Journal of Information Hiding and Multimedia Signal Processing ©2015 ISSN 2073-4212
Ubiquitous International Volume **6**, Number **3**, May **2015**

Multi-scale Shot Segmentation Based on Weighted Subregion Color Histogram

Yan-Chao Xing, En-qing Sun and Yang Lu
College of Communication and Electronic Engineering
Qingdao Technological University
Qingdao, 266033, P. R. China.
xingyanchao@yeah.net

Zhe-Ming Lu*
School of Aeronautics and Astronautics
Zhejiang University
Hangzhou, 310027, P.R.China
zheminglu@zju.edu.cn

Received August, 2014; revised February, 2015
(Communicated by Zhe-Ming Lu)

Abstract. *Shot segmentation algorithms must meet practical requirements on speed and accuracy, where video feature extraction and frame difference calculation are key issues. Subregion color histograms combine color and spatial information and can balance between effectiveness and robustness. Through analyzing the intra-shot and inter-shot video content characteristics, different weights were set to different subregions to better reflect the video foreground variations. The weighted sum of subregional differences outperformed the non-weighted version remarkably. Shot segmentation presents temporally multi-scale characteristics, both abrupt and gradual shot transitions could be detected with multi-scale analysis, together with confidence degree of the resulting shot boundaries. The selection for subregion sizes and temporal intervals was discussed. A fast multiple-level subregion histogram algorithm was presented. Experiments show promising detection effects and fast processing speed.*

Keywords: Video shot segmentation, Weighted subregion color histogram, Temporarily multi-scale analysis, Shot boundary detection.

1. **Introduction.** As a prerequisite for many content-based video applications, shot segmentation can produce the basic processing units for video abstraction, content-based video retrieval, video clip comparison, etc [1]. From the perspective of pattern recognition, shot segmentation has three basic technical issues: choose suitable video features to represent the video content; define effective measurements to reflect the difference between video frames; classify the above measurement curves into cut transition, kinds of gradual transitions, or none shot transitions [2]. There have been various techniques presented for video content representation, from simple pixel-based [3], histogram-based [4], edge-based and motion-based algorithms, to complicated block histogram method [5], spatio-temporal slices [6], SVD and pattern matching method [7], etc. There are also some techniques for compressed video or 3D video shot segmentation.

To make the system practical, shot segmentation must be robust and efficient. Among video features mentioned above, the subregion/block color histograms could balance well between invariance and sensitivity, they are also computationally efficient and could act as a practical measurement. But the differences between frames within one shot or between different shots are caused by camera actions and/or video object movements. So directly summing up all subregional differences could not represent shot transition conditions well. More investigations should be performed on the essence of subregional differences.

Simple thresholding the difference between adjacent frames makes no use of contextual information, so it could only detect abrupt shot transitions. There are some novels algorithms making use of contextual information which are summarized by Cooper [8]. In fact, shot transitions present obvious multi-scale characteristics in temporal domain. Different types of shot transitions with various durations could be observed on different time intervals. And robust gradual shot transitions must be stable across several time scales/intervals. With this observation, temporally multi-scale analysis could be taken into account for shot segmentation. In this paper, subregion color histogram was taken as the basic frame feature. Then subregion differences are analyzed from the perspective of camera action and visual object movements. With different weights to these subregions, the frame difference is defined. Through a joint analysis of multi-time-interval frame difference curves, the reliability and accuracy of the shot segmentation was improved, and the shot boundary confidence degrees were given too.

2. **Subregion Color Histogram.** Pixel-based features are too sensitive, while histogram is a popular alternative. Experiments showed that usually the simple histogram feature could achieve satisfactory results, which could not be outperformed by some complicated features, e.g. edge [9]. Since histogram does not contain the spatial distribution information of different colors, it's robust to local or small global movements, but with poor expressive ability to distinguish shots within a same scene. Subregion color histogram is a better tradeoff between pixel and global color histogram, where each frame is divided into several subregions, the histograms for all subregions will be extracted and compared with corresponding ones of other frames.

2.1. **Subregion Size vs. Influence of Movements on Histogram.** Subregion color histogram can keep the primarily spatial color distribution information, and have some degree of tolerance to movement at the same time. With subregion color histogram, visual object movements won't cause too big frame difference values, and frame differences between adjacent shots won't be too small even they have similar global color histogram. These characteristics are all beneficial for shot segmentations. How to select suitable subregion sizes selection is a key issue. Too big size can't keep spatial distribution information effectively, may even degenerate to the traditional histogram difference methods. To small size may suffer from inter-frame movements. As shown in figure 1: suppose the windows size is W, inter-frame displacement is A, the overlapped area before and after the movement is:

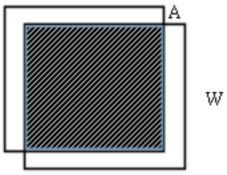


Figure 1. Overlapping area with displacement A

$$WA - (W - A)A = 2WA - A^2 \tag{1}$$

The proportion to the block area is:

$$\left(2WA - A^2\right)/W^2 = 2A/W - A^2/W^2 \approx 2A/W \tag{2}$$

For given value of A, since W>>A, the second item could be omitted. So the proportions of non-overlapping area are smaller with bigger W. In other words, the influence of movement on subregion histogram decreased with increasing W, so as the frame difference.

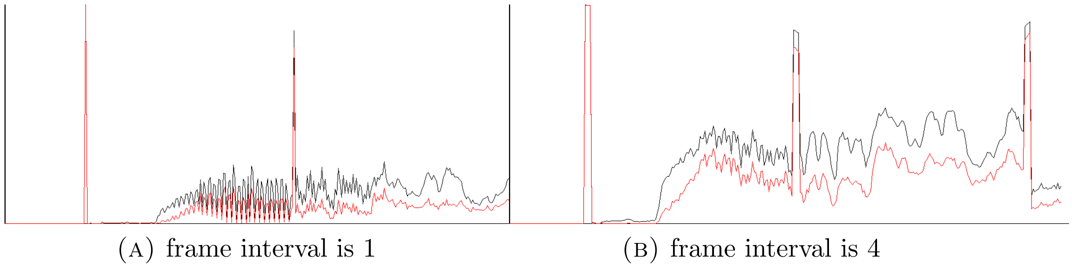
2.2. **Camera Actions vs. Visual Object Movements.** According to the definition, a shot is a series of continuous frames captured by a single camera action. The frame differences are mainly caused by two sources. One is the camera motion, e.g. tracking, zooming, rolling, etc. The other is video foreground content changes, including object movements, object appearing/disappearing, illumination changes, etc.

It's obvious that the intra-shot frame differences are caused by camera motions and foreground content changes. The camera usually moves gradually, and the resulting video content changes are smooth in both spatial and temporal domain. The frame differences of such type are global and smooth. The foreground changes are usually local, which may be smooth or abrupt. The intra-shot frame differences are mainly caused by visual object changes, since they belong to different shots. Such types of frame differences are both global and abrupt for all subregions.

With the above observations, a weighted subregion color histogram comparison scheme was proposed as below.

2.3. **Weighted Subregion Color Histogram Comparison.** Inter-shot frame differences are caused by global and abrupt visual changes, while intra-shot differences by local visual changes and camera motions. We can suppress intra-shot differences to emphasize inter-shot ones. One simple but effective method is to assign small weights to top N subregions with biggest changes. After this, locally abrupt changes are suppressed. For intra-frame conditions, only smooth changes are left which are mainly caused by camera motions. For inter-frame conditions, some visual content changes are suppressed, but most of them are still kept.

YCrCb was chosen for color histogram. To reduce noise interference and storage requirements, the sampling intervals for were selected as 32, 16 and 16 respectively. Figure 2 shows frame difference curves based on non-weighted and weighted subregion color histogram comparison. It shows that intra-shot frame differences are suppressed successfully, making the inter-shot frame differences more prominent.



(A) frame interval is 1 (B) frame interval is 4

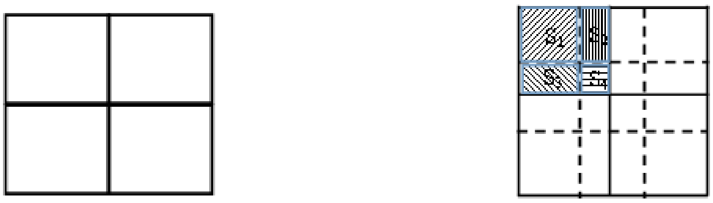
Figure 2. Non-weighted (black) vs. Weighted (red) frame difference curves

For temporally multi-scale analysis in section 4, different temporal intervals are used for frame difference calculation. Obviously, for the same movement velocity, lager intervals lead to larger displacement, the subregion sizes should be increased correspondingly to achieve the same tolerance for movements. This requires calculating multi-level subregion color histograms at the same time. In order to speed up computational, a fast algorithm of multi-level block histogram was designed as below.

3. **A Fast Multiple-level Histogram Calculation Algorithm.** First, use the minimal subregion size. For the larger subregion size, the subregion histograms can be simply calculated from previous histograms, without repeating the same calculation process. Fig 3.a shows the commonly used 1X1, 2X2, 4X4 multi-level partitioning methods. After calculating the 4X4 block histograms, one 2X2 subregion corresponds to 4 4X4 smaller subregions. According to the definition of histogram, the new histogram is just the mean of the 4 smaller subregion histograms. If it is like those in figure 3.b, from 3X3 to 2X2, the areas of corresponding multiple small subregions are different. According to the definition of histogram, we can also directly calculate the 2X2 block histogram easily:

$$H\left(Y^k, Cr^k, Cb^k\right) = \left(\sum_{i=1}^{4} S_i \cdot H\left(Y_i^k, Cr_i^k, Cb_i^k\right)\right) / \sum_{i=1}^{4} S_i \tag{3}$$

Where k is the index, Si is the corresponding small partition area. Usually, subregion areas in adjacent levels are less than 4 times. Most commonly used are those like Fig. 3.a. Since the calculation cost is much lower than the direct calculation, the multiple-level block histogram-based method is feasible and practical.



(a) Multi-level integer multiple partitions

(b) Multi-level non-integer multiple partitions

Figure 3. Diagrams for Multi-level partitioning histograms

4. **Temporally Multi-scale Analysis.** For cut or abrupt shot changes, where the current and the previous frames belong to different shots, the shot boundaries could be reliably located using frame difference curves, here the frame difference is between adjacent frames. But for gradual shot transitions, e.g. fading in and fading out, the process of changing from one shot into another is gradual, it may take more than several or dozens of frames. During the transition, the changes are gradual and the frame differences between adjacent frames are not very abrupt, so it is difficult to discover the shot transition using adjacent frame differences.

There are many types of gradual shot transitions, together with many new emerging types, especially computer animation special effects. To detect various gradual shot changes, we need to use some common features. Considering the temporally multi-scale characteristics: the gradual shot transitions on a small interval need a lot of frames to complete, but at some bigger interval, the gradual transitions will become “abrupt”. Since transiting durations of the shot change vary greatly, fixed time intervals cannot be suitable for all scenarios, multi-scale analysis must be considered. The multi-scale analysis can effectively detect the cut changes too.

For a specific video clips, i.e. a bit of China, Fig.4 shows the characteristics of multi-scale of gradual and cut shot transitions. Here the frame intervals used are {1, 2, 4, 8,

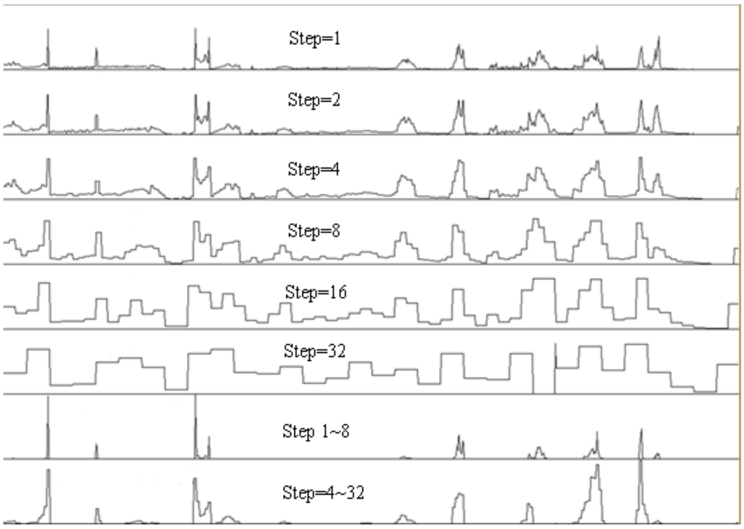


FIGURE 4. Multi-scale inter-frame difference curves for 1024 frames

16, 32}. To reduce computational cost, frame differences at larger scale are calculated at bigger intervals. Compared with direct frame difference, the total amount of calculation is $\sum_{i=0}^{5} \frac{1}{2^i} = 1.97$times, so the total computational cost is acceptable.

The cut changes are similar to unit step functions, for which reliable peaks of frame difference could be obtained with small intervals. For sudden interferences, e.g. flashlight or instantaneous blocking of the camera, small interval difference curve analysis may lead to false positives, while larger intervals could suppress such interferences. In order to achieve accurate boundary location and overcome the sudden disturbance at the same time, the frame differences of scales 1 to M are multiplied together, the result is shown as the "Step 1∼8" curve in figure 4. This curve has better peak-valley characteristics, ensuring the accuracy of the cut position. Numerical analysis is shown in table 1.

It is hard to detect the gradual shot changes with small time intervals. With the increase of interval steps, peaks corresponding to the gradual changes are gradually showing up. Stable gradual changes in successive time intervals are preferred, by multiplying several frame difference curves of adjacent scales, we get "Step 4 ∼ 32" curve in figure 4. With such curves, the gradual shot transitions could be detected more reliably. Numeric analysis is shown in table 1 as below.

Extracting curve peaks and valleys has been widely used in signal processing. The accuracy and reliability of the extraction results depends on the statistical characteristics of these curves, especially the means and the variances. If the standard deviation ratio to the mean is small, the curve's peak-valley characteristics is not obvious, the extraction may be hard and inaccurate. If the ratio is big, the extraction is easy and accurate because the peak-valley characteristics are very obvious. Table 1 lists the mean and variance of each scale (corresponding to the curves in figure 4). It is obvious that "Step 1∼8" and "Step 4∼32" curves have much better peak-valley characteristics.

The adapting thresholding method was used for final classification, where the threshold values are dynamically updated based on a sliding window of size 1024. Using the above curves, peaks could be located accurately, so as the shot segmentation results. Subregion sizes should match up with the time intervals. The basic principle is larger subregions should be used for large intervals. Specific values could be obtained from empirical data,

Table 1. Mean-Variance for difference frame curves in Figure 4

Step Length	Mean(m)	Variance(δ^2)	δ	δ/m
1	0.092759	0.018236	0.135043	1.45584795
2	0.152576	0.038592	0.196449	1.2875485
4	0.244578	0.071244	0.266916	1.091332826
8	0.371318	0.118606	0.344392	0.927485336
16	0.558961	0.189215	0.434988	0.77820814
32	0.73079	0.19443	0.440943	0.603378536
1∼8	0.034825	0.028027	0.167413	4.807264896
4∼32	0.21504	0.362163	0.6018	2.798549107

or automatically determined by analyzing video content. Our experiments used empirical values and achieved promising results.

There is a lot of fuzziness in the shot segmentation results, even for manually marked results. After shot segmentations, there are still many subsequent processing steps, such as key frame extraction, scene segmentation, etc. It's important to give out some confidence level for each shot boundary. According to the above discussion, multi-scale curve value can be used as the confidence degree for shot transitions.

5. **Analysis of Experimental Results.** The Internet Archive videos (IACC.2) for TRECVID 2013 were adopted, totally 30 videos in MPEG-1 format were chosen for tests. There were totally 9805 video transitions, 73.5% of them were cut transitions. First, basic pixel-based and histogram-based methods were compared to the proposed algorithm. For the first two methods, three fixed time intervals {4, 8, 16} were used. The results are shown in table 2, where $F1 = 2 * Rec * \Pr ec / \left(Rec + \Pr ec \right)$. The proposed algorithm got was the best result, especially for gradual transitions.

Table 2. Basic shot segment methods results

		Cuts			Gradual transitions		
		Recall	Precision	F1	Recall	Precision	F1
Histogram-based	Step=4	0.939	0.942	0.940	0.706	0.816	0.757
	Step=8	0.935	0.946	0.940	0.698	0.817	0.753
	Step=16	0.931	0.951	0.941	0.696	0.820	0.753
Pixel-based	Step=4	0.960	0.908	0.933	0.726	0.809	0.765
	Step=8	0.954	0.912	0.933	0.723	0.812	0.765
	Step=16	0.942	0.917	0.929	0.718	0.816	0.764
Algorithm of this paper		0.953	0.936	0.944	0.778	0.788	0.783

With the concept of multi-scale weighted subregion histogram, which was mainly for video representation, several classical methods were revised. The comparisons between revised and original versions were shown in table 3, including Kaabneh's [9], Gargi's [4], and Lu's [7]. And in Lu's method, only the first stage, e.g. candidate localization was touched. The revised versions had better recall rates, which means higher sensitivity, while the precision rates were only slightly decreased, especially for gradual transitions.

6. **Conclusion.** Subregion color histogram could balance the robustness and effectiveness for shot segmentation through adjusting the subregion number. Based on the analysis

Table 3. Shot segment results for revised classical methods

Algorithm		Cuts			Gradual transitions			All Transitions		
		Recall	Prcsn	F1	Recall	Prcsn	F1	Recall	Prcsn	F1
Kaabneh's	Original	0.930	0.941	0.935	0.701	0.831	0.760	0.872	0.916	0.893
	Revised	0.953	0.939	0.946	0.743	0.826	0.782	0.897	0.909	0.903
Gargi's	Original	0.946	0.913	0.929	0.760	0.780	0.770	0.899	0.880	0.889
	Revised	0.966	0.910	0.937	0.788	0.776	0.782	0.919	0.874	0.896
Lu's	Original	0.930	0.939	0.934	0.788	0.791	0.789	0.894	0.901	0.897
	Revised	0.964	0.935	0.949	0.805	0.786	0.795	0.922	0.896	0.908

of the sources for video content changes between frames, a weighted version was proposed to suppress local foreground changes. The resulting intra-shot frame differences were suppressed which makes the inter-shot ones more prominent. Shot transitions present temporally multi-scale characteristics. With multi-scale frame difference curves, we can achieve reliable and accurate host boundaries for gradual and cut shot transitions, and boundary confidence degree at the same time. To speed up, a fast multiple-level histogram calculation algorithm was presented. Without significant increase in computational complexity, compared with histogram difference of fixed step method and frame difference method, the algorithm could get better positioning accuracy and detection reliability. For each parameter in the algorithm of adaptive dynamic adjustment, further research is needed.

Acknowledgment. This work was partially supported by the National Natural Science Foundation of China under grant 61171150 and Zhejiang Provincial Natural Science Foundation of China under grant R1110006.

REFERENCES

[1] N. Dimitrova, H. J Zhang, B. Shahraray, I. Sezan, T. Huang and A. Zakhor, Applications of video content analysis and retrieval, *IEEE Multimedia*, vol. 9, no. 3, pp. 42-55, 2002.

[2] J. H. Yuan, H. Y. Wang, L. Xiao, W. J. Zheng, J. M Li, F. Z Lin and B. A. Zhang, A Formal Study of Shot Boundary Detection, *IEEE Trans. On Circuuits and Systems for Video Technology*, vol. 17, no. 2, pp. 168-186, 2007.

[3] H. J. Zhang, C. Y. Low and S. W Smoliar, Video parsing and browsing using compressed data, *Multimedia Tools Applications*, vol. 1, no. 1, pp. 89-111, 1995.

[4] U. Gargi, R. Kasturi and S. H. Strayer, Performance characterization of video-shot-change detection methods, *IEEE Trans. Circuits System Video Technology*, vol. 10, no. 1, pp. 1-3, 2000.

[5] J. Q. Yu, B. Tian and Y. Tang, Video Segmentation Based on Shot Boundary Coefficient, *Proc. of the 2nd Int'l Conference on Pervasive Computing and Applications*, pp. 630-635, 2007.

[6] C. W Ngo, T. C. Pong and R. T. Chin, Video partitioning by temporal slice coherency, *IEEE Trans. Circuits System Video Technology*, vol. 11, no. 8, pp. 941-953, 2001.

[7] Z. M. Lu and Y. Shi, Fast Video Shot Boundary Detection Based on SVD and Pattern Matching, *IEEE Trans. on Image Processing*, vol. 22, no. 13, pp. 5136-5145, 2013.

[8] M. Cooper, Video segmentation combining similarity analysis and classification *Proc. of the 12th annual ACM international conference on Multimedia*, pp. 252-255, 2004.

[9] K. Kaabneh, O. Alia, A. Suleiman and A. Abuirbaleh, Video Segmentation Via Dual Shot Boundary Detection (DSBD) *Proc. of 2nd Intl. Conference on Information and Communication Technologies*, vol. 1, pp.1530-1533, 2007.