Shot Boundary Detection with Spatial-Temporal Convolutional Neural Networks

Lifang Wu, Shuai Zhang, Meng Jian(✉), Zhijia Zhao, and Dong Wang

Faculty of Information Technology, Beijing University of Technology, Beijing, China
lfwu@bjut.edu.cn, zhangshuai0212@emails.bjut.edu.cn, jianmeng648@163.com, zhaozhijia0229@163.com, dwang@nlpr.ia.ac.cn

Abstract. Nowadays, digital videos have been widely leveraged to record and share various events and people's daily life. It becomes urgent to provide automatic video semantic analysis and management for convenience. Shot boundary detection (SBD) plays a key fundamental role in various video analysis. Shot boundary detection aims to automatically detecting boundary frames of shots in videos. In this paper, we propose a progressive method for shot boundary detecting with histogram based shot filtering and C3D based gradual shot detection. Abrupt shots were detected firstly for its specialty and help alleviate locating shots across different shots by dividing the whole video into segments. Then, over the segments, gradual shot detection is implemented via a three-dimensional convolutional neural network model, which assign video clips into shot types of normal, dissolve, foi or swipe. Finally, for untrimmed videos, a frame level merging strategy is constructed to help locate the boundary of shots from neighboring frames. The experimental results demonstrate that the proposed method can effectively detect shots and locate their boundaries.

Keywords: Shot boundary detection · Shot transition
Video indexing · Convolutional neural networks
Spatial-temporal feature

1 Introduction

With the development of multimedia and network technologies, a large amount of videos are uploaded to the internet, rapid video understanding and content-based retrieving become serious problems. To complete content-based video retrieval, that is, to quickly and efficiently retrieve the user's desired video content from the video database, the first task is to structure the unstructured video sequence. Typically, a top-down multi-level structure model can be used to hierarchically represent video. The top layer is the video layer, which is composed of multiple different scenes; Following is the scene layer, which is composed of a plurality of shot segments that are related to each other in terms of content; The third layer

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11257, pp. 479–491, 2018.
https://doi.org/10.1007/978-3-030-03335-4_42

is the shot layer, which consists of multiple consecutive frames; the bottom is the frame layer, which is the smallest unit of video.

A video shot is defined as a sequence of frames taken by a single camera. The quality of shot boundary detection will affect the efficiency of video retrieval and video semantic analysis. Shot boundary detection, also known as shot transition detection, is the basis of video retrieval. Shot transition refers to the transformation from one continuous video image sequence to another. These transformations can be classified into two main categories: abrupt and gradual shot transition, which is shown in Fig. 1. An abrupt shot change completed between two frames, while gradual transitions occur over multiple frames. Gradual transition can be further classified into dissolve, swipe, fade in and fade out.

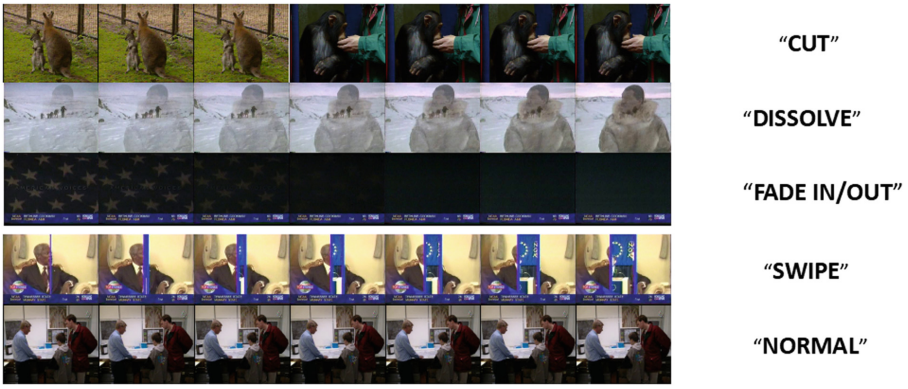


Fig. 1. Cut and gradual shot change.

Until now great efforts have been made in SBD, but most of them focus on abrupt shot change detection. Considering that two frames have great dissimilarity, these approaches usually extract features of consecutive frames and measure the similarity between features; When similarity exceeds a threshold, then an abrupt transition is detected. Compared with abrupt transition, gradual transition is more difficult because a gradual shot transition usually consists of a few frames and the difference between adjacent frames is not obvious. Traditional SBD techniques analyze hand-crafted features. The pixel comparison method is the simplest shot boundary detection algorithm and the theoretical foundation of most other algorithms. By comparing the density values or color difference values in adjacent video frames, this method is to calculate the absolute sum of the pixel differences and compare them with the threshold [1]. Obviously, this method is very sensitive to the motion of objects and cameras. Thus [2] uses a 3×3 filter to smooth the image, then traverses all the pixels in the image, and counts the number of pixel points where the pixel value difference is greater than a certain threshold, which is treated as the frame of the two frames. Compared to template matching methods based on global image feature (pixel difference)

comparison methods, block-based comparison methods [3] use local features to increase the robustness to camera and object motion. When the shot changes, the edges between frames of different shots also change. Zabih et al. proposed the use of Edge Change Ratio to detect the shot boundary [4]. Individual edge features are not suitable for detecting shot boundaries because sometimes the edge is not clear. Some studies have shown that edge features can be a good complement to other features. Because the edge features are insensitive to brightness changes, Kim and Heng avoid the false detection of shot boundaries caused by flashlight by detecting the edge features of candidate shot boundaries, and does not significantly increase the computation time [5,6]. Among the shot boundary detection algorithms, the histogram method is a method that is frequently used to calculate the difference between image frames [7]. Zhang et al. [2] compared the two methods of pixel values and histograms and found that the histogram can meet the requirements of the speed and accuracy of video edge detection. Ueda uses color histogram change rates to detect shot boundaries [8]. In order to improve the difference between two video frames, the author proposed to use x^2 histogram to compare the color histogram differences between two adjacent video frames [9]. There are several ways to calculate the difference between two histograms. Some studies have shown that the Manhattan distance and the Euclidean distance are simple and effective frame difference calculation methods [10]. Zuzana uses mutual information to calculate the similarity of images, the larger the mutual information, the more similar the two images are [11].

Compared with abrupt shot transition, gradual shot transition is more difficult because a gradual shot transition usually consists of a few frames and the difference between adjacent frames is not obvious. The dual threshold method [12] is a classic method of shot detection which requires setting two thresholds. But it has a major problem: the starting point of the gradual is difficult to determine. The basic principle of the optical flow method [13] is that there is no optical flow when the gradual transition is happening, and the movement of the camera should be suitable for a certain type of optical flow. This method can achieve better detection results, but its calculation process is quite complicated.

The methods mentioned above were relied on low-level features of visual information. Recently, the use of deep neural network has attracted extensive attention from researchers, and it has also achieved a major breakthrough in both accuracy and efficiency compared to traditional manual feature methods. However, relatively few studies have applied deep learning to the field of shot boundary detection. [14] present a novel SBD framework based on representative features extracted from CNN which achieves excellent accuracy. [15] present a CNN architecture which is fully convolutional in time, thus allowing for shot detection at more than 120x real-time. [16] presents DeepSBD model for shot boundary detection and build two large datasets, but did not distinguish the specific gradual shot type. Inspired by these works, we first use the histogram method to detect the abrupt shots; Based on this, C3D model is employed to extract time domain features and achieve gradual shot detection.

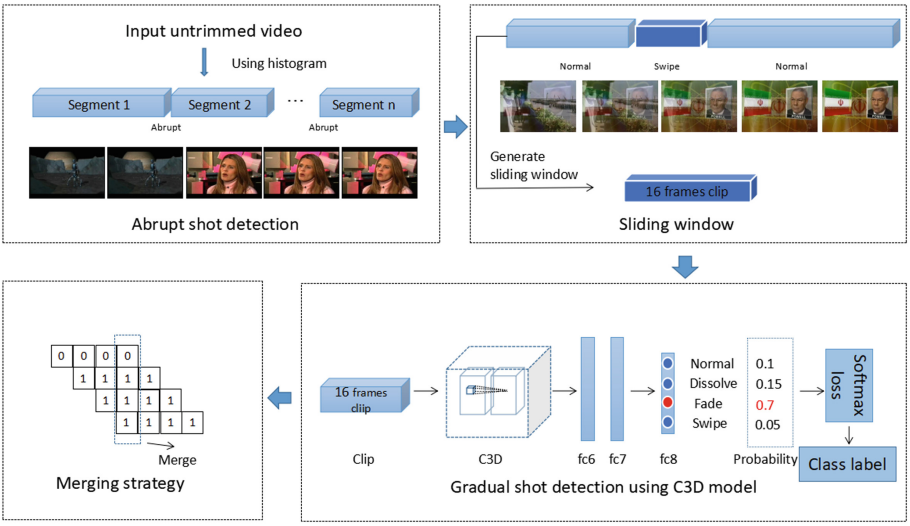


Fig. 2. Framework of the proposed method.

In this work, we investigate spatial and temporal features of videos jointly with a deep model and proposed shot boundary detection with spatial-temporal convolutional neural networks. Figure 2 illustrates the framework of the proposed shot boundary detection method. The proposed method implements shot boundary detection in a progressive way of detecting abrupt and gradual shots. The abrupt shots are firstly extracted from the whole video with histogram base shot filtering. Then, C3D deep model is constructed to extract features of frames and distinguish shot types of dissolve, swipe, fade in and fade out, and normal. The main contributions of this work are summarized as follows:

- Considering different changing characteristics of abrupt shots from gradual shot, a progressive method was proposed to distinguish abrupt and gradual shots separately.
- Joint spatial and temporal feature extraction with C3D model help effectively distinguish different gradual shot types from video segments.
- We further develop a frame level merging strategy to determine temporal localization of shot boundaries.

2 Proposed Method

In this section, we describe the details of the proposed progressive method for shot boundary detection. As illustrated in Fig. 2, the proposed method performs shot boundary detection in a hierarchical manner. Histogram based abrupt shot filtering is employed to detect abrupt shots, which further help divide the whole video into segments with the abrupt frames as boundaries. Then, a 16 frames

sliding window is conducted with 87.5% overlap, which will be the inputs of C3D model. Next, C3D model extracts spatial-temporal features of sequential frames from the segments to assign clips with a shot type of dissolve, swipe, fade in and fade out, and normal. Last, a frame level merging strategy is proposed to well locate shot boundaries.

Table 1. Utilization popularity of different features in shot boundary Detection algorithms [17].

Luminance and color	Histogram analysis	Edge	Transformation coefficients (DCT,DFT etc.)	Statistical measurement	Motion analysis	Object detection
4	28	3	7	16	13	2
8%	56%	6%	14%	32%	26%	4%

2.1 Abrupt Shot Detection

Considering imbalance of frames of abrupt shot compared with the other gradual shot, in this work we intuitively construct a hierarchical progressive framework to detect abrupt shot firstly and distinguish the other shot type with a deep model. According to [17], more than 50 investigation works were published on shot boundary detection from 1996 to 2014 with various feature descriptors. [17] provided a summary of utilization popularity summarization on feature descriptor in shot boundary detection as Table 1, which indicated that histogram based shot analysis is the most popular feature with 56% utilization rate compared with traditional luminance and color, edge, transformation coefficients, statistical, motion and object based descriptors. Therefore, we employ histogram in abrupt shot detection considering the lack of frames of the abrupt shot type for feature learning. Then we perform histogram based abrupt shot filtering by measuring difference of neighboring frames. Indeed, researches in similar domains have indicated that the simple Manhattan distance metric is highly effective. We conduct histogram based abrupt shot filtering with Manhattan distance to measure the difference of neighboring frames as follows.

$$D_{Manhattan}(H_i, H_j) = \sum_{k=1}^{binNum} |h_{ik} - h_{jk}| \tag{1}$$

where H_i and H_j denote histograms of the $i-$th frame and $j-$th frame of an video, h_{ik} is the value of the k$-$th bin of the histogram H_i, and binNum is the number of bins of the histogram. In this work, we take 64 bins for each channel of R,G,B. The smaller the distance is, the more similar the histograms of the two frames are. Otherwise, the difference between the two frames is greater. If the difference between frames exceeds a given threshold, the two frames is treated as an abrupt shot.

2.2 Gradual Shot Detection

Pre-processing. With the detected abrupt shot frames, the whole video has been divided into several segments of frames. As mentioned in the proposed framework, the divided segments are given to a deep learning model C3D for feature extraction and shot type classification. The networks are set up to take video clips as inputs and predict the shot category which belong to 4 different shot types, i.e., dissolve, swipe, fade in and fade out, and normal. All video frames are resized into 128×171. Videos are split into variable length−overlapped 16−frame clips which are then used as input to the network. The input dimensions are $3 \times 16 \times 128 \times 171$. We also use jittering by using random crops with a size of $3 \times 16 \times 112 \times 112$ of the input clips during training.

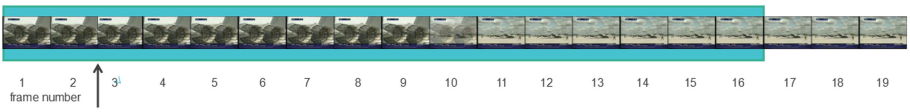


Fig. 3. A varied-length sliding window used for sampling video clips.

Sliding Window. Considering that the length of each gradual shot transition type is different, we need to select the clip length of each type and the step size of the window according to the length of the shot type. Therefore, we calculate the length of the gradual transition events used by TRECVid 2003-2007. The statistical results show that the average length of SWIPE, DIS and FOI are 12.5, 21.9 and 29.9 frames, respectively. Therefore, 16 frames used by Trans [18] is the appropriate length. Based on the length of different events, a sliding window with step size of 2, 7, 5 is used to sample the video clips to ensure the equalization of the samples, as shown in Fig. 3.

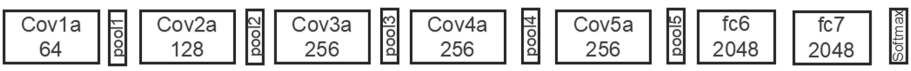


Fig. 4. The structure of C3D model for spatial temporal feature learning.

C3D Based Gradual Shot Detection. Most previous works for detecting gradual shots are mainly based on hand crafted features. These low level visual features are lack of describing ability on high level semantic information. Feature extraction based on deep neural network is favored by researchers because it can better reflect the information of the data itself and does not require researchers to

have a large number of domain related information as manual features. The shot transition features that need to be extracted are in time series, so the solution to the SBD task is more dependent on the extracted temporal features. Therefore, we use a three dimensional CNN–C3D mode to extract temporal features automatically. Compared with a two−dimensional CNN, Three−dimensional CNN not only divides video into frames, but also applies convolution kernel to both spatial domain and temporal domain. Combining with each other, the video features can be obtained better. As depict in Fig. 4, the C3D mode includes five convolution layers, five pool layers, two fully connection layers to learn features, and one softmax layer to provide the predicted class. The number of convolution kernels of the five convolution layers is 64, 128, 256, 256, and 256, the optimal size of the convolution kernel is $3 \times 3 \times 3$, After one convolution operation, the features are downsampled by one pooling operation to gradually reduce the scale of feature map, reduce the difficulty of training and improve the accuracy of training. In this paper, the convolution kernel size of the pooling layer from the second to the fifth are $2 \times 2 \times 2$, while that of the first layer is $1 \times 2 \times 2$, so that in the early stage, the time domain information in the network can be preserved to the maximum extent. After several convolution and pool operations, the feature map is abstracted into a 2048 dimensional feature vector to mark the classification information of the sample.

Post-processing. We tested on both trimmed videos and untrimmed videos. For trimmed videos, we used a method similar to video classification, which took a part of the sample as test set and got the classification result; For untrimmed videos, We proposed frame level merging strategy, which is shown in Fig. 5. After detecting the location of the abrupt shot and divide the video into video segments, we conduct temporal sliding windows of 16 frames with 87.5% overlaps over the video segment. We visualize the output of the model's softmax layer so that we get a classification result every 16 frames, which take the label with the highest probability as the prediction label for this video clip; Then each frame in the video clip is given the same label as the video clip, so that each frame will be combined with the results of multiple clips to get a prediction result. Each frame will have a maximum of 8 labels, we take label with maximum amount as the frame label.

3 Experiment

TRECVid Dataset. From 2001 to 2007, NIST held a competition from 2001 to 2007 and provided data sets and shot type labels for the contest. The data for each year have been a representative, usually random, sample of approximately 6 h of the video The origins and genre types of the video data have varied widely from the initial NIST and NASA science videos in 2001, to the Prelinger Archive's antique, ephemeral video, to broadcast news from major US networks in the mid−1990's to more recent Arabic and Chinese TV news programming. Editing styles have changed and with them the shot size and distribution of shot

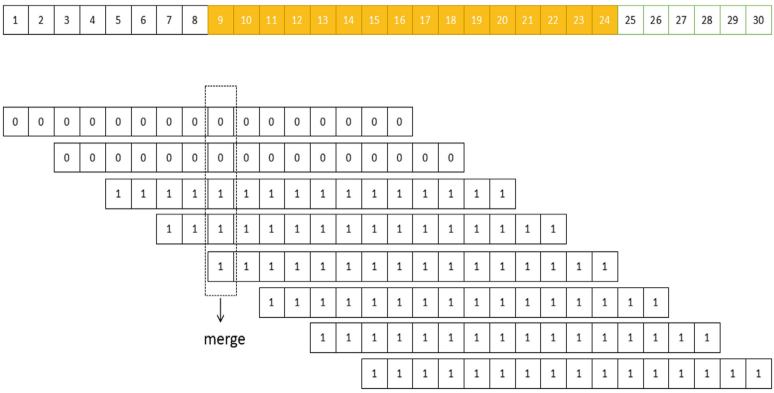


Fig. 5. Frame level merging strategy, where the frames 9 to 24 represent a gradual shot change. By moving sliding window with 87.5% overlap, Each frame will be combined with the results of multiple clips to get a prediction result. Each frame will have a maximum of 8 labels, we take label with maximum numbers as the frame label.

transitions types. The TRECVid dataset is not public available and requires an application from NIST to obtain it. We collected all the SBD related dataset from the year 2003 to 2007.

Implementation Details. The proposed framework is built on the deep learning model of caffe C3D. The whole network is trained from scratch and the parameter are set as mini-batch size of 20, base learning rate to 1×10^{-4}, momentum to 0.9, weight decay to 5×10^{-5}, and maximum number of training iterations to 60000. The learning rate is divided by 10 every 10000 iterations. All experiments are conducted on Nvidia Titan X GPU with Intel(R) Xeon(R) CPU E5-2683 v3 @ 2.00GHz, running a Ubuntu 14.04 LTS environment and python 2.7.12.

Abrupt Shot Change Detection. For abrupt shot detection, We took 10 videos from the TRECVid data set for testing. The threshold is set to 0.2, which is based on the results of the experiment. Detecting results are partly shown in Table 2 (due to limitation of length). The overall F score is 0.899, which is effective; What's more, the execution time took only 0.11 of real time, which is very fast.

Gradual Shot Change Detection. For gradual shot change detection, we use five years of data from the TRECVid competition to train and test. The data of TRECVid 2005 is used for testing and is not included in the training set. By sampling windows with different lengths, the data set is sample-equalized. The number of training set, verification set, and testing set are 7015, 2618, and

Table 2. The results of abrupt shot detection using the histogram comparison method.

Video	Total transition	Detected	Misdetected	Recall	Precision
BG_2408	121	100	2	82.64%	98%
BG_9401	90	79	7	87.78%	92%
...	...	...	...	...	...
BG_14213	111	91	3	81.98%	97%
BG_34901	224	204	20	91.07%	91%
Total	1146	1019	97	88.92%	91%

2752 video segments, respectively, and the ratio is roughly 3:1:1. Table 3 is data distribution of the data set.

Table 3. Data distribution of the data set.

Shot Type	Train	Val	Test
NOR	1799	677	630
DIS	1821	543	752
FOI	1847	670	504
SWIPE	1548	728	866

The results of the shot detection after trimming are shown in Table 4. The results show that the proposed method can effectively extract and identify the time domain features of different shot transition types. The average detection accuracy is 89.4%.

Table 4. Confusion matrix for trimmed video.

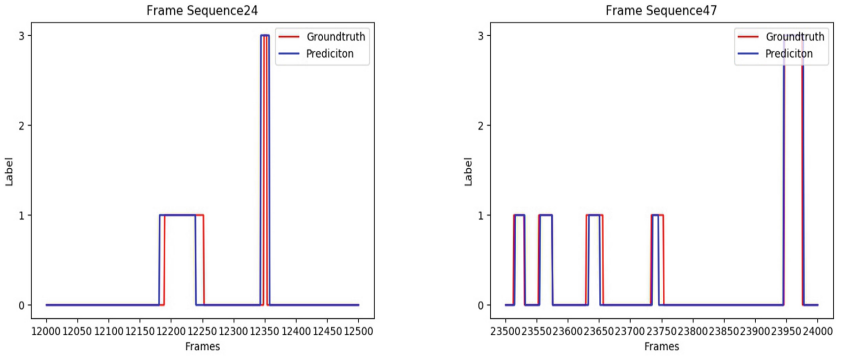
	NOR	DIS	FOI	SWIPE	Accuracy
NOR	528	12	18	72	83.81%
DIS	32	618	60	42	82.18%
FOI	0	0	493	11	97.82%
SWIPE	10	11	23	822	94.97%

Further, we tested on untrimmed video of TRECVid 2005. The frame level comparison results are shown in Table 5, where the overall detection accuracy is 88.8%. For more intuitive representation, we randomly selected several segments, as shown in Fig. 6, where the x-axis indicates the frame number, and the y-axis indicates shot types. The labels 0, 1, 2, 3, 4 represent normal shots, dissolves, fades, swipes, and abrupt shots respectively.

It is easy to find that the accuracy of localizing might affects the results to some extent. In this regard, we have performed statistics on localizing accuracy of TRECVid 2005 data set. The results are shown in Fig. 7. Where the x-axis represents the degree of overlap, and the y-axis represents the proportion of each gradual shot change type. Experiments show that the overlapping degree between the prediction result and Groundtruth is centered on 0.9-1, and the overlap degree of FOI is all greater than 0.6. This indicates that this method not only can identify different shot types, but also can accurately locate events.

Table 5. Confusion matrix for untrimmed video.

	NOR	DIS	FOI	SWIPE	CUT	Accuracy
NOR	374769	11785	2571	30011	112	89.39%
DIS	771	6482	351	1877	44	68.05%
FOI	67	66	1918	68	26	89.42%
SWIPE	205	147	553	2720	18	74.66%
CUT	97	27	25	42	3059	94.12%



(a) (b)

Fig. 6. Frame level comparison. The x-axis indicates the frame number, and the y-axis indicates shot types; The red line represents the Groundtruth result, and the blue line represents prediction result by our model. (Color figure online)

We use Trecvid 2005 data as a test set and compare it with other methods. The comparison results are shown in Table 6. The Best TRECVid performer [19] make use of support vector machine classifiers (SVMs) to help detect either cuts or gradual transitions and use color features to train the classifier, and the accuracy is 78.6%. In comparison, Our method extracts the deep features that

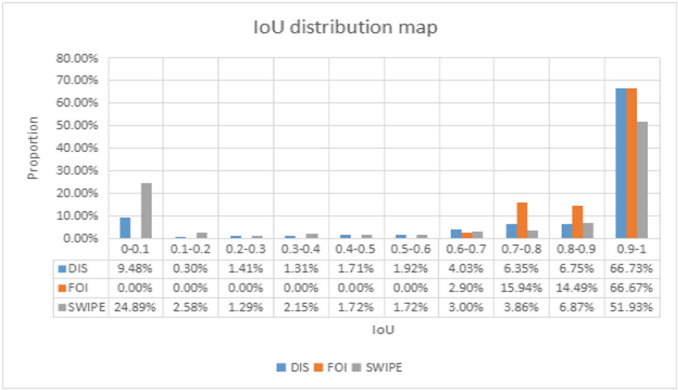


Fig. 7. IoU distribution of test video, where the x-axis represents the degree of overlap and the y-axis represents the proportion of events.

combine the relationships between frames through convolutional neural network, which improves the result by nearly 10%; Compared with the results of another deep learning model LSTM, Our proposed method increases the accuracy by 18%; DeepSBD [16] divides the shot transition types into abrupt transition, gradual transition and no-transition. Instead, the progressive framework we used first detects abrupt shots and only distinguish three types of gradual transition and no-transition, which effectively avoid the misdetection caused by the abrupt shots which lack of time domain information and increase the accuracy by 4%.

Table 6. Comparing against different techniques.

MODEL	Accuracy
CNN+LSTM	0.708
Best Trecvid Performer	0.786
DeepSBD	0.844
OURS(untrimmed)	**0.888**
OURS(trimmed)	**0.894**

4 Conclusion

In this paper, we have proposed a progressive method for shot boundary detection task. Our method employs histogram comparison method to detect abrupt shot changes, which effectively avoid the misdetection caused by the abrupt shots which lack of time domain information. Moreover, C3D model performs to

extract temporal features and distinguish different types of gradual shot transitions. The experiments are conducted on TRECVid data set and the results demonstrate that the proposed method are feasible and effective for detecting shot boundaries and its position.

Acknowledgements. This work was supported in part by Beijing Municipal Education Commission Science and Technology Innovation Project under Grant KZ201610005012, in part by Beijing excellent young talent cultivation project under Grant 2017000020124G075 and in part by China Postdoctoral Science Foundation funded project under Grant 2017M610027, 2018T110019.

References

1. Wang, J., Li, J., Gray, R.: Unsupervised multiresolution segmentation for images with low depth of field. IEEE Trans. Pattern Anal. Mach. Intell. **2**(5), 99–110 (2002)
2. Zhang, H., Kankanhalli, A., Smoliar, S.: Automatic partitioning of full-motion video. Multimed. Syst. **1**(1), 10–28 (1993)
3. Lefèvre, S., Vincent, N.: Efficient and robust shot change detection. J. R.-Time Image Process. **2**(1), 23–34 (2007)
4. Zabih, R., Miller, J., Mai, K.: Feature-based algorithms for detecting and classifying scene breaks. Proc. ACM Multimed. **7**(2), 189–200 (1995)
5. Sang, H., Kim, R.: Robust video indexing for video sequences with complex brightness variations (2002)
6. Wei, J., Ngan, K.: High accuracy flashlight scene determination for shot boundary detection. Signal Process. Image Commun. **18**(3), 203–219 (2003)
7. Feng, H., Yuan, H., Wei, M.: A shot boundary detection method based on color space. In: Proceedings of the International Conference on E-Business and E-Government, pp. 1647–1650. IEEE (2010)
8. Ueda, H., Miyatake, T., Yoshizawa, S.: IMPACT: an interactive natural-motion-picture dedicated multimedia authoring system. Proc. Chi **7**(7), 343–350 (1991)
9. Nagasaka, A., Tanaka, Y.: Automatic video indexing and full-video search for object appearances. Ipsj J. **33**, 113–127 (1992)
10. Cheng, C., Lam, K., Zheng, T.: TRECVID2005 Experiments in The Hong Kong Polytechnic University: Shot Boundary Detection Based on a Multi-Step Comparison Scheme. TREC Video Retrieval Evaluation Notebook Papers (2005)
11. Cernekova, Z., Pitas, I., Nikou, C.: Information theory-based shot cut/fade detection and video summarization. IEEE Trans. Circuits Syst. Video Technol. **16**(1), 82–91 (2005)
12. Wang, J., Luo, W.: A self-adapting dual-threshold method for video shot transition detection. In: Proceedings of the IEEE International Conference on Networking, Sensing and Control, pp. 704–707. IEEE (2008)
13. Zhang, H., Wu, J., Zhong, D.: An integrated system for content-based video retrieval and browsing. Pattern Recognit. **30**(4), 643–658 (1997)
14. Xu, J., Song, L., Xie, R.: Shot boundary detection using convolutional neural networks, In: Proceedings of Visual Communications and Image Processing, pp. 1–4. IEEE (2016)
15. Gygli, M.: Ridiculously Fast Shot Boundary Detection with Fully Convolutional Neural Networks (2017)

16. Hassanien, A., Elgharib, M., Selim, A.: Large-scale, fast and accurate shot boundary detection through spatio-temporal convolutional neural networks (2017)
17. Pal, G., Rudrapaul, D., Acharjee, S.: Video shot boundary detection: a review. In: Satapathy, S., Govardhan, A., Raju, K., Mandal, J. (eds.) Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2. Advances in Intelligent Systems and Computing, vol. 338, pp. 119–127. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-319-13731-5_14
18. Du, T., Bourdev, L., Fergus, R.: Learning spatiotemporal features with 3D convolutional networks, pp. 4489–4497 (2014)
19. Smeaton, A., Over, P., Doherty, A.: Video shot boundary detection: seven years of TRECVid activity. Comput. Vis. Image Underst. **114**(4), 411–418 (2010)