A Video Shot Boundary Detection Approach based on CNN Feature

1st Rui Liang
School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu, Sichuan
1662723894@qq.com

2nd Qingxin Zhu
School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu, Sichuan
1429171743@qq.com

3rd Honglei Wei
Sport department & School of Economics and Manageme
Southwest Jiaotong University
Chengdu, Sichuan
529996146@qq.com

4rd Shujiao Liao
School of mathematics and statistics
Minnan Normal University
Zhangzhou, Fujian
sjliao2011@163.com

***Abstract*—In nowadays,as the development of digital photographic technology, video files grow rapidly, there is a great demand for automatic video semantic analysis in many scenes,such as video semantic understanding,content-based analysis,video retrieval. Shot boundary detection is a key basic technology and first step for video analysis. However,recent methods are time consuming and performs bad in the gradual transition detection. In this paper we proposed a new approach which used CNN model to extract features of video sequence parallelly based on GPU, so we can simplify the expression of video and reduce the calculation time for shot detection, and took local frame similarity and dual-threshold sliding window similarity into consideration to increase recall and precise of shot detection. The experimental result shows that the proposed method can achieve a high $F1$ score and excellent detection speed.**

***Index Terms*—Shot boundary detection,Shot transition,CNN feature,Dual-threshold sliding window.**

I. Introduction

With the development of multimedia and network technologies,massive videos are uploading to the internet,repid video understanding and content-based retrieving become serious problems. A video shot is definde as a sequence of frames taken by a single camera. Shots are the basic semantic units of video, the detection of shot's boundary is the basis of video segmentation,management,retrieving. The quality of shot boundary detection will affect the efficiency of video retrieving and video semantic analysis. After shot boundary detection, shots can be extracted from video.The essence of shot boundary detection is to distinguish the switching process between two shots,which is a semantic change process [1],called transition. There are two types of shot transitions:cut and gradual.A cut transition changes sharply between the shots boundary.A gradual transition changes in a mild way such as fading, dissolving,wiping and zooming and other gradual effects.

This paper is financially supported by the National Natural Science Foundation of China (Grant No. 61300192). The Fundamental Research Funds for the Central Universities (Grant No. ZYGX2014J052)

II. Related Work

Early methods of shot boundary detection mainly focused on cut detection because there is great difference between two adjacent frames.When the distance measure of two adjacent frames exceeds a threshold, a cut transition is detected.While gradual transition is much complex and hard to detected. In order to detect graudal transition, a classical way of shot boundary detection is based on dual-threshold method. The main idea is that frame distance on the edge of gradual transition is larger than internal frames,but is much smaller than cut transtion. Pixel based shot segmentation [2] which calculated the change of gray degree between two frames;histogram based shot segmentation method [3] which counts the number of pixels in different gray degree;X^2 histogram [4] was widely used because it can enlarge the difference of frames and the algorithm is stable;blocked X^2 histogram [5] [6] was improved from X^2 histogram which is moving objects tolerance,but is time consuming. Because of complexity of gradual transition, some algorithms were specially proposed to detect gradual transition, [7] detect gradual transition by evaluating the degree to which the transition matches the corresponding model; [8] mutual information and joint entropy are used for cut and fade transtion detection;edge contour change rate shot boundary detection method [9] [10]which used canny operator to get the edge of images and calculated the total dispalcement to determine the boundaries of shots; [16] proposed a unified SBD framework based on graph partition model, which formulated SBD in the pattern recognition perspective, although they achieved good results, still not as good as machine learning opproaches.

Beside accuracy and recall, high computation complexity of detection algorithms hindered the real-time applications. Li *etal* [11]. employed preprocessing techniques to segment

978-1-5386-2937-6/17 $31.00 © 2017 IEEE
DOI 10.1109/ISM.2017.97

video and select candidate segments which are considered as suspect shot change fragments,then they improved the LTD algorithm [12] to detect gradual transitions,their method greatly speeded up shot boundary detection. In [13],researchers applied candidate segment selection to improve the singular value decomposition based method [15] and speeded up shot boundary detection.

III. Proposed Approach

Problem Analysis. Giving a video sequence $X = (x_1, x_2, ..., x_n)$, dividing the sequence into sub-sequences $X_1, X_2, ..., X_m$,call each sub-sequence a shot,every shot contains an independent semantic content or event.The purpose of shot boundary detection is to divide a video into meanful segments. In this paper, we proposed a novel method for shot boundary detection, this method can be divided in four main parts.Firstly,we extracted CNN feature for each frame and used the feature to represent each frame;secondly, calculate consine similarity to describe the similarity of a pair of frames;thridly,a local frames' similarity based method for cut transition detection;lastly,a window similarity and dual-threshold based method for gradual transition.

Feature Extraction. The first step is extracting feature for each frame in a video sequence.In this paper used pre-trained deep CNN(Convolutional Neural Network) models to extract feature of each frame in a video sequence.The reason of using deep CNN model for extracting feature is that these famous deep CNN models express the semantics of images well,and has the advantages of stablility of deformation such as translation,zoom,tilt; these models have achieved very good performance on the tasks of image classisfication,object detection;feature extraction process can be accelerated by GPU. So these features are suitable for the measure the similarity between two frames. In this paper,we tried pre-trained AlexNet [17] models to extract fc7 layer feature and pre-trained ResNet-152 [18] model to extract pool5 layer feature, the reason we chose these models were these two models are accurate enough and have fewer parameters, so they are highly time efficiency.

Similarity metric. Similarity metric calculate the degree of similarity between two images.In contrast to distance metric,the smaller the similarity metric value,the smaller the similarity between images,the greater the difference.Common similarity metrics include:Consine similarity,Pearson correlation coefficient,Jaccard similarity coefficient,Adjusted cosine similarity.In the paper we used cosine similarity:

$$sim(X, Y) = \cos\theta = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{\| x \| \cdot \| y \|} \tag{1}$$

as the similarity metric of a pair of frames.In equation 1, X and Y are the feature vectors of two images,which is the CNN feature in this paper.

A. Shot boundary detection based on CNN feature

In order to detect cut and gradual transitions, we introduced three threshold,T_c is the threshold of cut transition;T_{gh} indicates the start of a gradual transition;T_{gl} indicates the lowest threshold of a gradual transition.

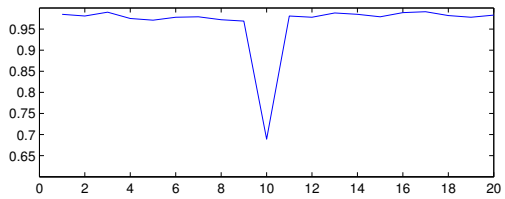


Fig. 1. Similarity between adjacent frames of cut transition

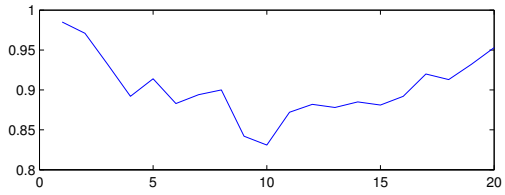


Fig. 2. Similarity between adjacent frames of gradual transition

Cut transition detection. In a cut transition, frame changes charply, there is a great difference between the two adjacent frames, in the perspective of similarity, the cosine value would be small,Based on this fact, we proposed a necessary condition for detect cut transition boundary:

$$S(t, 1) < T_c \tag{2}$$

$S(t, 1)$ stands for the similarity of tth,$t + 1$th frame.There always are some exceptions,such as flashlight may course a great change of feature,so we calculated another two similarity$S(t, 2), S(t + 1, 1)$. When the previous necessary condition is satisfied and the similarity between current frame and the frame after the next frame is smaller than T_c, and the similarity between current frame and the previous frame is higher than T_{gh} which means these frames are not in a gradual transition. When all the conditions are satisfied, a cut transition detected between tth,$t + 1$th frame.

$$S(t, 2) < T_c, \quad S(t - 1, 1) > T_{gh} \tag{3}$$

Gradual transition detection. Different from cut transition,the gradual transition process lasts for many frames.We proposed a graudal boundary detection method based on dual-threshold sliding window.In this method,the detection steps can be described as:

(1).The similarity between adjacent frames is in $[T_{gl}, T_{gh}]$,which means transition continues in the window;

$$T_{gl} \leq S(t, i) \leq T_{gh}, \qquad \forall i \in [1, w] \tag{4}$$

(2).The window similarity which stands for the similarity between the first and last frame in the window should be smaller than the threshold of cut transition T_c,and the similarity between the last two frames in the window should be bigger than T_{gh};

$$S(t, w) < T_c, \qquad S(t + w - 1, 1) > T_{gh} \tag{5}$$

(3).If (1),(2) are both satisified,then frames in the window are considered as gradual transition frames,a gradual transition is detected.

(4).If (1) is satisified,(2) is not satisified,the tail of the sliding window move to the next frame.

(5).Other condition,reset the window position and move to next frame, util (1) is satisified and set the head and tail of the window to current position.

As 3 shows, the window similarity should be in a decreasing trend util the gradual transition completed. The flow the

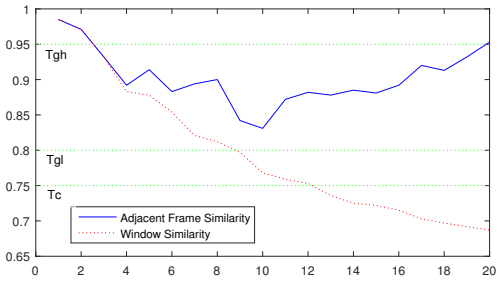


Fig. 3. Gradual shot boundary detection besed on sliding window

proposed method above can be described in Algorithm 3-1.Our method can delect cut and gradual transition in only one round of travel,the time comlexity is $O(n)$.

Algorithm 3-1 CFSSBD(CNN Feature Similarity based Shot Boundary Detection)

```
//Input: featureVectors
//Output:shotBoundaries
Initialize:
    frameCount ← N;
    threshW ← T_c;
    threshGH ← T_gh;
    threshGL ← T_gl;
    featureVectors ← F;
End
Begin:
    Input featureVectors ;
    t ← 0 ;// feature index
    w ← 0 ;// size of sliding window
    gp ← −1;// position where gradual transition starts
    While t < N − 1:
        w = t − gp;
        s_{t,1} ← S(t, 1);   //calculating the similarity between Nth
                and N − 1th frame,S is the similarity function
        ···   //calculating other similarities,here we omit them.
        If s_{t,1} < T_c and s_{t,2} < T_c and s_{t−1,1} > T_gh
        and gp == −1:
          Output a Cut Boundary [t, t + 1];//a cut transition boundary.
        Else If s_{t,1} <= t_gh and s_{t,1} >= t_gl
        and gp == −1:
            gp ← t ;   //a gradual transition starts.
        Else If s_{t,1} > t_gh and s_{t,w} < T_c and
        s_{t+w−1,w} > T_gh and gp! = −1:
            Output a Gradual Boundary [gp, t];
            gp ← gp − 1;
        Else If gp! = −1:
            gp ← −1;//if the gradual transition ending condition
        is not satisfied,reset the gradual transition initial position.
         t ← t + 1;
    End While;
End
```

IV. Experiments and Results

A. Dataset

In order to evaluate our method and compare with other methods, we downloaded 3 video clips(respectively:sport, movie, cartoon) from the internet and extracted frame sequence of each video.Then we manually marked the shot boundary indexes of transitions (cut or gradual) of each video.The statistical numbers of the 3 videos are listed in table I.

TABLE I
Transition statistical numbers.

Video Name	Total Frames	Transitions	Cuts	Graduals
Sport	2897	43	22	21
Movie	1867	156	92	64
Cartoon	2382	193	145	48

B. Evaluation

For each video, an array of shot boundary indexes was gotten ,comparing detection boundary results for each video to the shot boundary reference data constructed manually.then reporting the Precise/Recall/$F1$ score to evaluate the shot segmentation method.

Precise: Among the transitions (cut or gradual) detected ,how many were true transitions.

Recall: For all possible transitions (cut or gradual) we marked manually, how many were detected .

$F1$**:**Defined as the comprehensive evaluation index of precise and recall.

A good method should have high precise,high recall. If we let d be the number of transition frames detected,r be the number of transition frames we manually marked, and d_z be the number of transition frames which are detection are in the manually marked as transition frames, then precise and recall can be written as:

$$Precise = \frac{d_r}{d} \qquad Recall = \frac{d_r}{r} \tag{6}$$

$$F1 = \frac{2 * Precise * Recall}{Precise + Recall} \tag{7}$$

C. Experimental Result

While numerous experiments based on different methods were conducted, we campared our method to another six methods in pixel domain, within four classic methods, and two new methods [14] [13],the main thought of these methods are described as follows:

PSBD(Pixel based Shot Boundary Detection).This method processes the brightness,grayscale and color values of images, the principle is to calculate gray scale change of each pixel between two frames.The gray scale difference of pixel on the same position in two adjacent frames is:

$$f_d(i,j) = \mid f_{n+1}(i,j) - f_n(i,j) \mid \tag{8}$$

$f_n(i,j)$,$f_{n+1(i,j)}$ stand for the gray scale of the nth,$(n+1)$th frame.The total difference of two adjacent frames is:

$$F_d(f_n, f_{n+1}) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} f_d(i,j) \tag{9}$$

M, N is the size of an image.If the total frames difference is bigger than a threshold,then record a transition.

HSBD(Histogram based Shot Boundary Detection).It is the most common shot boundary detection method,it is simple and convenient, and for most videos it can get good results.It is developed from HSBD,it divides the gray scale into N grades for each pixel,statisitics the number of pixels for each gray scale grade to make histograms for two images,the histogram based frames' difference formula is described as follows:

$$D = \frac{1}{2N}\sum_{i} \mid h_m(i) - h_n(i) \mid \tag{10}$$

N is the number of pixels in an image.$h_m(i) - h_n(i)$ stand for the distance on the ith histogram.

X^2**HSBD**(X^2 histogram based Shot Boundary Detection).It amplies the difference of two frames,the difference of histogram can be normalized like:

$$X^2 = \begin{cases} \sum_{i=1}^{k} \frac{(h_m(i)-h_n(i))^2}{max(h_m(i),h_n(i))}, \\ (h_m(i) \neq 0 \vee h_n(i) \neq 0) \\ 0, \qquad\qquad \text{else} \end{cases} \tag{11}$$

k is the number of histogram grade.$h_m(i) - h_n(i)$ stands for the distance of ith histogram.The bigger X^2 is, the bigger the difference between two frames is, the converse is also true.

B-X^2**HSBD**(blocked X^2 histogram based Shot Boundary Detection).It is improved from X^2HSBD. In order to reduce the effect of motion and illumination,each frame had been divided into blocks and compared the histogram difference of each block,abandoned block with the maxinum difference,the formula is described as follows:

$$X^2 = \begin{cases} \sum_{i=1}^{k}\sum_{j=1}^{p} \frac{(h_m(i,j)-h_n(i,j))^2}{max(h_m(i,j),h_n(i,j))}, \\ (h_m(i,j) \neq 0 \vee h_n(i,j) \neq 0) \\ 0, \qquad\qquad \text{else} \end{cases} \tag{12}$$

Sun's method [14].They detect cut and gradual transitions in different at different scan of frames.They calculate a distance like 13-17 as the measurement of frames. The used a threshold to detect cuts, when the distance between current frame and a few consecutive frames are all bigger than the threshold,mark the current and the next frame as a cut boundary. Take frames in a one second segment to adopt a rule of calculating distance between different frames, when all conditions are satisified,then a gradual transition is detected.They used optical flow to exclude wrong detections caused by motion movement, and used SIFT to exclude wrong detections caursed by luminance mutation.

$$D_{t,n} = \frac{\sum_{x=0}^{W-1}\sum_{y=0}^{H-1} d_{t,n}(x,y)}{WH} \tag{13}$$

$$d_{t,n}(x,y) = \min_{i^2+j^2 \leq r^2} \frac{1}{3}[d_{R_{t,n}}(x,y,i,j) + d_{G_{t,n}}(x,y,i,j) + d_{B_{t,n}}(x,y,i,j)] \tag{14}$$

$$d_{R_{t,n}}(x,y,i,j) = |R_t(x,y) - R_{t+n}(x+i, y+j)| \tag{15}$$

$$d_{G_{t,n}}(x,y,i,j) = |G_t(x,y) - G_{t+n}(x+i, y+j)| \tag{16}$$

$$d_{B_{t,n}}(x,y,i,j) = |B_t(x,y) - B_{t+n}(x+i, y+j)| \tag{17}$$

$R_t(x,y), G_t(x,y), B_t(x,y)$ and $R_{t+n}(x,y), G_{t+n}(x,y)$, $B_{t+n}(x,y)$ represent the R, G, B value of pixel (x,y) respectively, W, H are width and height.

We compared the our similarity with Sun's distance between adjacent frames,shown in Fig.4,we have normalized their distance in range [0,1] to make it comparable with our similarity. The position of difference are basically consistent,but our polyline is much clear at the transition position according to the video frames.

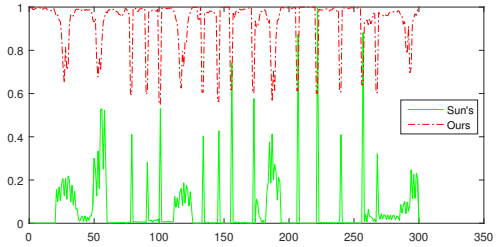


Fig. 4. Comparison of our similarity and Sun's distance

Lu's method. The adopted the normalized hue-saturation-value (HSV) color histograms as frame features, and applied candidate segment selection to improve the singular value decomposition based method [15] and speeded up shot boundary detection.More detail is described in [13].

The presents experimental results in Fig.5, provide interesting and meaningful contrasts. We can see that the method of shot boundary detection we proposed outperforms other methods on all 3 videos.Within PSBD,HSBD,X^2HSBD,B-X^2HSBD perform the worst, so we compared Lu et al. [13] and Sun et al. [14] with our method with the best parameter,the results list in II-IV,it is not difficult to see that our method has obvious advantages,the mean value of $F1$ in cut transition detection is 0.961,in gradual transiton detection is 0.886,overall is 0.914.

We also compared the calculation performance,see table V,as we used GPU to accelerate the calculation,and PSBD,HSBD,X^2HSBD,Sun's are accelerated by multithreading . Result shows that the speed of our method is close to the simplest method PSBD.

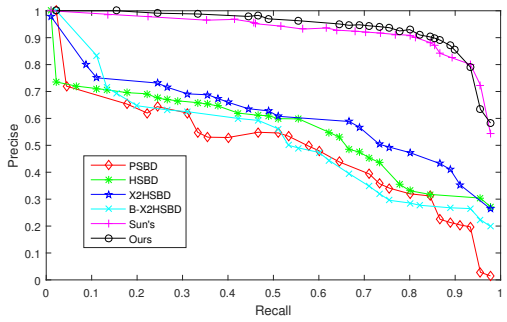
V. Conclusion

In this paper, we proposed a new video shot boundary detection method based on CNN feature, for cut boundary detection,we based the similarity of local frames to get more accuracy cut detection and proposed dual-threshold sliding

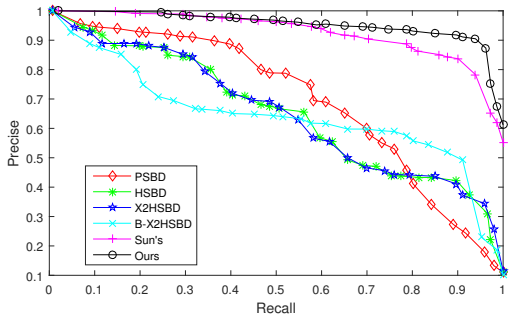
window for gradual transition detection.The experimental results shows the outstanding performance at $F1$ score and speed .In the future, it may be productive to add motion features or different kind of gradual transition feature to this task in order to get more accurate shot detection for gradual transition.

VI. Acknowledgement

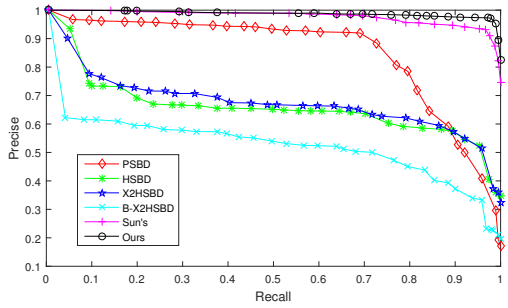
This work was partially sponsored by The National Natural Science Foundation of China(Grant No. 61300192) and the Fundamental Research Funds for the Central Universities(Grant No.ZYGX2014J052).



(a).PR curve of the 'sport' video.



(b).PR curve of the 'movie' video.



(c).PR curve of the 'cartoon' video.

Fig. 5. Comparison of Shot Boundary Detection Algorithm.

References

[1] A. SenGupta. A Formal Study of Video Segmentation [J] International Journal of Innovations in Engineering and Technology(IJIET), 2015,5(2):54-60.

[2] Hanjalic A. Shot-boundary detection:unraveled and resolved.IEEE Circuits and Systems for Video Technology, 2002,12(2):90-105

TABLE II
Comparison of cut transition detection performance.

Videos	Cut Transitions								
	Recall			Precise			F1		
	Sun's	Lu's	Ours	Sun's	Lu's	Ours	Sun's	Lu's	Ours
Cartoon	0.966	0.972	**0.979**	0.933	0.959	**0.972**	0.949	0.965	**0.976**
Sport	**0.954**	0.864	0.909	**0.875**	0.826	0.870	**0.913**	0.844	0.889
Movie	0.950	0.937	**0.965**	0.931	0.904	**0.951**	0.941	0.920	**0.958**
Mean	0.958	0.948	**0.968**	0.928	0.924	**0.955**	0.943	0.936	**0.961**

TABLE III
Comparison of gradual transition detection performance.

Videos	Cut Transitions								
	Recall			Precise			F1		
	Sun's	Lu's	Ours	Sun's	Lu's	Ours	Sun's	Lu's	Ours
Cartoon	0.886	0.905	**0.924**	0.842	0.888	**0.890**	0.863	0.897	**0.907**
Sport	0.872	0.824	**0.896**	0.807	0.769	**0.848**	0.838	0.795	**0.872**
Movie	0.834	0.806	**0.900**50.724	0.726	**0.845**	0.775	0.764	**0.874**	
Mean	0.863	0.848	**0.910**	0.787	0.796	**0.863**	0.823	0.821	**0.886**

TABLE IV
Comparison of overall performance.

Videos	Cut Transitions								
	Recall			Precise			F1		
	Sun's	Lu's	Ours	Sun's	Lu's	Ours	Sun's	Lu's	Ours
Cartoon	0.932	0.944	**0.956**	0.895	0.929	**0.937**	0.913	0.937	**0.947**
Sport	0.893	0.834	**0.899**	0.825	0.783	**0.854**	0.858	0.808	**0.876**
Movie	0.901	0.881	**0.939**	0.837	0.826	**0.905**	0.868	0.852	**0.922**
Mean	0.89	0.879	**0.917**	0.859	0.863	**0.911**	0.875	0.871	**0.914**

TABLE V
Speed of process frames.

Algorithm Name	Frame Per Second
PSBD	124.25
HSBD	28.59
X^2HSBD	28.86
B-X^2HSBD	17.82
Sun's	0.32
Lu's	438.27
Ours	85.35

[3] Sun J.F,Li Y.X.Some methods of automatic video shot segmentation. Journal of South China University of Technology(Natural Science),2003,31(8):10-14.

[4] Hou G.H,Shi.P.The research of video segementation and scene clustering algorithms.Journal of Communication University of China Science and Technology,2006,13(2):32-37

[5] Tsamoura E, Mezaris V, Kompatsiaris I. Gradual transition detection using color coherence and other criteria in a video shot meta-segementation framework. IEEE, 2008:45-49.

[6] Priya G G L, Domnic S. Video cut detection using block based histogram differences in RGB color space[C]// International Conference on Signal and Image Processing. 2010:29-33.

[7] Yang L, Lu H, Wang B, et al. Shot Boundary Classification and Refinement Using Inter-Frame Similarity Patterns[C]// International Conference on Information Communications & Signal Processing. IEEE, 2005:673-677.

[8] Cernekova Z, Pitas I, Nikou C. Information theory-based shot cut/fade detection and video summarization[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2006, 16(1):82-91.

[9] Zabih R, Miller J, Mai K. A Feature-Based Algorithm for Detecting and Classifying Production Effects[J]. Multimedia Systems, 1999, 7(2):119-128.

[10] Zhang S.J.Image Enhancement Method of vehicle Plate based on Canny

Edge Detection.Journal of ChongQing JiaoTong University(Nature Science),2012:1-4
[11] Li Y N, Lu Z M, Niu X M. Fast video shot boundary detection framework employing pre-processing techniques[J]. Image Processing, IET, 2009, 3(3):121-134.
[12] Grana C, Cucchiara R. Linear Transition Detection as a Unified Shot Detection Approach[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2007, 17(4):483-489.
[13] Lu Z M, Shi Y. Fast Video Shot Boundary Detection Based on SVD and Pattern Matching[J]. Image Processing IEEE Transactions on, 2013, 22(12):5136-5145.
[14] J. Sun, Y. Wan. A novel metric for efficient video shot boundary detection[M]. 2015, 4548
[15] Cernekova Z. Video shot-boundary detection using singular-value decomposition and statistical tests[J]. 2007, 16(4):043012.
[16] Jinhui Yuan, Huiyi Wang, Lan Xiao, Wujie Zheng, Jianmin Li, Fuzong Lin, Bo Zhang: A Formal Study of Shot Boundary Detection. IEEE Trans. Circuits Syst. Video Techn. 17(2): 168-186 (2007).
[17] KRIZHEVSKY A, SUTSKEVER I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in Neural Information Processing Systems, 2012, 25(2):1106-1114.
[18] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[EB/OL]. [2016-09-14].