Coarse-to-Fine Dissolve Detection based on Image Quality Assessment

Weigang Zhang[1,2], Chunxi Liu[3], Qingming Huang[3,4], Shuqiang Jiang[4], Wen Gao[1,5]

[1] School of Computer Science and Technology, Harbin Institute of Technology,
Harbin 150001, China
[2] School of Computer Science and Technology, Harbin Institute of Technology at Weihai,
Weihai 264209, China
[3] Graduate University of Chinese Academy of Sciences, Beijing 100190, China
[4] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[5] Institute ofDigital Media, Peking University, Beijing 100871, China

{wgzhang, cxliu, qmhuang, sqjiang, wgao}@jdl.ac.cn

Abstract. Although many approaches have been proposed for video shot boundary detection, dissolve detection remains an open issue. For a dissolve, we could find that the video frames reveal a "clarity–blur–clarity" visual pattern. Accordingly, the image quality in the dissolve also reveals a "high–low–high" pattern. Based on the above observation, in this paper a novel coarse-to-fine dissolve detection approach based on image quality assessment is presented. Firstly, the normalized variance autofocus function is employed to calculate the image quality value for its good performance and the image quality feature curve is obtained. The grooves on the curve, which are monotone decreasing to a local minimum and then are monotone increasing to a normal value, are detected by using a simple threshold-based method and deemed as dissolve candidates. After obtaining the coarse results, some refined features are extracted from these dissolve candidates and the final dissolve detection is accomplished with the help of the support vector machine based on a new dissolve length normalization method. The experimental results show that the proposed method is effective.

Keywords: dissolve detection, coarse-to-fine, image quality assessment, dissolve length normalization, shot boundary detection.

This work was supported in part by National Natural Science Foundation of China: 61025011, 60833006 and 61070108, and in part by Beijing Natural Science Foundation: 4092042.

1 Introduction

The last decade has witnessed the great advance of multimedia technology, the fast increase of the computer performance, and the significant improvement of the Internet, which led to the mass production and easily accessible of digital videos all over the world. However, when facing this huge amount of video information it is not easy for users to find their interested content. Therefore, there is a high demand for video content management techniques, including efficient video indexing, browsing and retrieving, *etc*. In the past few years, video content analysis attracted extensive attention of the researchers and many technical papers have been published. Among these video processing technologies, partitioning a video sequence into shots is deemed as the first step toward video-content analysis and content-based video browsing and retrieval [1]. A shot is a series of interrelated consecutive pictures taken contiguously by a single camera and representing a continuous action in time and space. Usually, a video consists of a series of shots.

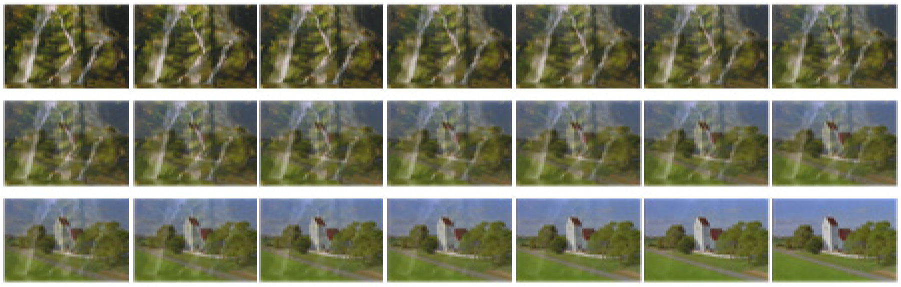

Fig. 1. A dissolve transition example

Generally speaking, there are two kinds of transition from one shot to another, which are hard cut and gradual transition. A hard cut is an instantaneous transition from one shot to the next. There are no transitional frames between the two consecutive shots. For the gradual transition, a video transition effect is added between the two consecutive shots to improve the visual experience. There are many gradual transition types including dissolve, wipe, fade out/in, *etc*. Among them, dissolve is the most popular gradual transition type. Dissolve is the video effect that the content of the first shot gradually disappears while the content of the second one becomes more and more visible, just like the video frames shown in Fig.1. Compared with the gradual transition detection, cut detection is relatively easy. Most of the presented techniques for cut detection could perform well and have a good precision [2]. However, for gradual shot transition, the more complex representations of the video content in the transitions bring the more trouble.

We mainly focus on detecting the most popular gradual transition type—dissolve. During the past few years, some algorithms have been proposed for dissolve detection. R. Lienhart[3] adopted several features, including edge change ratio (ECR) and edge-based contrast (EC), and applied some threshold selection strategies to detect dissolves. Su *et al.* [4] considered that object motion and camera motion are the main causes for error detection of dissolves. They built a nonlinear dissolve model and adopted the sliding window technology to improve the dissolve detection performance. Huang *et al.* [5] presented a dissolve detection approach based on contrast context histogram and local keypoint matching of video frames. B. Ionescu *et al.* [6] proposed a straightforward intensity-based dissolve method based on the amount of fading-out and fading-in pixels. Yuan *et al.* [7] conducted a formal study of the shot detection problem with a general formal framework, a comprehensive review of the existing approaches and a unified system based on graph partition model. Although many approaches have been proposed for dissolve detection, the reported results are not satisfactory. Until now, video dissolve transition is still an open issue.

In this paper, we attempt to detect the dissolve transition from the viewpoint of image quality assessment, and propose a coarse-to-fine approach for dissolve detection. In the coarse step, the image quality assessment based method is employed to detect the dissolve candidates. While in the fine step, the final dissolve detection is accomplished with the help of the Support Vector Machine (SVM). From Fig.1 we can easily find which frame belongs to the dissolve. We carefully checked many dissolves and found that all the video frames in the dissolve are mixed with the same two clear frames. One is the last frame of the previous shot, the other the first frame of the subsequent shot. The frames between the transitions are fused and blurred, and according to the human vision system, the quality of these images is relative low. The middle frame of a dissolve is usually the most blurred and its quality is the lowest. The starting and ending frames of a dissolve are relative clear and have high visual quality. In summary, the video frames in a dissolve transition reveal a "clarity–blur–clarity" pattern. Correspondingly, the image quality of the dissolve reveals a "high–low–high" pattern. If a shot transition shows the similar pattern, it could be considered as a dissolve candidate. In order to adopt the image quality assessment to detect the dissolve candidates, the first step is to select an appropriate evaluating function for image quality. Based on the image quality values, we could adopt a simple threshold based method to detect the dissolve candidates. Then, some new refined features are extracted from the dissolve candidates based on a dissolve length normalization method, and the support vector machine (SVM) is implemented to get the final dissolve detection results. The main contribution of the paper can be summarized as follows:

(1) We propose to detect dissolve transition from the viewpoint of the image quality analysis.

(2) We propose a novel coarse-to-fine approach for dissolve detection.

(3) We propose a novel dissolve length normalization approach to deal with the variable length problem of the dissolve transition.

The rest of this paper is organized as follows. Section 2 presents the selection of the evaluating functions for image quality assessment; section 3 provides the coarse-to-fine dissolve detection approach with SVM in detail; section 4 discuses the experimental results; section 5 concludes this paper.

2 Image Quality Assessment

The dissolve transition reveals a very clear visual pattern. If we can find a criterion by which could be indicated whether a frame belong to the dissolve or not, then the dissolve detection will become relatively easy. However, finding this criterion is as hard as dissolve detection. In this paper, instead of finding the ideal criterion, we try to measure the image quality of the video frames with the assumption that the image quality of the frames in a dissolve will be relatively low.

Actually, image quality assessment has been a hot research topic for a long time [8]. The situation we encounter is a typical no-reference image quality assessment problem. The dissolve reveals a “clarity–blur–clarity” pattern. In order to evaluate the image quality of the video frames, we adopt an autofocus function, which is used frequently for digital image blur measure. Autofocus functions are usually used to measure the focusing performance of micro-imaging systems. If the obtained images from these systems are blurred, they will output low values, which indicate that the quality of these images is low. Conversely, high autofocus function value indicates that the image is clear and the quality is high. As we known the frames in a dissolve transition show a “clarity–blur–clarity” pattern. Therefore, the autofocus function values should reveal a corresponding “high–low–high” quality pattern.

There are many autofocus functions available [9,10], such as Brenner gradient, Tenenbaum gradient (Tenengrad), energy Laplace and normalized variance, *etc.* A.Santos *et al.*[9] made a lot of comparative experiments on 13 autofocus functions. According to their qualitative evaluation, relative (semiquantitative) evaluation and quantitative absolute evaluation, they draw the conclusion that among these autofocus functions the normalized variance could achieve good performance. Sun *et al.*[10] also made a comprehensive comparison study of 18 focus algorithms in which a total of 139,000 microscope images were analyzed. The experimental results show that the normalized variance function performs the best and is the optimal function to evaluate the blur of images. Therefore, we adopt it to calculate the quality value for each frame. The calculation of the normalized variance function is as below:

$$f(i)=\frac{1}{H\times W\times \mu_i}\sum_{x=1}^{W}\sum_{y=1}^{H}\left(I_i(x,y)-\mu_i\right)^2 \tag{1}$$

where H and W are the height and width of the video frame respectively. $I_i(x,y)$ is the gray value of the pixel (x,y) in the grayscale of the original color video frame i. μ_i is the average gray value of the frame and is calculated as below:

$$\mu_i=\frac{1}{H\times W}\sum_{x=1}^{W}\sum_{y=1}^{H}I_i(x,y) \tag{2}$$

By using equation (1), a normalized image quality curve of a short video clip is shown in Fig.2. This clip consists of six dissolves. From the figure, we could see that the groove patterns of these dissolves are rather clear.

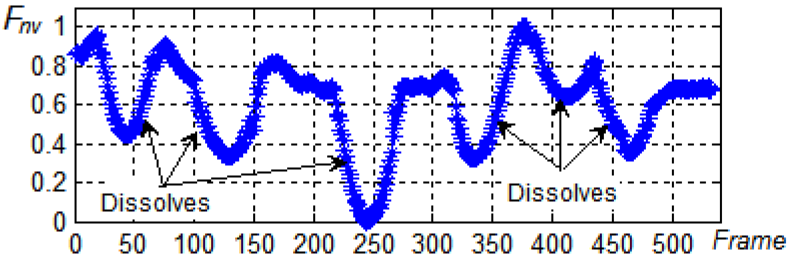

Fig. 2. The groove patterns on the normalized image quality curve of a short video clip which has six dissolves by using the normalized variance autofocus function

3 Coarse-to-Fine Dissolve Detection

In this section, we describe the proposed coarse-to-fine approach for dissolve detection. In the coarse step, a threshold method based on image quality assessment is employed to detect the dissolve candidates. In the fine step, the final dissolve detection is accomplished with SVM.

3.1 Dissolve Candidate Detection

Based on the normalized variance function we can obtain the image quality curve. In this section, we will detect the dissolve candidates based on the quality curve. As can be seen from Fig.2, the image quality of different shot frames is not at the same level. Therefore, directly using a threshold to detect the dissolve is not feasible. In this section, we try to use the gradient of the normalized variance to normalize the image quality value. There are many way for gradient calculation. In the proposed method, the gradient calculation of the normalized variance is as follow:

$$\partial f(i) / \partial i = f(i+\beta) - f(i-\beta) \tag{3}$$

where $f(i)$ is calculated by equation (1). β is a parameter to amplify the gradient value and in this paper β is set as 4 according to experiments. Fig.3 shows the gradient curve of the image quality values in Fig.2.

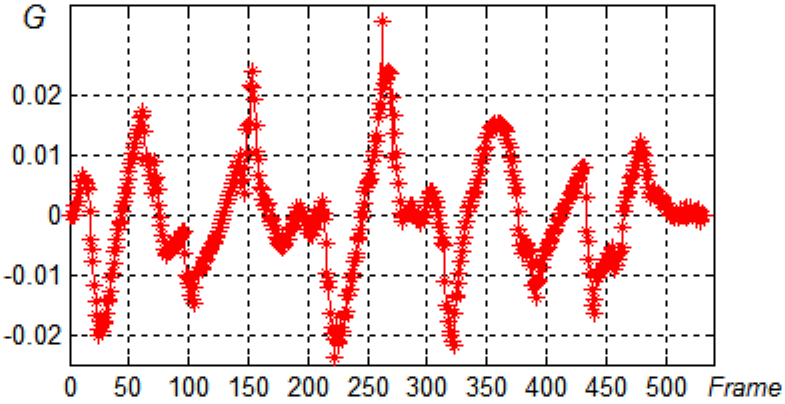

Fig.3. The gradient curve based on the original image quality values

After gradient calculation, the dissolve groove pattern changes into another interesting pattern. The gradient value of the images in the same shot will become very small and near zero. However, for a dissolve the gradient value will first be smaller than zero. When the original normalized variance value reaches the minimum, the gradient value should be zero. After that, the gradient value increases and is bigger than zero. Thus, we use a heuristic method to detect the dissolve candidate patterns. First, two thresholds $\delta_1 > 0$ and $\delta_2 > 0$ are used to detect the gradient parts of dissolves. δ_1 and δ_2 are set small enough to catch all the dissolve patterns. Then, we label the curve into a step curve with -1, 0 and 1. The labeling rule is defined as below:

$$\begin{cases} D(i) = 1, & \partial f(i) / \partial i >= \delta_1; \\ D(i) = -1, & \partial f(i) / \partial i <= -\delta_2; \\ D(i) = 0, & otherwise. \end{cases} \tag{4}$$

The label results of the curve in Fig.3 are shown in Fig.4. After labeling, we search through the curve along the original video timeline and merge the part whose value is below zero with another part whose value is above zero according to their distance in the timeline. By comparing Fig.2 and Fig.4 we could know that almost all the dissolve patterns will be detected by the above method. The obtained dissolve candidates will be further analyzed to achieve the final detection results.

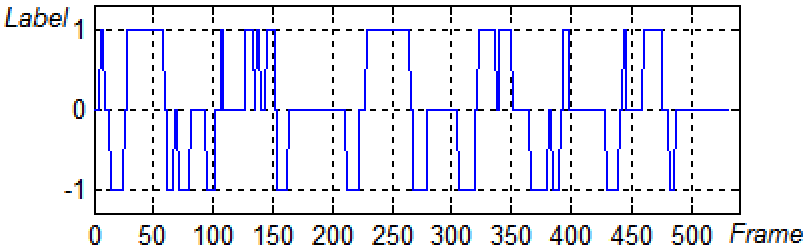

Fig.4. The label step curve based on the gradient curve of the original image quality values

3.2 Dissolve Length Normalization

In our view, one difficulty for dissolve detection is that the dissolve length is variable. On the other hand, many exiting machine-learning algorithms, such as the SVM, require the inputted features have the same length. Therefore, many methods utilize the multi-resolution approach [7], and build different models in different length scales to detect the dissolves. Another reason why they build so many models in different length scales is that there are no dissolve candidates in their approach and they have to slide different length windows along the timeline to detect the dissolves. Different from the existing dissolve detection approaches, we could obtain the dissolve candidates firstly. Then, a novel length normalization approach for dissolves could be employed to avoid the trouble of multi-scale modeling.

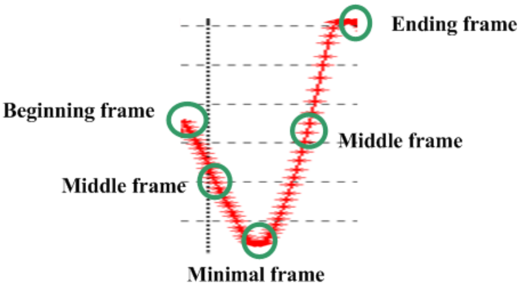

Fig.5. The length normalization rule of dissolves

After obtaining the dissolve candidates, we can get the beginning and ending frame for each candidate. As described on above, the image quality in the dissolve reveals a “high–low–high” groove pattern, and there is a minimal image quality value in each dissolve. We propose to use the beginning, ending and minimal frame of the dissolve candidate to help normalizing the lengths of dissolve candidates. After obtaining these three frames, we extract two other frames. One is the middle frame of the beginning and the minimal frame; another is the middle frame of the minimal and ending frames of the dissolve. The length normalization rule is shown in Fig.5. In all we use the extracted five frames to represent each dissolve and extract the refined features from these frames for further processing.

3.3 Dissolve Detection based on SVM

After normalizing the dissolve candidates, we extract some new refined features from each candidate, including the HSV histogram distance (10 features) and the mutual information (10 features) among the five frames, the original image quality of the five frames (5 features), and the first order difference of the quality value (4 features). Totally 29 features are extracted. These features are extracted for the reason that they can reveal the special pattern of the dissolve and are useful for dissolve detection. The calculation of the histogram distance is as bellow,

$$D(X,Y) = (\sum_{b=1}^{B} \left| H_X(b) - H_Y(b) \right|^p)^{1/p} \tag{5}$$

where $H_X(b)$ and $H_Y(b)$ denote the bth bin value of the normalized HSV histograms of the X and Y frame in the selected five frames. B is the total bin number of the histograms. The calculation of the mutual information is as in equation (6). $Entropy(X)$ and $Entropy(Y)$ represent the entropy of frame X and Y respectively. $Entropy(X,Y)$ represents the joint entropy of frame X and Y .

$$MI(X,Y) = Entropy(X) + Entropy(Y) - Entropy(X,Y) \tag{6}$$

$$Entropy(X) = -\sum_{x \in A_X} P_X(x) \log P_X(x)$$

$$Entropy(Y) = -\sum_{y \in A_Y} P_Y(y) \log P_Y(y)$$

$$Entropy(X,Y) = -\sum_{x \in A_X, y \in A_Y} P_{XY}(x,y) \log P_{XY}(x,y)$$

After obtaining these refined features, we fed them into the pre-trained SVM models to get the final dissolve detection results. SVM algorithm tries to classify different classes with the help of the support vector by maximizing the margin between these classes. It is a very effective method for classification. For more details about the SVM , refer to [11].

4 Experiments

Some videos in the TRECVID 2005 dataset are used to test the performance of the proposed method. Totally, there are 8 news videos from different channels. The resolution of the videos is 352×240. Dissolve takes up 30.5% of the all transitions and 77.8% of the gradual transitions.

In order to show the normalized variance function is really better than other functions for image quality assessment, a qualitative experiment is designed. We presents the normalized image quality curves in Fig.6, which are obtained by using other three autofocus functions such as Brenner gradient, Tenengrad and energy

Laplace. From the figure, we could see that the Tenengrad curve does not reflect the groove pattern of the first dissolve and the Brenner gradient curve and energy Laplace curve both have a wrong groove pattern at the beginning. They may incur miss detection or wrong detection of dissolve. Furthermore, these curves have several glitches that will bring trouble to dissolve detection. In contrast, the normalized variance function is really a good choice for the image quality evaluation.

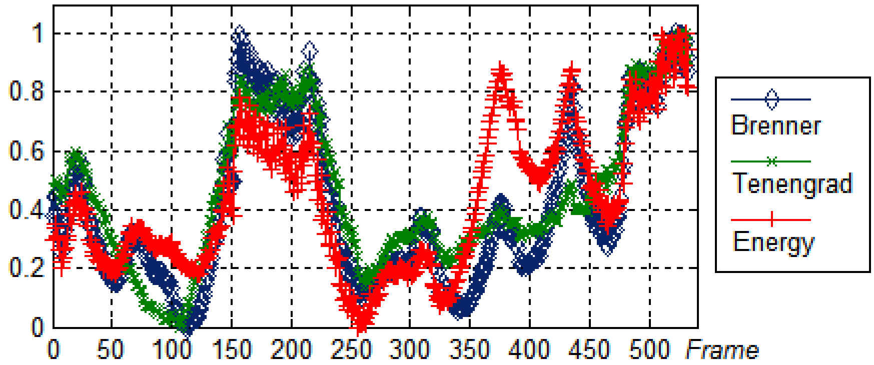

Fig.6. The normalized image quality curves of Brenner gradient, Tenengrad and energy Laplace functions

In the experiment, in order to not miss any dissolve transition, δ_1 and δ_2 in equation (4) are both set as 0.005 according to experience. The distance that controls the merging of two labeled parts in Fig.4 is set as 20 frames. For SVM, the RBF kernel is selected for its good generality. We divide the video set into two sets. 4 videos are for SVM model training and the other 4 for testing. We adopt the precision and recall to show the dissolve detection result.

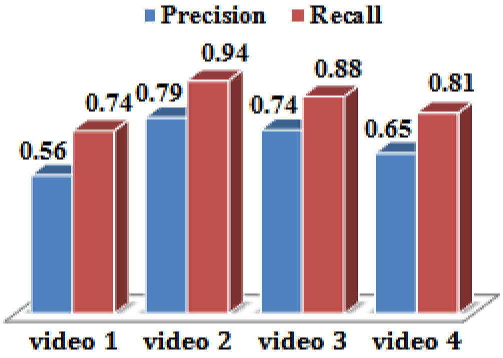

Fig.7. The dissolve detection results of 4 videos

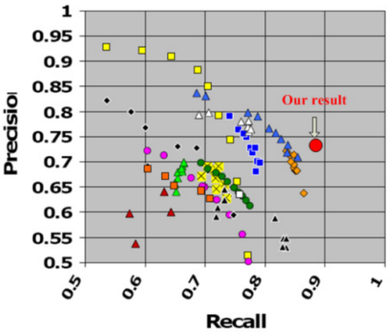

Fig.8. The comparison of our results with the results reported in TRECVID

After the threshold based dissolve candidate detection, nearly all of the right dissolves are detected. The reason is that the thresholds δ_1 and δ_2 are set small enough. In SVM model training, the percentage of the support vector is 30.7%, which means the extracted features are effective for dissolve detection. The final test precision is 73.3% and recall is 88.6%. The detection result for each test video is shown in Fig.7.

The dissolve detection comparison between our results and those reported in TRECVID is shown in Fig.8. Our dissolve detection results are comparable to the results reported in the TRECIVD test. Especially, the recall in our approach is better than other methods. The evaluation and comparison results show that the proposed dissolve detection approach is effective.

5 Conclusions

In this paper, we try to detect dissolves from the viewpoint of image quality analysis and propose a novel coarse-to-fine framework for dissolve detection. In the coarse step, the dissolve candidates are detected with a simple threshold based method. In the fine step, the SVM is used to make the final detection decision. Because we detect dissolve candidates in advance, we can normalize the length of the dissolves easily. The experimental and comparison results show that the proposed method is effective and is comparable to the results reported in the TRECVID. In the future, we will further improve the proposed framework to get better results.

References

1. A. Hanjalic: Shot-Boundary Detection: Unraveled and Resolved?. IEEE Transactions on Circuits and Systems for Video Technology, vol. 12, no. 2. (2002)
2. http://www-nlpir.nist.gov
3. Rainer Lienhart: Reliable Transition Detection in Videos: A Survey and Practitioner's Guide. International Journal of Image and Graphics, 1(3):469-486 (2001)
4. Chih-Wen Su, Hong-Yuan Mark Liao, Hsiao-Rong Tyan, Kuo-Chin Fan, Liang-Hua Chen: A Motion-Tolerant Dissolve Detection Algorithm. IEEE Transactions on Multimedia, 7(6):1106-1113 (2005)
5. Chun-Rong Huang, Huai-Ping Lee, Chu-Song Chen: Shot Change Detection via Local Keypoint Matching. IEEE Transactions on Multimedia, 10(6):1097-1108 (2008)
6. Bogdan Ionescu, Constantin Vertan, Patrick Lambert: Dissolve Detection in Abstract Video Contents. ICASSP 2011: 917- 920 (2011)
7. Jinhui Yuan, Huiyi Wang, Lan Xiao, Wujie Zheng, Jianmin Li, Fuzong Lin, Bo Zhang: A Formal Study of Shot Boundary Detection. IEEE Transactions on Circuits and Systems for Video Technology, 17(2): 168-186 (2007)
8. Zhou Wang, Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: from Error Visibility to Structural Similarity. IEEE Transactions on Image Processing, vol.13, no.4 (2004)
9. Santos A, Ortiz de Solórzano C, Vaquero JJ, Peña JM, Malpica N, del Pozo F: Evaluation of Autofocus Functions in Molecular Cytogenetic Analysis. Journal of Microscopy, 188, 264-72 (1997)

10. Sun Y, Duthaler S, Nelson BJ: Autofocusing in Computer Microscopy: Selecting the Optimal Focus Algorithm. Microscopy Research and Technique, 65(3), 139-149 (2004)
11. V. Vapnik: Statistical Learning Theory, Wiley. (1998)