BUPT at TRECVID 2007: Shot Boundary Detection

Zhi-Cheng Zhao, Xing Zeng, Tao Liu, An-Ni Cai
Multimedia Communication and Pattern Recognition Labs,
School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract

In this paper we describe our methodologies and evaluation results for the shot boundary detection at TRECVID 2007. We submitted 10 runs results based on SVM classifiers and several separate detectors.

BUPT_01	Default SVM parameters and a low threshold for motion detector
BUPT_02	Default SVM parameters and a low threshold for edge detector
BUPT_03	Make high penalty for false cuts to increase the precision
BUPT_04	Make more penalty for false cuts to increase the precision
BUPT_05	Make high penalty for false gradual transitions (GT) to increase precision
BUPT_06	Make high penalty for missing GT to increase recall
BUPT_07	Enhance motion threshold to increase the precision of GT
BUPT_08	Extend the filtering window of motion to increase both recall and precision of GT
BUPT_09	Increase the recall of both cuts and GT
BUPT_10	decrease motion threshold to increase the recall of GT

Evaluation results showed that our CUT algorithm can achieve a satisfying result while the GT dose not work as well as our previous testing.

1. Introduction

This is the first time that our group (Beijing University of Posts and Telecommunications, BUPT) participated in TRECVID. We take part in the shot boundary detection (SBD) task, and present a corresponding detection system to realize a fast, effective and tractable SBD. Our system consists of five components, including a MPEG decoder and features vector generation module, a CUT detector, a FOI detector, a gradual transition (GT) detector, and a motion detector. Two support vector machines (SVM) are used to detect the CUT and GT respectively. After these detectors make decisions successively, the locations of shot boundaries and types are obtained. 40 runs are generated from the system with different parameters. Among them, random 10 runs are submitted for evaluation.

The rest of this paper is organized as follows. Section 2 introduces our system framework. In Section 3, a detailed methodology of every module is presented. Section 4 is evaluation and conclusion.

2. System Framework

Our system framework is shown in Fig.1. The flowchart consists of five parts: the 1st one is an initial process which includes the MPEG decoding and feature vector generation. In order to implement it at real-time, we consider detecting shot boundaries in a sliding window L, which means to detect while video is decoded.

And then, several visual-aural features such as HSV histogram, RGB histogram, phase histogram of motion vectors (MV), MFCC, ordinal measures of luminance [1] and so on, are extracted to form the feature vector. Meanwhile, the inter-frame differences in L are statistically computed by different distance metrics [1-2]. In addition, considering the content correlation among the shots and the inner of shots, we respectively construct feature ties for the CUT and GT detector. The feature tie is a group of features of all frames in a local sliding window with length L_1 and centered at the current frame. The rest modules will be described in next section in details.

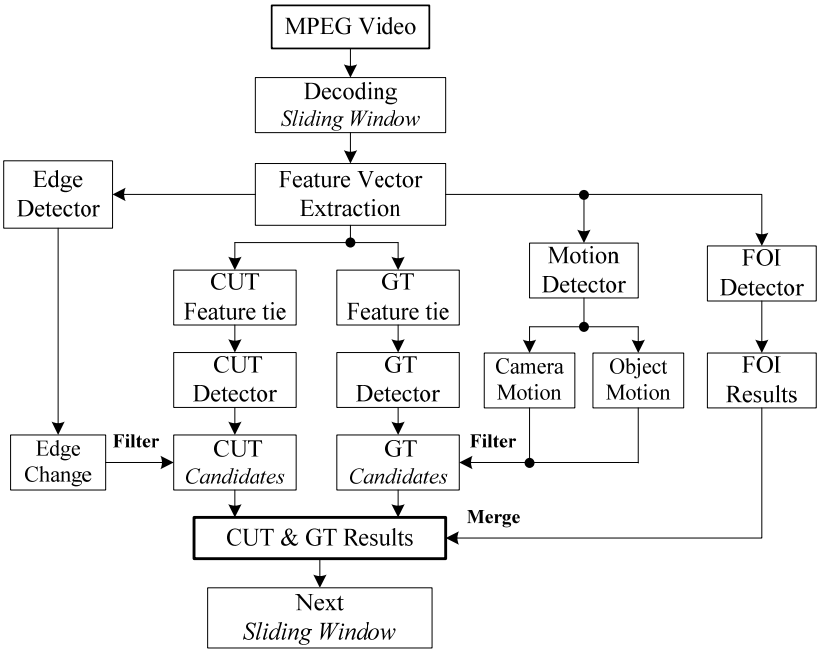

Figure.1 The proposed framework of SBD

3. Shot Boundary Detection

3.1 CUT detector

So far, a lot of CUT detection approaches have been presented and a good performance has been achieved [3]. However most of methods are sensitive to flashlight and a kind of edit effect so-called "video-in-video". In this paper, we adopt a post-processing to remove influence of them. The final CUT detection process is implemented by the following steps:

1. Training. A training dataset of more than 20000 samples based on the feature tie is formed. The ratio of positive samples to negative one is about 1:5. Then, a SVM is trained as the cut detector.

2. Testing. Within a sliding window, we utilize the trained SVM to get the candidates of CUT boundaries.

3. Filtering. A filtering process based on edge change detection [4] is used to eliminate the influence of flashlight and video-in-video.

3.2 FOI detector

Fadein/fadeout is a typical GT, where one shot fades to monochrome (usually black) frame (with silence), and then fades in to another shot. According to this characteristic, we detect it by three separate steps:

1. Monochrome-block determination

(a) For each frame, the HSV color histogram is extracted-16 bins for each channel and total 48 bins.
(b) The sum of the first 4 bins in each channel is computed. If the sum is less than the threshold T_{FOI}, we consider it a monochrome frame (MF), and a group of consecutive MFs forms a block.
(c) The average sound energy, A_S, of each frame in the block is calcualted. If $A_S < T_S$ threshold, the current frame is labeled as the silence frame.
(d) A frame in a candidate FOI should meet both (b) and (c).

2. FOI boundary detection

For a candidate FOI, we first determine its left boundary by using the accumulated histogram method:
(a) For each candidate block, starting from the previous frame of the block, we consecutively compare the difference of the sum of the first 4 bins of each channel between the current frame and its previous frame the difference is less a threshold T_{diff}.
(b) The right boundary is determined in the similar way.

3. Merging and filtering

The FOI results will be merged into the Final GT results, and will also be used to filter out false ones in the final CUT results.

3.3 GT detector

GT detection is still a challenging work. Though numerous approaches were presented, the results are far from satisfactory. The main reason is that GT is sensitive to motion which cause significant false detections. Hence, in this paper, we use the motion feature, including the camera motion (CM) and the object motion (OM) to eliminate their influence.

The training and testing processes of the GT detector are similar to the CUT detector except different features and feature ties are used. We briefly introduce it as the following:

1. For all features in sliding window L, an alpha-trimming filtering is first used to smooth abnormal features, which is in favor of training.
2. A GT detector with SVM is generated, and then the candidate GT set.
3. The over-segmented segments are merged in candidate GT.
4. Motion filtering is implemented, which is discussed in next subsection.

3.4 Motion detector

As we well known, the inter-frame difference during a gradual transition can be of the same magnitude as that caused by CM and OM, which makes it difficult to distinguish changes caused by a gradual transition from those caused by such motions. Hence, in this paper, in order to resolve this, we analyze the camera motion such as pan, tilt and zoom, and object motion.

1. Camera motion detector

(a) Our previous work [5] is used to extract three camera motions, panning, tilting and zooming, in the video.
(b) The average value, I_{CM}, of each CM segment is computed.
(c) If $I_{CM} > T_{CM}$ threshold, we remove the corresponding GT from the candidate GT set.

2. Object motion detector

(a) In this work, MVs are decoded from MPEG stream, which contains both CM and OM, so MVs of OM should be compensated by CM. i.e., $MV_{OM}=MV-MV_{CM.}$
(b) Median filtering with a w×w spatial window is applied.
(c) The phase histogram of MV_{OM}s are computed and sorted from big to small.
(d) The MV_{OM}s in the biggest three histogram bins are preserved.
(e) All 8-connected MBs with non-zero MV_{OM}s in a frame are merged, and corresponding motion activity I_{OM} of each frame in candidate GT is calculated.
(f) If $I_{OM}>T_{OM}$ threshold, we remove the corresponding GT from the candidate GT set.

4. Evaluation and Conclusion

Our training and testing dataset is from the collections of 2004, 2005 and 2006 SBD task and other sources in different genres of videos such as ad, sports, movie and documentary. An effective SVM tool [6] and C-SVM [7] are adopted to train the shot boundary detectors. After a small scale cross-validation process, we select the best parameter settings: RBF kernel, penalty factor C and kernel parameter γ. By tuning the ratio of C of positive to negative examples, we can control the precision vs. recall of each detector's output. Finally, 40 runs are yielded based on 2007 SBD task, and ten runs are randomly selected and submitted for evaluation. The evaluation result is depicted in Table 1.

Table 1. Evaluation results of the ten submissions

Runs	*ALL*		*CUTS*		*GRADUAL*			
	Recall	Prec.	Recall	Prec.	Recall	Prec.	*Frame*	
							Recall	Prec.
1	0.913	0.900	0.965	0.956	0.345	0.321	0.648	0.832
2	0.914	0.885	0.966	0.938	0.340	0.320	0.653	0.830
3	0.902	0.933	0.962	0.969	0.238	0.355	0.660	0.852
4	0.894	0.856	0.932	0.985	0.481	0.227	0.576	0.815
5	0.892	0.937	0.965	0.944	0.092	0.514	0.894	0.824
6	0.939	0.745	0.964	0.958	0.665	0.164	0.570	0.792
7	0.937	0.758	0.965	0.964	0.641	0.167	0.733	0.732
8	0.948	0.706	0.965	0.959	0.752	0.150	0.675	0.716
9	0.946	0.702	0.965	0.955	0.743	0.147	0.644	0.723
10	0.935	0.788	0.965	0.963	0.612	0.190	0.575	0.811

From the evaluation, we can see that our CUT algorithm can achieve a satisfying result while the GT dose not work as well as our previous testing. The reasons for this are found out after comparing the results with the ground truth.

(1) In GT, lots of false detections happened. That is, because our motion detector didn't work as well as we expected, a more elaborate design of motion detector is needed.

(2) The inconsistent annotation of videos is noticeable. Due to the lack of standard rules, there exists misunderstanding of true GT frames in the datasets.

In summary, much improvable room, especial for GT, for our algorithm exists.

References

[1] D. N.Bhat, S K.Nayar. Ordinal measures for image correspondence, IEEE Trans. on Pattern analy. and machine Intell.,vol.20, no.4, 1998, pp: 415-423.

[2] Jesús Bescós. Real-Time Shot Change Detection over Online MPEG-2 Video. IEEE Trans. Circuits Syst.Video Technol., vol. 14, no.4 pp. 475–484, April 2004.

[3] W. Kraaij, P. Over, T. Ianeva and A. Smeaton. TRECVID 2006 - An Overview. In TRECVID 2006 Workshop, Gaithersburg, MD, US, November, 2006.

[4] R. Lienhart. Reliable transition detection in videos: A survey and practitioner's guide. International Journal of Image and Graphics, vol.1, no.3, 2001, pp: 469-486,

[5] Z.C. Zhao, A.N. Cai. A video retrieval scheme based on global motion for scenery videos, Journal of BUPT, vol.29, 2006, pp: 18-23.

[6] C.C. Chang, C. J. Lin LIBSVM: a library for support vector machine. 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[7] V.Vpanik, The Nature of Statistical Learning Theory. New York: Springer Verlag, 1995.