach provides acceptable recall rate in different kinds of videos, it has bad precision performance when there are significant motions in sports video programs. This result shows the reliability of the proposed framework, which takes spatial information and motion-based refinement into account. The proposed approach generally has better precision but worse recall rate. In the current framework, the weights of color vector angle and adjacency histograms are not sedulously adjusted for each video sequence to achieve the best performance. They actually could be assigned by the user to meet different performance requirements, such as higher recall rate with slight degradation in precision.

Table 1. Performance of shot boundary detection

Video Frames	Cut (Correct, False)	Gradual (Correct, False)	Recall/Precision (Cut)	Recall/Precision (Gradual)
News (31015)	201 (170,1)	25 (12,16)	84.57/99.42	48/42.86
-after refinement	201 (170,1)	25 (20,27)	84.57/99.42	80/42.55
Movie (32123)	312 (265,5)	29 (11,46)	84.94/98.15	37.93/19.30
-after refinement	312 (265,5)	29 (10,29)	84.94/98.15	34.48/25.64
Sports (30520)	197 (177,10)	37 (26,23)	89.84/94.65	70.27/53.06
-after refinement	197 (177,10)	37 (24,10)	89.84/94.65	64.86/70.59
Commercial	101 (90,1)	9 (3,3)	89.11/98.9	33.33/50
(5677) -after refinement	101 (90,1)	9 (6,4)	89.11/98.9	66.67/60
Total (with ref.)	811 (702,17)	100 (49,58)	86.56/97.64	59.32/50.69

Table 2. Comparison of abrupt cut detection between (a) the proposed framework and (b) conventional approach

Video	Recall (a)	Precision (a)	Recall (b)	Precision (b)
News	84.57	99.42	95.6	91.6
Movies	84.94	98.15	99.01	85.71
Sports	89.84	94.65	85.83	39.39
Commercial	89.11	98.9	91.09	92

5 Conclusion

We have presented a framework which integrates spatial color descriptors and motion-based post-refinement techniques to detect various types of shot boundaries. The spatial color descriptors effectively represent the adjacency between colors in video frames and provide robustness to substantial appearance changes. The post-refinement process which exploits structure tensor to detect camera motion is seamlessly combined to improve the detection accuracy of gradual transition. The evaluation results show that this approach provides satisfactory performance in different kinds of videos and is robust to rapid motion and dazzling spotlights. Future work may include improving the performance of motionanalysis and conquering the limitation of spatial color descriptors described in Section 4.

Acknowledgement

This work was partially supported by the CIET-NTU(MOE) and National Science Council of R.O.C. under the contract No. NSC93-2622-E-002-033, NSC93-2752-E-002-006-PAE, NSC93-2213-E-002-006.

References

[1] H.Y. Lee, H.K. Lee, and Y.H. Ha, "Spatial Color Descriptor for Image Retrieval and Video Segmentation," IEEE Transactions on Multimedia (2003), Vol. 5, No. 3, 358-367.

[2] U. Gargi, R. Kasturi, and S.H. Strayer, "Performance Characterization of Video-Shot-Change Detection Methods," IEEE Transactions on Circuits and Systems for Video Technology (2000), Vol. 10, No. 1, 1-13.

[3] R. Lienhart, "Comparison of Automatic Shot Boundary Detection Algorithms," SPIE Storage and Retrieval for Still Image and Video Databases VII (1999), Vol. 3656, 290-301.

[4] Z. Cernekova, C. Nikou, and I. Pitas, "Shot Detection in Video Sequences Using Entropy-Based Metrics," Proceedings of International Conference on Image Processing (2002), Vol. 3, 421-424.

[5] B.L. Yeo and B. Liu, "Rapid Scene Analysis on Compressed Video," IEEE Transactions on Circuits and Systems for Video Technology (1995), Vol. 5, No. 6, 533-544.

[6] G.H. Granlund and H. Knutsson, "Signal Processing for Computer Vision," Norwell, MA: Kluwer (1995).

[7] A. Hanjalic, "Shot-Boundary Detection: Unraveled and Resolved?" IEEE Transactions on Circuits and Systems for Video Technology (2002), Vol. 12, No. 2, 90-105.

[8] H. Lu, and Y.P. Tan, "An Effective Post-Refinement Method for Shot Boundary Detection," Proceedings of International Conference on Image Processing (2003), Vol. 2, 1013-1016.

[9] C.W. Ngo, T.C. Pong, and H.J. Zhang, "Motion Analysis and Segmentation Through Spatio-Temporal Slices Processing," IEEE Transactions on Image Processing (2003), Vol. 12, No. 3, 341-355.

Biography

▲Name: Wei-Ta Chu

Address: R501, CSIE Building, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan

Education & Work experience: Wei-Ta Chu received the B.S. and M.S. degrees in Computer Science and Information Engineering from National Chi Nan University in Nantou, Taiwan, in 2000 and 2002. He is currently pursuing his Ph.D. degree in the Department of Computer Science and Information Engineering, National Taiwan University, Taiwan. As he is in the Communication and Multimedia Laboratory, his research interests include digital content analysis, multimedia indexing, digital signal process, and pattern recognition.

Tel: +886-2-23625336 ext 501

E-mail: wtchu@cmlab.csie.ntu.edu.tw

Other information:

▲ Name: Wen-Huang Cheng

Address: R501, CSIE Building, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan

Education & Work experience: Wen-Huang Cheng received the B.S. and M.S. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 2002 and 2004, respectively, where he is currently pursuing the Ph.D. degree in the Graduate Institute of Networking and Multimedia. His research interest includes multimedia data management and analysis.

Tel: +886-2-23625336 ext 501

E-mail: wisely@cmlab.csie.ntu.edu.tw

Other information:

▲ Name: Sheng-Fang He

Address: R506, CSIE Building, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan

Education & Work experience: Sheng-Fang He is a college student of department of computer science, National Taiwan University.

Tel: +886-2-23625336 ext 506

E-mail: b89075@csie.ntu.edu.tw

Other information:

▲ Name: Chia-Wei Wang

Address: R506, CSIE Building, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan

Education & Work experience: Chia-Wei Wang is a graduate student of department of computer science, National Taiwan University.

Tel: +886-2-23625336 ext 506

E-mail: nacci@cmlab.csie.ntu.edu.tw

Other information:

▲ Name: Ja-Ling Wu

Address: R415, CSIE Building, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan

Education & Work experience: Ja-Ling Wu received the B.S. degree in electronic engineering from TamKang University, Tamshoei, Taiwan, R.O.C., in 1979, and the M.S. and Ph.D degrees in electrical engineering from Tatung Institue of Technology, Taipei, Taiwan, in 1981 and 1986. Since 1987, he has been with the Department of Computer Science and Information Engineering, National Taiwan University, where he is presently a Professor. He has published more than 200 journal and conference papers. His research interests include algorithm design for DSP, data compression, digital watermarking and multimedia systems. Prof. Wu was the recipient of the Excellent Research Award from NSC, Taiwan, in 1999, 2001 and 2004.

Tel: +886-2-23625336 ext 415

E-mail: wjl@cmlab.csie.ntu.edu.tw

Other information: