Neurocomputing 116 (2013) 102–111

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

An adaptive video shot segmentation scheme based on dual-detection model

Xinghao Jiang [a,d], Tanfeng Sun [a,d,*], Jin Liu [b,c,d,**], Juan Chao [a], Wensheng Zhang [c]

[a] School of Information Security Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] State Key Laboratory of Software Engineering, School of Computer, Wuhan University, Wuhan 430072, China
[c] Key Laboratory of Complex System & Intelligence Science, Institute of Automation, Chinese Academy of Science, Beijing 100190, China
[d] Key Laboratory of Shanghai Information Security Management and Technology Research, Shanghai 200240, China

ARTICLE INFO

Available online 9 October 2012

Keywords:
Video shot segmentation
Scale invariant feature transform
Dual-detection model
Adaptive binary search

ABSTRACT

Efficient segmentation of the video shots is the important and foundational work for the research of video content retrieval and analysis. A video shot segmentation scheme based on a dual-detection model is proposed, which includes the pre-detection and re-detection processes. The concepts of uneven blocked color histogram difference and uneven blocked pixel value difference based on human visual features are introduced, which are used as the main descriptors of the pre-detection process to enlarge the importance of central areas and to reduce the noises of background movements and logos. And the adaptive binary search method is introduced to fast detect boundaries with a time complexity of $O(\log_2 n)$. In the re-detection round, the scale invariant feature transform is applied to re-detect boundaries so as to improve the detection precision rate. Experiments show that this algorithm can improve both the recall rate and the precision rate of video shot boundary detection, especially for gradual shot boundaries.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

With the development of digital multimedia technologies and equipment, videos are becoming popular in our life and on the Internet. The content-based video retrieval and analysis is currently a hot research topic [1], and the video shot segmentation is one of the foundational works in the multimedia processing field.

There are mainly two kinds of video shot boundaries [2,3]: abrupt shot boundary and gradual shot boundary. An abrupt shot boundary, also known as cut shot boundary, is the direct concatenation of two adjacent video scenes without transitional frames. A gradual shot boundary is an artificial shot transformation effect which can last for several or even tens of frames. Common used gradual shot boundaries include dissolve, fade in and fade out, and wipe, etc. As we know, to correctly detect gradual shot boundaries is usually more difficult than to detect abrupt shot boundaries. Lienhart has presented an in-depth analysis on why detecting gradual shot boundaries is more difficult than detecting abrupt shot boundaries in [4].

Many researchers are following this filed and various methods have been proposed and compared [5–9], which can be classified into four major categories: edge based method [10,11], pixel based method [12], histogram based method [13,14] and motion vector based method [15,16]. The common edge based method uses the edge change ratio to detect shot boundaries, it is efficient in detecting abrupt, dissolve and fade-in/out shot boundaries, but it is sensitive to strong motions caused by the movements of large objects/camera, meanwhile extracting the edge information is more difficult than pixel values and histograms. The pixel based method commonly calculates the pixel difference of adjacent frames and compares it with one or two thresholds to decide shot boundaries, it is the easiest way of the four methods, but it does not perform very well in detecting gradual shot boundaries, and it is sensitive to large objects/camera movements. The histogram based method is easy and the most common method in shot boundary detection. It can detect abrupt shot boundaries and other short-lasting gradual transitions, but it is not efficient in detecting long-last gradual shot boundaries. The motion based method usually uses the motion vectors extracted from blocking matching to detect shot boundaries, it can be performed directly in the compressed domain, but it is more complicated than the pixel based method and histogram based method, meanwhile it is not efficient in distinguishing gradual shot boundaries and camera motion.

* Corresponding author at: School of Information Security Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. Tel.: +86 13564454752; fax: +86 21 34206657.
** Corresponding author at: State Key Laboratory of Software Engineering, School of Computer, Wuhan University, Wuhan 430072, China. Tel.: +86 13520960916; fax: +86 11 68776837.
E-mail addresses: xhjiang@sjtu.edu.cn (X. Jiang), tfsun@sjtu.edu.cn, jxh5000@gmail.com (T. Sun), mailjinliu@yahoo.com (J. Liu), chaojuan0103@sjtu.edu.cn (J. Chao), wensheng.zhang@ia.ac.cn (W. Zhang).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2011.11.037

The major challenges to video shot segmentation algorithms [17] include the effects of gradual transition, abrupt illumination change and large objects/camera movements. Gradual transitions are slow and the differences of adjacent frames are not obvious as abrupt shot boundaries, so they are usually hard to be detected, and the detection recall rate of gradual shot boundaries is always lower than that of abrupt shot boundaries. For abrupt illumination, the sudden changes caused by flashlight of the camera would change the pixel values and histogram of a frame suddenly, so it can induce false abrupt detections in pixel based methods and histogram based methods. Movements of large objects/camera would cause gradual changes in adjacent frames and they are similar to gradual transitions, so they are usually falsely detected as gradual shot boundaries.

In the shot boundary detection field, the recall rate and precision rate are two main evaluations to shot boundary detection algorithm. In addition, an ideal algorithm is of small computing complexity and real-time. Nowadays though many shot boundary detection algorithms have been proposed, there rarely exists an algorithm compatible for the detection of all the different transition forms. For example, the methods in [18,19] focus on detecting abrupt shot boundaries, Lienhart [4] introduced an algorithm to detect dissolve transitions, also Fernando [20] proposed an algorithm to detect fade and dissolve transitions. Qian [21] proposed an algorithm for fades and flashlight detection, and some researches focus on a special kind of videos, such as news [22,23], sports [24,25] and movie [26,27].

Scale invariant feature transform (SIFT) was proposed by Lowe and has been demonstrated as a robust descriptor by Mikolajczyk and Schmid [28]. It has been widely used in object recognition, video joint and motion tracking [29,30], and nowadays it has been applied to shot segmentation [31]. In our method, SIFT is also applied to re-detect video shot boundaries so as to improve the detection precision rate.

In this paper, a video shot segmentation scheme based on dual-detection model is proposed, which includes the pre-detection process and the re-detection process. Our method can perform well in the detection of both abrupt and gradual shot boundaries. The paper is organized as follows: In Section 2, the dual-detection model and the whole framework of the algorithm are introduced. The detailed procedure is given in Section 3. Section 4 describes relative simulation and comparison experiments, and presents some comparative experimental results. We conclude with a brief discussion and some future work in Section 5.

2. The whole framework with dual-detection model

2.1. The dual-detection model

The dual-detection model is the core of our method, as shown in Fig. 1. In the pre-detection round, the uneven blocked mechanism based on human visual features is proposed. Based on this mechanism the uneven blocked color histogram difference (UBCHD) and uneven blocked pixel value difference (UBPVD) are presented and applied to the adaptive binary search (ABS) procedure which is implemented in moving windows (MW) to detect shot boundaries. The ABS uses different judgment mechanisms for abrupt and gradual shot boundaries. The pre-detection process can list all possible abrupt and gradual shot boundaries for further detection.

The re-detection round applies the SIFT feature matching algorithm to these probable boundaries given by the pre-detection round, it can efficiently exclude those false detection results and improve the detection precision rate. Meanwhile, the re-detection round can adaptively modify relative thresholds used in the whole algorithm to achieve better detection recall rate and precision rate.

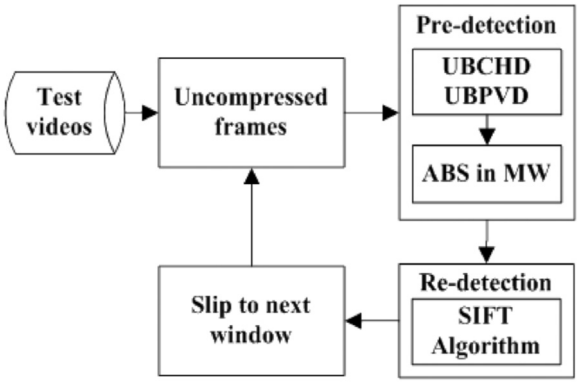


Fig. 1. The dual-detection model.

2.2. The whole framework of our algorithm

The whole procedure of this algorithm is listed as follows:

(1) According to the moving window, video frames are extracted and uncompressed from the video, and then are divided into two sub-windows. The size of the moving window is set to 17 frames; the reason is discussed in Section 3.3.
(2) Calculate the UBCHD and UBPVD of each sub-window according to Sections 3.1 and 3.2; compare them according to the ABS method in Section 3.3.
(3) If the result of step (2) suggests an abrupt shot boundary, re-detect the adjacent frames around the abrupt shot boundary as described in Section 3.4, and move to step (4). If it suggests a gradual shot boundary, re-detect the first frame and last frame of the window, move to step (4). If it suggests no shot boundary in this window, and then directly move to step (6).
(4) If the matching ratio of SIFT feature which can be calculated according to Formula 14 in Section 3.4 is lower than the threshold, we can confirm the pre-detection result. Else it denies the result of pre-detection, directly move to step (6).
(5) For each detected gradual shot boundary, if there is a gradual shot boundary in its former window, then it suggests that a whole gradual shot boundary has been divided into two windows, so just merge them together to one gradual shot boundary.
(6) Slide to next window: for an abrupt shot boundary, the next window starts from the next frame of the boundary. For a gradual shot boundary, it starts from the next frame of the gradual window. For no shot boundary, it starts from the 17th frame of current window.
(7) Repeat from step (1) until the video ends.

3. Description of our algorithm

3.1. Uneven blocked color histogram difference (UBCHD)

The UBCHD is proposed based on human visual features. As we know, a person always fixes his attention in the central area of a picture at first sight, and he pays more attention to foreground and moving objects instead of the static background [32].

As shown in Fig. 2, a frame is divided into 9 uneven blocks with a proportion of 1:3:1 in both its width and height, where (1, 3, 7 and 9) are in Group 1, and (2, 4, 6 and 8) are in Group 2, and the biggest block in the center indexed by 5 belongs to Group 3. For a picture, Group 3 is the focus of the human visual system, and it is also the central area of the video camera so the main objects should be included in this area. To the corners, the four

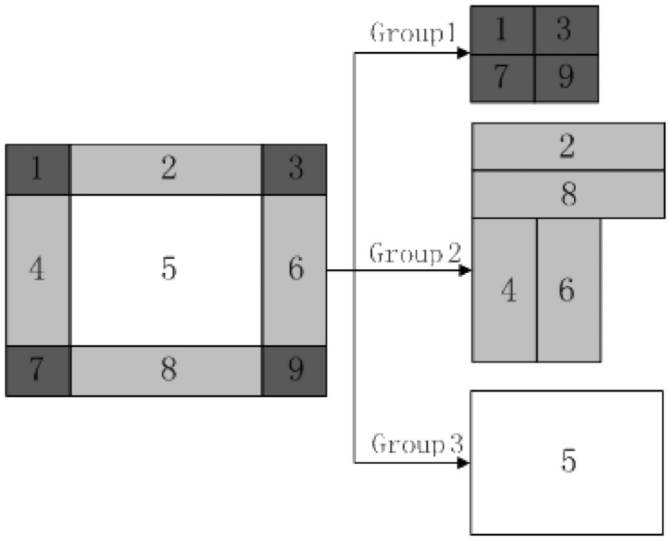


Fig. 2. Uneven blocked mechanism.

blocks in Group 1 are filled with logos in most cases and can be paid less attention to. Meanwhile the four blocks in Group 2 are usually background environment which are of less importance. According to the uneven blocked mechanism, the importance of each group can be easily judged by their locations, so different weighting coefficients can be set to each group. This can efficiently reduce the effect of the logos, subtitles, movements of the camera and large objects.

Calculate the histogram difference of each block group as follows.

$$DH_{(k,c,(i,j))} = \sum_{n=1}^{n=N} \left| \frac{b_{(k,c,i,n)} - b_{(k,c,j,n)}}{S_{(k,c)}} \right|^2 \tag{1}$$

where $DH_{(k,c,(i,j))}$ is the histogram difference; the subscript k is the group index (k=1,2 and 3), c is the color component (for YUV color space, *Y: c=1, U: c=2* and *V: c=3*); i and j are frame indexes; N is the total level of the histogram (usually N=256); b is the bin value in histogram; $S_{(k,c)}$ is the area of Component c in Group k. The areas of component Y, U and V are not equal and should be computed individually. For the common YUV color space, each pixel contains a color component Y, and for each 2×2 pixels, there is only one color component U and one color component V, so the $Y{:}U{:}V$=4:1:1.

Then we can calculate the total difference of the three block groups by adding the three group differences together with different weighting coefficients:

$$DH_{(i,j)} = \sum_{k=1}^{3} \left(\alpha_k \left(\sum_{c=1}^{3} DH_{(k,c,(i,j))} \right) \right) \tag{2}$$

where $DH_{(i,j)}$ is the UBCHD of the ith frame and the jth frame; α_k is the weighting coefficient of Group k, where α_1=0.15, α_2=0.25 and α_3=0.6.

The UBCHD can filter noise and other interference in corners and boundary areas efficiently. In Section 4.3 the comparative experiments are given to show its advantage.

3.2. Uneven blocked pixel value difference (UBPVD)

In our method, the UBPVD is also proposed based on human visual features. As described above in Fig. 2, a frame is divided into 9 uneven blocks, and then the pixel gray level difference of each block is calculated as follows:

$$DP_{(k,c,(i,j))} = \sum_{b_k=1}^{N_k} \sum_{m=1}^{W_{(b_k,c)}} \sum_{n=1}^{H_{(b_k,c)}} \left| \frac{p_{(b_k,c,i,m,n)} - p_{(b_k,c,j,m,n)}}{S_{(k,c)}} \right|^2 \tag{3}$$

where $DP_{(k,c,(i,j))}$ is the pixel difference of the ith frame and jth frame, Group k, color Component c; W is number of pixels in each row; H is the number of pixels in each column; N_k is the number of total blocks in Group k (N_1=4, N_2=4 and N_3=1); b_k is the block index in Group k; $S_{(k,c)}$ is the area of Group k with Component c; p is the gray level of a pixel; (m, n) is the pixel coordinate.

$$DP_{(i,j)} = \sum_{k=1}^{3} \left(\alpha_k \left(\sum_{c=1}^{3} DP_{(k,c,(i,j))} \right) \right) \tag{4}$$

$DP_{(i,j)}$ is the UBPVD between the ith frame and the jth frame, α_k is the same coefficient as Formula (2).

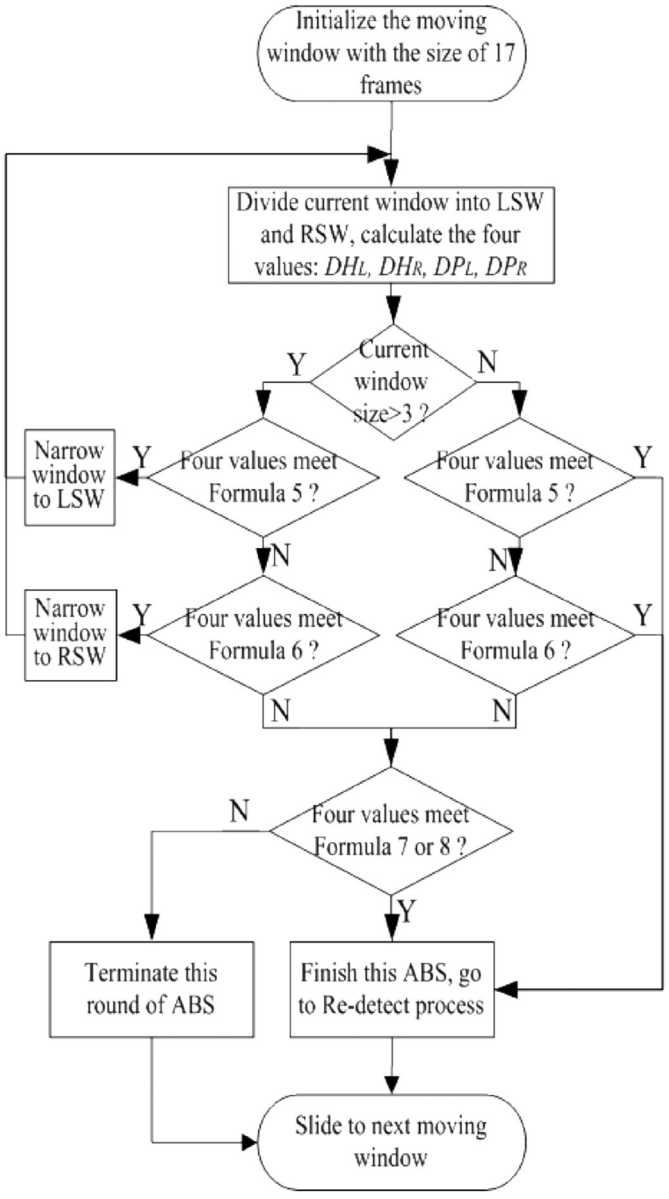


Fig. 3. Procedure of the ABS algorithm.

In Section 4.3 comparisons are designed to demonstrate the advantage of the uneven blocked and weighing coefficients mechanism.

3.3. Adaptive binary search (ABS)

In our method, the adaptive binary search in moving windows is proposed. The initial moving window size is set to 17 frames. There are two reasons. One is that, a video shot commonly lasts for at least 0.5 s to express its contents fully, the frame rate for NTSC videos is 29.97 f/s, thus a video shot lasts for at least 15 frames. The other reason is that, in general binary search a window is divided into two sub-windows $[1, N_w/2]$ and $[N_w/2+1, N_w]$, where N_w is the size of a window. However, in shot boundary detection, there may be an abrupt shot boundary between the Frame $N_w/2$ and Frame $(N_w/2+1)$, so the middle frame should be added to both sub-windows, that is to say $\log_2(N_w\text{-}1)$ should be an integer, so 17 is chosen as the window size. For an initial moving window with 17 frames, the window size would be changed as follows {17, 9, 5 and 3} during one binary search procedure. A number of statistical data have also verified the rationality of the window size of 17.

The adaptive binary search mechanism is as shown in Fig. 3; more details are discussed as follows:

(1) Divide the moving window into two equal parts as left sub window (LSW) and right sub window (RSW), where the middle frame belongs to both sub-windows. So the initial size of each sub-window is 9. Calculate the UBCHDs and UBPVDs between the first frame and the last frame in LSW and RSW, namely DH_L, DH_R, DP_L, DP_R. If the current window (including both the LSW and RSW) size is larger than 3, move to step (2), or else to step (3).

(2) As we know, gradual shot boundaries are made artificially in the video post-process, so there are much more abrupt shots in videos than gradual shots. In our method abrupt shots are detected with a priority to gradual shot boundaries. If the four values of step (1) meet Formula (5), it suggests a possible shot boundary in LSW. If they meet Formula (6), then a shot boundary exists in RSW. In such cases, narrow the window to the probable sub-window and go to step (1). Else if the four values meet Formulas (7) or (8), a gradual shot boundary may exist, finish the ABS process and go to the re-detection round. If the four values meet none of the formulas, it suggests no boundary in this window, terminate the ABS process and slide to the next window.

$$(DH_L > \beta_h DH_R) \quad or \quad (DP_L > \beta_p DP_R) \tag{5}$$

$$(DH_R > \beta_h DH_L) \quad or \quad (DP_R > \beta_p DP_L) \tag{6}$$

$$(DH_L > TH_g) \quad and \quad (DH_R > TH_g) \tag{7}$$

$$(DP_L > TP_g) \quad and \quad (DP_R > TP_g) \tag{8}$$

β_h, β_p are adaptive thresholds modified with Formulas (9) and (10), where D will be replaced by DH and DP accordingly, N_a is the number of detected abrupt shot boundaries, $\beta_{(Na)}$ is the value of β when the number of detected abrupt shot boundaries is N_a. The initial values: $\beta_h=10$, $\beta_p=10$, $N_a=0$.

$$\beta_{temp} = \begin{cases} D_L/D_R, & when(D_L \geq D_R) \\ D_R/D_L, & when(D_L < D_R) \end{cases} \tag{9}$$

$$\beta_{(N_a)} = \begin{cases} \beta_{temp}, & when(N_a = 1) \\ (\beta_{(N_a-1)}(N_a-1)+\beta_{temp})/N_a, & when(N_a \geq 2) \end{cases} \tag{10}$$

where TH_g and TP_g are the adaptive gradual thresholds, which are modified with Formulas (11) and (12), where T will be replaced by TH and TP; N_g is the number of detected gradual shots, $T_{(Ng)}$ is the value of T when the number of detected gradual shots is N_g. The initial values: $TH_g=0.02$, $TP_g=100$, $N_g=0$.

$$T_{temp} = \frac{(D_L+D_R)}{2} \tag{11}$$

$$T_{(N_g)} = \begin{cases} T_{temp}, & when(N_g = 1) \\ (T_{(N_g-1)}(N_g-1)+T_{temp})/N_g, & when(N_g \geq 2) \end{cases} \tag{12}$$

If the four values of step (1) meet Formula (5), it suggests that an abrupt shot boundary may exist in LSW. If they meet Formula (6), then an abrupt shot boundary may exist in RSW. In such cases, finish the ABS process and go to the re-detection round. Else if the four values meet Formulas (7) or (8), a gradual shot boundary may exist, finish the search process and directly go to the re-detection round. Otherwise it suggests no boundary in this window, terminate the ABS process and slide to the next new one.

Fig. 4 gives an example of the ABS while an abrupt shot boundary is detected. As shown in Fig. 4, at the beginning the moving window (Frame: 1–17) is divided into LSW (Frames: 1–9) and RSW (Frames: 9–17), where the 9th frame belongs to both LSW and RSW. Then the UBCHD and UBPVD of each sub window can be calculated. Assume DH_L, DH_R, DP_L and DP_R meet Formula (6), so we can narrow the current window to the RSW (Frames: 9–17) and go on searching it with the ABS procedure. Divide the RSW into Sub-LSW (Frames: 9–13) and Sub-RSW (Frames: 13–17), and repeat the searching processes until the window size is 3 (assume Frames: 11–13). If the four values in window 11–13 meet Formula (5), then there is an abrupt shot boundary in the LSW (Frames 11 and 12), finish the ABS in this window (Frames: 1–17) and go on to the re-detection round by matching the 11th frame and 12th frame with SIFT algorithm. If the re-detection round denies the abrupt shot boundary, the next window begins with the 17th frame, else if it confirms the abrupt shot boundary, adjust the relative abrupt threshold with Formulas (9) and (10) adaptively, and slide to next window which begins with the 13th frame. Here the reason to start the next window from the 13th frame instead of the 17th frame is that: there may be another boundary between the detected shot boundary and the 17th frame. Though according to the assumption before, there is at most one shot boundary in 15 frames, if the detected boundary is between the first frame and the second frame of the window, there are still 15 frames between the detection

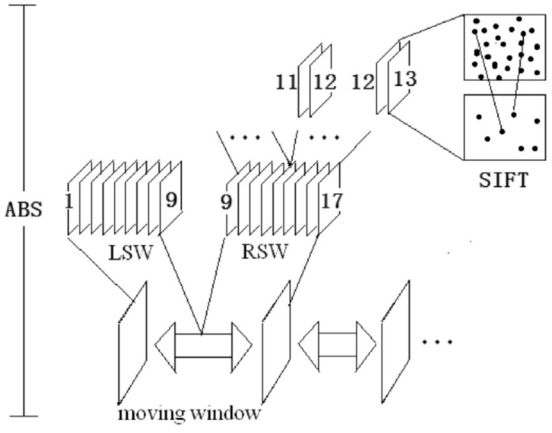


Fig. 4. Detecting an abrupt shot boundary with the ABS method.

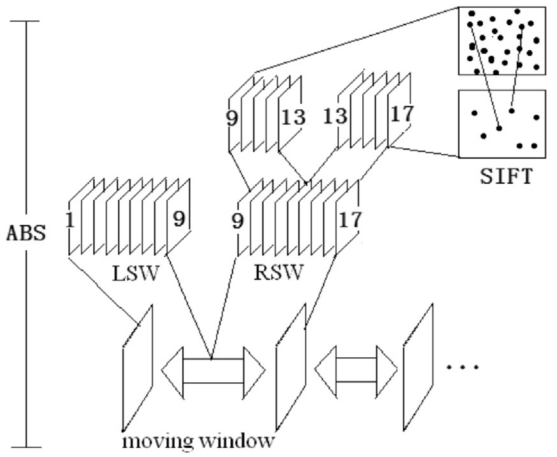


Fig. 5. Detecting a gradual shot boundary with the ABS method.

boundary (the second frame) and the 17th frame. Therefore the next shot boundary shot should be started from the next frame of the detected shot boundary instead of the 17th frame to improve the recall rate of the algorithm, which will be given in Section 4.1.

Fig. 5 gives an example of the ABS while a gradual shot boundary is detected. As shown in Fig. 5, at the beginning the moving window (Frame: 1–17) is divided into LSW (Frames: 1–9) and RSW (Frames: 9–17). Then the UBCHD and UBPVD of each sub window can be calculated. Assume that DH_L, DH_R, DP_L and DP_R meet Formula (6), so we can search the RSW (Frames: 9–17) with the ABS procedure. If DH_L, DH_R, DP_L and DP_R in Sub-LSW (Frame: 9–13) and Sub-RSW (Frame: 13–17) meet Formulas (7) or (8), then it suggests that there may be a gradual shot boundary in the RSW (Frame: 9–17), jump out of the pre-detection round and match the 9th frame and 17th frame with the SIFT algorithm. If the re-detection round confirm the gradual shot boundary, adjust the relative gradual threshold with Formulas (11) and (12) adaptively, and slide to next window which begins with the 18th frame, otherwise the next window begins with the 17th frame.

With ABS method the abrupt and gradual shot boundaries can be detected separately and the computing complexity is reduced from $O(n)$ to $O(\log_2 n)$. A comparative experiment is conducted to illustrate the efficiency of ABS mechanism and is presented in Section 4.4.

3.4. The re-detection process based on SIFT feature matching

SIFT is a robust descriptor in video process. It is used to match two pictures so that to detect or follow the tracks of a given object. Generally speaking, if two pictures are well the same or similar, then the SIFT key-points extracted from them will be matched well. Contrary, if the SIFT key-points of two frames are much different, it suggests that the two frames vary a lot, so there is a shot boundary between them. As a result, the matching rate is applied to measure the similarity of two frames in a video sequence. In our method, SIFT feature matching is used in the re-detection round to exclude possible false detection of the pre-detection round. The procedure is described as follows:

(1) Extract the SIFT feature key-points of the given two frames, and name their numbers as N_1 and N_2. Formula (13) gives the possible relationship between N_1 and N_2, which is discrete value RS.

$$RS=\begin{cases}1, & when(N_1=0 \quad and \quad N_2=0)\\ 2, & when(N_1=0 \quad and \quad N_2\neq 0)\\ 3, & when(N_1\neq 0 \quad and \quad N_2=0)\\ 4, & when(N_1>T_mN_2 or N_2>T_mN_1)\\ 5, & else\end{cases} \tag{13}$$

If $RS=1$ as is shown in Formula (13), it means that both frames are monochrome frames, commonly pure black or white frames, and there is no shot boundary between them. If $RS=2$, there is a fade-in boundary. If $RS=3$, it is a fade-out boundary. If $RS=4$, it means that the key-points of the two frames are much different and there is a shot boundary between them, where $T_m=5$ in our algorithm and it is an empirical value after a number of video tests and analysis. Otherwise N_1 and N_2 are similar and it is not sure whether the pre-detection result is reliable, and then perform further detection with SIFT feature matching, move to step (2).

(2) Match the key-points of these two frames; calculate the R_m with following formula:

$$R_m=N_m/(N_1+N_2-N_m)\times 100\% \tag{14}$$

where R_m is the matching rate of two video frames, N_m is the number of correctly matched key-points with SIFT feature matching. If R_m is lower than the static matching threshold (in our algorithm, the threshold is 2%), then there is really a shot boundary between the two frames, else deny the detection results of the pre-detection round.

In Section 4.4, another comparative experiment between our algorithm and the single-round detection method without SIFT feature matching is performed to demonstrate that SIFT feature matching can improve the precision rate.

4. Experiments and results

4.1. Experiment environment and evaluation standards

All the experiments in the paper are based on the Visual Studio 2005 platform and the FFMPEG and OpenCV function Library with C and C++. The FFMPEG-3.0 and OpenCV-2.0 function library are both open source, they include a series of video processing functions. Additional tools include Elecard StreamEye to view independent video frames, and BRVideoConverter to convert different videos into MPEG-4 format with the frame ratio of 30 f/s. Before starting the test, we treated a number of tests to adjust relative parameters (the uneven blocked proportion 1:3:1, the different weighting coefficients $\alpha_1=0.15$, $\alpha_2=0.25$ and $\alpha_3=0.6$, the initial value of $\beta_h=10$, $\beta_p=10$, $TH_g=0.02$, $TP_g=100$, and the static matching threshold 2%, $T_m=5$ in Formula (13)) in our algorithm to the best.

To evaluate the detection efficiency of our algorithm, the recall rate (R_r) and precision rate (R_p) [6] are used. The recall rate is the percentage of correctly detected boundaries among all existing boundaries, and the precision rate is the percentage of correctly detected boundaries among all the detected ones.

$$R_r=N_c/(N_c+N_l)\times 100\% \tag{15}$$

$$R_p=N_c/(N_c+N_f)\times 100\% \tag{16}$$

where N_c is the number of correctly detected shot boundaries, a correctly detected shot boundary is a really existing shot boundary that has been detected; N_l is the number of lost shot boundaries, a lost boundary is a really existing shot boundary

Table 1
Simulation experiment results with our algorithm.

Video type	News		Movie		Cartoon	
Total frames	109,534		142,745		141,940	
Boundary type	A	G	A	G	A	G
Number of shots	442	65	1594	77	1595	105
N_c	426	61	1535	71	1429	88
N_f	10	28	3	13	24	35
N_l	16	4	59	6	166	17
R_r (%)	96.38	93.85	96.30	92.21	89.59	83.81
R_p (%)	97.71	68.54	99.80	84.52	98.35	71.54

but has not been detected; N_f is the number of falsely detected shot boundaries, i.e., there is no shot boundary among these frames but the algorithm result returns one.

4.2. Simulation

The TRECVID [33] video database is chosen as a main part of the test video dataset, but we have to add some other videos from Internet since there is no enough cartoon videos and sample videos with gradual shot boundaries in the TRECVID database. There are 394,219 frames, 3878 shot boundaries in our test video database. Table 1 lists the basic information of the test video data and the experimental results of our algorithm, the boundary type *A* in the third row denotes abrupt shot boundaries and *G* is short for gradual shot boundaries.

As is shown in Table 1, the simulation experiment has chosen a large number of test videos of news, movie and cartoon. From the number of shots, it is clearly that there are more abrupt shot boundaries than gradual shot boundaries, which corresponds to the analysis in Section 3.3. In average there is an abrupt shot boundary for each 109 frames and a gradual shot boundary for each 1596 frames. Meanwhile for different video types the last time of each shot also varies, for news videos a shot lasts 216 frames, for movies videos it lasts 85 frames, and for cartoon videos it lasts 83 frames. These data correspond with the objective facts. Generally speaking, in news videos a shot may last long to explain an issue clearly, while in movie and cartoon, which are kind of entertainments, a shot will last shorter to attractive audiences.

From Table 1, it can be found that the detection results of abrupt shot boundaries are always better than those of gradual shot boundaries. For example for news videos, the abrupt precision rate is 29.17% higher than the gradual precision rate, and for cartoon videos the abrupt recall rate is 6.78% higher than the gradual recall rate. In addition, for abrupt shots, the recall rate is lower than the precision rate, and the greatest difference is only 8.76% for cartoon. While for gradual shots the recall rate is higher than the precision rate, with the greatest difference of 25.31% for news, 12.27% for cartoon and the shortest distance can also reach 7.69% for movie.

In addition, the detection results of different video types also vary a lot. The gradual recall rate of news videos is 10.04% higher than that of the cartoon videos, while the gradual precision rate of movie videos is 15.98% higher than that of the news videos. The abrupt recall rate of news videos is 6.79% higher than that of the cartoon videos; while the abrupt precision rate of movie videos is 2.09% higher than that of the news videos. In a word, different video types have their own advantages in detecting different boundaries.

4.3. Comparisons with even blocked mechanism under different weighting policies

In this section, two comparative experiments directed at the uneven blocked mechanism are performed to demonstrate the efficiency of uneven blocked mechanism and different weighting policies.

4.3.1. 3×3 even blocked mechanism with equal weighting coefficients

Compared with the uneven blocked mechanism, the experiment divides each frame into 3×3 equal-size blocks, and three even blocked differences are added with the same weighting coefficient in the pre-detection. The even blocked mechanism is shown as follows.

In Fig. 6, a video frame is divided into 9 equal blocks, where block (1, 3, 7, 9) belong to the Group 1; block (2, 3, 6, 8) belong to the Group 2; and block 5 belongs to the Group 1. The color histogram difference and pixel value difference are calculated with the same Formulas (1–4), except that the weighting coefficients of each group is the same, where $\alpha_1=1/3$, $\alpha_2=1/3$ and $\alpha_3=1/3$.

4.3.2. 3×3 even blocked mechanism with different weighting coefficients

The second comparative experiment is the same with the first comparative experiment except that the three groups are added together with different weighting coefficients, where $\alpha_1=0.15$, $\alpha_2=0.25$ and $\alpha_3=0.6$.

Table 2 lists the test results of our algorithm and the two other methods with the 3×3 even blocked mechanism using different weighting policies. For abrupt shot boundaries, the recall rates and precision rates are more or less the same among the three methods. However for gradual shot boundaries, the recall rates are also close among the three methods, but the precision rates among the three methods are much different. The precision rate of our algorithm is as much as 34.72% higher than that of the method with 3×3 even blocks and equal weighting coefficients.

Contrast our algorithm with the method of 3×3 even blocked mechanism using different weighting coefficients, where the two methods are the same except the blocked mechanism. All these six evaluation values of our algorithm are higher than those of the comparative one, with a biggest difference of 21.02% in the gradual precision rate. It testifies that by dividing a frame into uneven sub-blocks, it can keep the integrity of the main object which usually appears in the central area of the video frames, it corresponds to the center-focusing feature of human's visual system and can reduce false detections caused by background changes and camera movements.

Comparing the effects of different weighting policies, as the results shown in Table 2, the detection results of abrupt shot

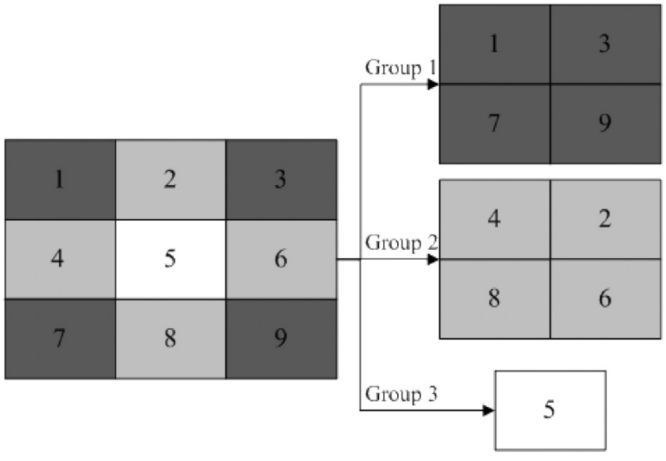


Fig. 6. 3×3 even blocked mechanism.

Table 2
Comparison results of our algorithm and the 3×3 even blocked mechanism with different weighting policies.

Algorithm	Abrupt		Gradual		Total	
	R_r (%)	R_p (%)	R_r (%)	R_p (%)	R_r (%)	R_p (%)
Our algorithm	93.36	98.92	89.07	74.32	93.09	96.96
3×3 even blocked, with equal weighting coefficients	93.86	99.13	87.85	39.60	93.48	90.94
3×3 even blocked, with different weighting coefficients	92.10	97.95	88.26	53.30	91.85	93.17

Table 3
Comparative results between ABS and frame-to-frame method.

Algorithm	Abrupt		Gradual		Total	
	R_r (%)	R_p (%)	R_r (%)	R_p (%)	R_r (%)	R_p (%)
Our algorithm with ABS method	93.36	98.92	89.07	74.32	93.09	96.96
Our algorithm with frame-to-frame method	91.60	97.17	69.23	90.48	90.17	96.82

Table 4
Comparative results with single-round detection without SIFT.

Algorithm	Abrupt		Gradual		Total	
	R_r (%)	R_p (%)	R_r (%)	R_p (%)	R_r (%)	R_p (%)
Our algorithm with dual-detection	93.36	98.92	89.07	74.32	93.09	96.96
Single-round detection without SIFT	82.92	72.14	86.64	71.33	83.16	72.08

boundaries by the equal weighting coefficients are better than those of the different weighting coefficients, but the detection results of the gradual shot boundaries are the opposite. The policy of different weighting coefficients sets a larger weighting coefficient for the central area of the frame and enlarges the importance of the central area, thus then efficiently exclude the noise of background caused by movements of the camera or large objects.

4.4. Comparison between ABS method and frame-to-frame method

In this comparative experiment, the ABS method in our algorithm is replaced by the frame-to-frame method, which calculates the difference between adjacent frames and compare it with static thresholds to estimates shot boundaries. If the difference between two frames is larger than a given static abrupt threshold, then it suggests an abrupt shot boundary. Else if the differences of several adjacent frames are larger than a static gradual threshold and lower than the static abrupt threshold, then it means a gradual shot boundary.

In Table 3, five values of our algorithm with ABS method are higher than those of the frame-to-frame method. The recall rate of gradual shot boundaries by the ABS method is 19.84% higher than the frame-to-frame method, which demonstrates that by calculating the difference of the first frame and last frame in a window, the ABS method can efficiently enlarge the changes of gradual shot boundaries, thus can reduce missing detections and improve the recall rate of gradual shot boundaries.

However, the ABS method calculates the difference of the first frame and last frame in a window instead of the two adjacent frames. It will expand the difference which is caused by movements of the camera or large objects and induce somewhat false detections, so the precision rate of gradual shot boundaries with ABS method is lower than that of the frame-to-frame method.

In conclusion, the ABS method in our algorithm can well improve the detection recall rate, but it is not perfect in distinguishing large objects/camera movements and gradual shot boundaries, which can induce false gradual detections. In addition, the ABS method can efficiently improve the time complexity of our algorithm to $O(\log_2 n)$, compared to the complexity of the frame-to-frame method which is $O(n)$.

4.5. Comparison between dual-detection and single-round detection

In this section another comparison between our dual-detection mechanism and single round detection without SIFT is performed. For the single-round detection, the re-detection process based on SIFT feature matching is canceled, which means that if the pre-detection round suggests a shot boundary, just trust it without any checking mechanism. The comparative experiment is performed to demonstrate that the SIFT algorithm can exclude the false detections of the pre-detection and thus can improve the precision rate.

From the results in Table 4, for the single-round detection without SIFT, both the recall rate and precision rate drop down dramatically, where the abrupt precision rate is 26.78% lower, and the gradual precision rate is 2.99% lower than that of our algorithm. Since two similar frames can be well matched through the SIFT algorithm, the re-detection round can efficiently remove those false detection in the pre-detection round, thus improve the precision rate. In return, with the SIFT algorithm excluding false detections, the pre-detection threshold can be set a little lower so as to enhance the recall rate. As shown in Table 4, the abrupt shot boundaries recall rate of our algorithm is 10.44% higher and the gradual recall rate is 2.43% higher.

4.6. Comparison with Li's algorithm

A comparison with Li's algorithm in [31] is designed, it uses 4×4 blocked color histogram differences and the number of SIFT

key points as the main descriptors to detect shot boundaries with SVM method.

Fig. 7 shows the comparative results of our algorithm and the method presented in [31]. The recall rate of gradual shot boundaries by our algorithm is 89.07%, which is 18.22% higher than that of the method presented in [31], which suggests that the uneven blocked mechanism and the ABS method can efficiently enlarge the difference of the slow changes in gradual transitions, thus they improve the recall rate of gradual shot boundaries. Meanwhile the SIFT feature matching algorithm can efficiently exclude the false detections of the pre-detection, so the precision rates of both abrupt and gradual shot boundaries are obviously improved compared to [31]. For abrupt shot boundaries, the precision rate is 98.92% and it is 4.88% higher than that of the method presented in [31], which is 94.04%. For gradual shot boundaries the precision rate of our algorithm is 74.32%, it is 4.32% higher than that of the method presented in [31].

However, the recall rate of our algorithm is 93.36%, which is 1.88% lower than that of the method presented in [31]. It may be caused by the large initial values of the adaptive threshold β_h, β_p. If the initial values of these two thresholds decrease a little it can improve the recall rate of abrupt shot boundaries. But it can also induce much false detection, which would increase the times of SIFT feature matching and waste a lot of time.

4.7. Comparison with Kucuktunc's algorithm

Another comparison with Kucuktunc's algorithm presented in [14] is conducted. The algorithm produces a histogram with 15 color bins for each video in the Lab color space, and it is based on the fuzzy logic theory. If the difference of two adjacent frames is larger than a static abrupt threshold, it suggests an abrupt shot boundary; else if for many continuous frames, their differences are larger than a static gradual threshold and are lower than the static abrupt threshold, then it suggests a gradual shot boundary.

Fig. 8 shows the comparative results between our algorithm and the method presented in [14], the six values of our algorithm are all higher than those of the method presented in [14]. For abrupt shot boundaries the recall rate of by our algorithm is 4.96% higher and the precision rate is 10.66% higher than that of the method presented in [14]. For gradual shot boundaries the advantages of our algorithm is much more obvious, where the recall rate is 26.72% higher and the precision rate is 25.89% higher. The result shows that our algorithm improves the recall rate of gradual shot boundaries obviously; it is because that the uneven blocked mechanism and different weighting coefficients in

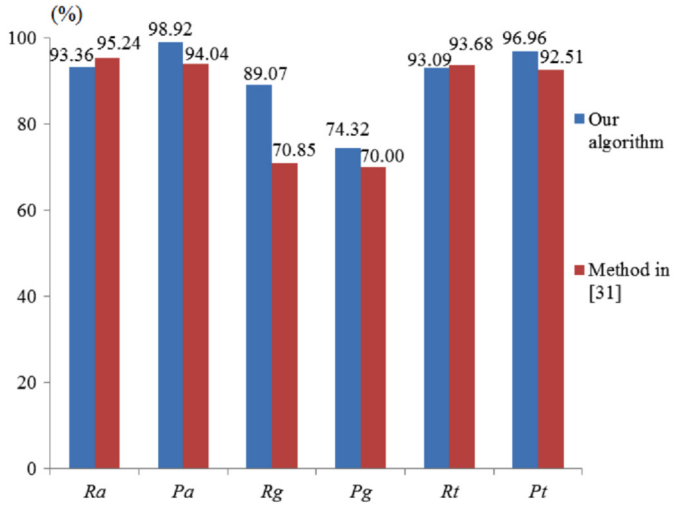


Fig. 7. Comparisons between our algorithm and the method presented in [31].

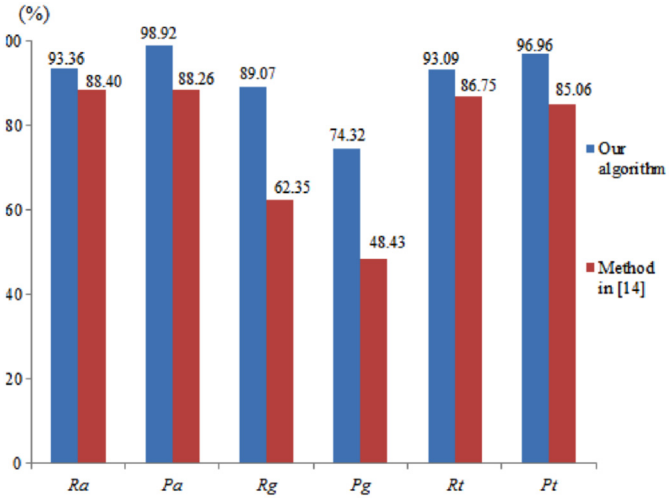


Fig. 8. Comparisons between our algorithm and the method presented in [14].

our method can improve the detection of gradual shot boundaries as well as reduce the false detection caused by the movements of large objects or the camera. In addition, the re-detection method based on SIFT feature matching can well exclude the false detection and improve the precision rate as is shown in Fig. 8 where the total precision rate of our algorithm is 11.90% higher than that of the method presented in [14].

4.8. Summary for experiments

Based on all these experiments, conclusion can be drawn safely according to the experimental results above:

(1) The average recall rate of our algorithm reaches 93.47%, and the precision rate attains 98.93%. In general, it achieves a better performance than all the comparative experiments in the recall rate as well as the precision rate.
(2) The detection results of abrupt shot boundaries are always much better than gradual shot boundaries because abrupt shot boundaries change faster than gradual shot boundaries. The precision rate of abrupt shot boundaries is 98.93%, while it is 74.58% for gradual shot boundaries with a distance of 24.35%.
(3) The uneven blocked mechanism can efficiently reduce the missed detections and can improve the recall rate, especially in gradual shot boundaries. The detection computing efficiency is much improved by the ABS method, and the detection precision rate has been advanced much by the SIFT feature matching algorithm.

5. Conclusion

In this paper a novel dual-detection shot segmentation algorithm to detect the abrupt and gradual shot boundary simultaneously is proposed. In the pre-detection round of this algorithm the uneven blocked mechanism based on human visual system is put forward, and consequently proposed the uneven blocked color histogram difference and uneven blocked pix value difference as the main descriptors for pre-detection in moving windows, by combining both the pixel based method and histogram based method together, the algorithm achieves good performance in detection abrupt and short-lasting boundaries with simple descriptors. In addition, a new boundary searching method called Adaptive Binary Search is proposed. It can improve the detection recall rate greatly especially for gradual shot boundaries, and also

can reduce the time cost, the time complexity of our algorithm is $O(\log_2 n)$ compared to $O(n)$ *of* the frame-to-frame searching method. In the re-detection round, the SIFT feature matching algorithm is used to exclude false detection of the pre-detection round so as to improve the precision rate.

Experiments have indicated that our algorithm performs well in both abrupt and gradual shots detection in the detection of abrupt and gradual shot boundary for movie, news and cartoon videos. Collectively, the uneven blocked and weighed difference mechanism in UBCHD and UBPVD can reduce the effect by the movements of large objects or the camera in a frame, obviously improve the gradual shot boundary detection precision rate and search a well balance among abrupt and gradual shot boundary detection. The ABS mechanism can efficiently enlarge the difference of shot boundary; as a consequence it improves the recall rate much, especially in the advance of gradual recall rate. And the re-detection round based on SIFT feature matching algorithm can effectually remove the false results produced by the pre-detection and increase the precision rate greatly.

Future works include reducing the residual shot boundary while in the pre-detection round to increase the recall rate, improving the matching efficiency, and reducing the computation complexity.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Nos. 60802057, 61071153, 61070013, 61272439, 61272249), Sponsored by Shanghai Rising-Star Program (10QA1403700), Program for New Century Excellent Talents in University (NCET-10–0569), the 111 Project (B07037) and the Science and Technology Commission of Wuhan Municipality 'Chenguang Jihua' (201050231058).

References

[1] A.F. Smeaton, P. Over, A.R. Doherty, Video shot boundary detection: seven years of TRECVid activity, J. Comput. Vision Image Understand. 114 (2010) 411–418.
[2] I. Koprinska, S. Carrato, Temporal video segmentation: a survey, J. Signal Process. Image Commun. 16 (2001) 477–500.
[3] R. Lienhart, Reliable transition detection in videos—a survey and practitioner's guide, Int. J. Image Graph. 1 (2001) 469–486.
[4] R.W. Lienhart, Reliable dissolve detection, in: Proceedings of SPIE Storage and Retrieval for Media, San Jose, vol. 4315, 2001, pp. 219–230.
[5] R. Lienhart, Comparison of automatic shot boundary detection algorithms, J. Proc. SPIE 3656 (1999) 290–301.
[6] J.S. Boreczky, L.A. Rowe, Comparison of video shot boundary detection techniques, J. Proc. SPIE 2664 (1996) 170–179.
[7] A. Hanjalic, Shot boundary detection: unraveled and resolved? J. IEEE Trans. Circuit Syst. Video 12 (2002) 90–105.
[8] C.R. Huang, H.P. Lee, C.S. Chen, Shot change detection via local keypoint matching, J. IEEE Trans. Multimedia 10 (2008) 1097–1108.
[9] C. Cotsaces, N. Nikolaidis, I. Pitas, Video shot detection and condensed representation, J. IEEE Signal Process. Mag. 23 (2006) 28–37.
[10] H.W. Yoo, H.J. Ryoo, D.S. Jang, Gradual shot boundary detection using localized edge blocks, J. Multimedia Tools Appl. 28 (2006) 283–300.
[11] D. Adjeroh, M.C. Lee, N. Banda, U. Kandaswamy, Adaptive edge-oriented shot boundary detection, EURASAP, J. Image Video Process. (2009) 13, http://dx.doi.org/10.1155/2009/859371.
[12] S.G. Lian, Automatic video temporal segmentation based on multiple features, J. Soft Comput. 15 (2011) 469–482.
[13] U. Gargi, R. Kasturi, S.H. Strayer, Performance characterization of video-shot-change detection methods, J. IEEE Trans. Circuit Syst. Video 10 (2000) 1–13.
[14] O. Kucuktunc, U. Gudukbay, O. Ulusoy, Fuzzy color histogram-based video segmentation, J. Comput. Vision Image Understand. 114 (2010) 125–134.
[15] H.Y. Zhou, A.H. Sadka, M.R. Swasha, J. Azizia, U.A. Sadiqa, Feature extraction and clustering for dynamic video summarisation, Neurocomputing 73 (2010) 1718–1729.
[16] A.M. Amel, B.A. Abdessalem, M. Abdellatif, Video shot boundary detection using motion activity descriptor, J. Telecommun. 2 (2010) 54–59.
[17] J.H. Yuan, H.Y. Wang, L. Xiao, W.J. Zheng, J.M. Li, F.Z. Lin, B. Zhang, A formal study of shot boundary detection, J. IEEE Trans. Circuit Syst. Video 17 (2007) 168–186.
[18] Z. Cernekova, I. Pitas, C. Nikou, Information theory-based shot cut/fade detection and video summarization, J. IEEE Trans. Circuit Syst. Video 16 (2006) 82–91.
[19] T. Barbu, Novel automatic video cut detection technique using Gabor filtering, J. Comput. Electr. Eng. 35 (2009) 712–721.
[20] W.A.C. Fernando, C.N. Canagarajah, D.R. Bull, Fade and dissolve detection in uncompressed and compressed video sequences, in: Proceedings of the International Conference on Image Processing, Kobe, vol. 3, 1999, pp. 299–303.
[21] X.M. Qian, G.Z. Liu, R. Su, Effective fades and flashlight detection based on accumulating histogram difference, J. IEEE Trans. Circuits Syst. Video 16 (2006) 1245–1258.
[22] T.S. Chua, S.F. Chang, L. Chaisorn, W.H. Hsu, Story boundary detection in large broadcast news video archives: techniques, experience and trends, in: Proceedings of the 12th ACM International Conference on Multimedia, New York, 2004, pp. 656–659.
[23] H. Lee, J. Yu, Y. Im, J.M. Gil, D. Park, A unified scheme of shot boundary detection and anchor shot detection in news video story parsing, J. Multimedia Tools Appl. 51 (2011) 1127–1145.
[24] L.Y. Duan, M. Xu, Q. Tian, C.S. Xu, J.S. Jin, A unified framework for semantic shot classification in sports video, J. IEEE Trans. Multimedia 7 (2005) 1066–1083.
[25] S.G. Lian, Y.A. Dong, H.L. Wang, Efficient temporal segmentation for sports programs with special cases, in: Proceedings of the Advances in Multimedia Information—PCM 2010 PT I. LNCS, Shanghai, vol. 6297, 2010, pp. 381–391.
[26] A. Hanjalic, R.L. Lagendijk, J. Biemond, Automated high-level movie segmentation for advanced video-retrieval systems, J. IEEE Trans. Circuits Syst. Video 9 (1999) 580–588.
[27] L.H. Chen, Y.C. Lai, H.Y.M. Liao, Movie scene segmentation using background information, J. Pattern Recognition 41 (2008) 1056–1065.
[28] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, J. IEEE Trans. Pattern Anal. 27 (2005) 1615–1630.
[29] D.G. Lowe, Distinctive image features from scale-invariant key-points, Int. J. Comput. Vision 60 (2004) 91–110.
[30] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Vol. 2(8), 1999, pp. 1150–1157.
[31] J. Li, Y.D. Ding, Y.Y. Shi, W. Li, A divide-and-rule scheme for shot boundary detection based on SIFT, J. Digit. Content Technol. Appl. 4 (2010) 202–214.
[32] W. Osberger, A.J. Maeder, Automatic identification of perceptually important regions in an image, in: Proceedings of the Fourteenth International Conference on Pattern Recognition, Brisbane, Queensland., Australia, vol. 1, 1998, pp. 701–704.
[33] A.F. Smeaton, P. Over, W. Kraaij, Evaluation campaigns and TRECVid, in: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, Santa Barbara, California, 2006, pp.321–330.

Xinghao Jiang received the PhD. Degree in Electronic Science and Technology from Zhejiang University, Hangzhou, PR China in 2003. He is an associate professor at the School of Information Security Engineering at Shanghai Jiao Tong University, Shanghai, PR China. His current research interests include multimedia security and image retrieval, intelligent information processing, cyber information security, information hiding and watermarking.

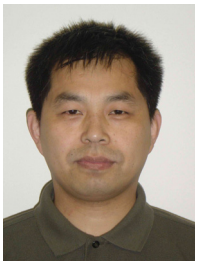
Dr. Jiang is an IEEE member.

Tanfeng sun received the PhD. Degree in Information and Communication Engineering from Jilin University, Jilin, PR China in 2003. He is a lecturer at the School of Information Security Engineering at Shanghai Jiao Tong University, Shanghai, PR China. His current research interests include multimedia security and image retrieval, information hiding and watermarking.

Jin Liu received the Ph.D. Degree from the State Key Lab of Software Engineering at Wuhan University, China in 2005. Since 2007 he is an associate Professor in the State Key Lab of Software Engineering, Wuhan University. His areas of research include software modeling on the Internet and intelligent information processing.

Dr. Liu is an IEEE member.

Juan Chao received the Bachelor Degree in the School of Information Security Engineering from Shanghai Jiao Tong University, Shanghai, PR China in 2010. She is now a Graduate student at the School of Information Security Engineering at Shanghai Jiao Tong University, Shanghai, PR China. Her current research interests include shot segmentation and digital video forgery detection.



Wensheng Zhang received the Ph.D. Degree in Pattern Recognition and Intelligent Systems from the Institute of Automation, Chinese Academy of Sciences (CAS), in 2000. He is a Professor of Machine Learning and Data Mining and the Director of Research and Development Department, Institute of Automation, CAS. His research interests include computer vision, pattern recognition, artificial intelligence and computer human interaction.