JDL AT TRECVID 2006 SHOT BOUNDARY DETECTION

Chunxi Liu[1], Huiying Liu[2], Shuqiang Jiang[2], Qingming Huang[1], Yijia Zheng[1], Weigang Zhang[3]

[1]Graduate University of Chinese Academy of Sciences, Beijing 10080, P.R. China
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, P.R. China
[3]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001
{cxliu, hyliu, sqjiang, qmhuang, yjzheng, wgzhang}@jdl.ac.cn

ABSTRACT

In this paper we proposed a simple but effective method for simultaneously detecting video shot transitions of various types by means of combining threshold and SVM based method. Our method first selects the suspicious transition candidates using a low threshold method and then judges the candidates by using the SVM base method. In selecting the low threshold, we use only two simply features: histogram and mutual information. Due to this low threshold, we do not need to extract complex features around every frame changes, so our method can run over real time. The most important problem of the system is its fail to distinct gradual transition and motion in some times. This is because we use little information about the camera motion and object motion. In the future we will use more information about the camera motion to alleviate this problem. The test result on TRECVID benchmark shows our approach is effective for shot boundary detection.

1. INTRODUCTION

Video shot boundary detection is the first step toward the high level video content analysis. In general video shot transition consists of two parts: abrupt transition and gradual transition. Abrupt transition is defined as that there is no intersection frames between two shots. Compared with gradual transition, abrupt shot transition detection is relatively an easy work. However until now even this easy work has not been solved totally. According to the TRECVID evaluation work in 2005 the best performance for abrupt shot transition detection can not pass 95% in precision and recall at the same time. Compared with abrupt shot transition, the gradual shot transition detection is more difficult. According to the editorial fashion, there are hundreds of transition type, including fade in fade out, wipe, dissolve and so on. How to detect the gradual shot transition in a uniform framework is really a challenging work.

Shot transition detection is not easy, because the performance of the shot boundary detection is affected by many factors. The two main negative factors are illumination change and motion. These two factors are the two classical problems in the video processing not only exist in shot detection but also exist in other video processing area such as tracking and so on.

In the past few years, many research works have been paid to shot transition detection. Many methods [1][2][3] use pixel and block based metrics to detection shot transition. While these method perform well on simple videos, they are very sensitive to object and camera motion. A step further towards reducing sensitivity to camera and object movement can be done by using the statistical information of the image—histogram [2][3][4][5]. Histogram is invariant to image rotation and now there exist many variance of histogram based method. Other low level feature such as edge change ratio [6] has also been used. A Comparison of existing shot boundary methods are given in [2] and [3]. In order to achieve good performance in the TRECVID task [9], many participates combine many different features and use SVM based classifier or other classifiers, which are verified effective in other area, to finish the work.

In this paper we summarize our approach for shot boundary detection in the TRECVID shot boundary detection task 2006. We employed only two simple features for cut detection. For gradual transition we add motion vector to compensate for the error caused by motion. The framework for abrupt shot transition and gradual transition is the same which is shown in Figure 1.

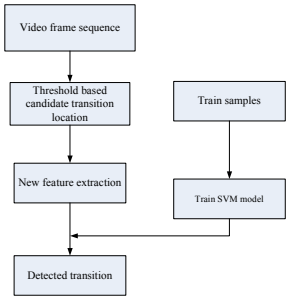


Figure 1: Shot boundary detection framework

The rest of the paper is organized as follows. In section 2, we describe our method for feature extraction and discuss our approach for abrupt and gradual shot boundary detection. Evaluations of the proposed techniques on the

TRECVID 2006 benchmark are shown in section 3. The conclusions and future work is presented in Section 4.

2. SHOT BOUNDRY DETECTION

2.1 Feature extraction

Video is a multi-dimensional media and contain huge information including both visual and auditory information. In our approach, we mainly employ three feature, they are histogram, mutual information between two consecutive frames and motion vector in the frame.

2.1.1 Histogram Dissimilarity Metrics

Histogram comparison is a relatively robust way to detect shot change. Compared with pixels based method, it is less sensitive to object and camera motion, because it ignores the spatial changes in a frame. In this condition, it may also happen the following phenomena that two totally difference images have similar histograms. However in actual video sequence this rarely happen. Let $H_i(k)$ denote the gray-level or color histogram for the *i*th frame, where *k* is one of the *L* possible colors or grey levels. Then the histogram difference *HDM* between the *i*th and its successor is computed as bellow:

$$D_H(i) = (\sum_{k=1}^{l} |H_i(k) - H_{i+1}(k)|_p)^{1/p} \qquad (1)$$

If the overall difference $D_H(i)$ is larger than a give threshold *T*, a shot boundary is detected. For convenience of processing, we first normalize the HDM into the interval [0, 1]. Figure 2 shows a plot of the normalized HDM (with p = 1) for a video clip.

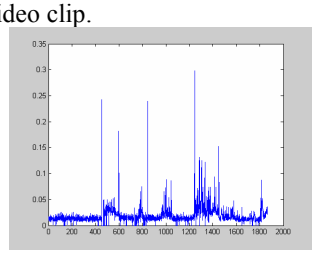


Figure 2: a normalized histogram difference metric with the parameter p =1.

There are 5 cut, 1 fade in the video with high motion. The five big values in Figure 1 are the positions of the cuts. And the last big value over 1800 in the X axis is the position of the fade. Although the histogram difference metric is less sensitive to motion compared with the pixel based method, but from Figure 1 we can see that it cannot totally remove the effect of the motion and in some case it will result in a false detection of a camera break.

2.1.2 Mutual Information

Let X be a discrete random variable with a set of possible outcomes $A_X = \{a_1, a_2, ...a_N\}$ having probabilities $\{p_1, p_2, ...p_N\}$, with $p_X(x = a_i) = p_i, p_i \geq 0$ and $\sum_{x \in A_X} p_X(x) = 1$. Entropy measures the information content or uncertainty of X

$$H(X) = -\sum_{x \in A_X} P_X(x) \log P_X(x) \qquad (2)$$

Two random variables, A and B, with marginal probability distributions, $P_X(x)$ and $P_Y(y)$, and joint probability distribution $P_{XY}(x, y)$. The joint entropy of the X, Y is expressed as bellow

$$H(X,Y) = -\sum_{x,y \in A_X, A_Y} P_{XY}(x,y) \log P_{XY}(x,y) \qquad (3)$$

where $P_{XY}(x, y)$ is the joint probability density function. For two random variables X and Y, the conditional entropy of Y given X is written as $H(Y/X)$ and is defined as

$$H(Y/X) = -\sum_{x,y \in A_X} P_X(x) H(Y \mid X = x) = -\sum_{x,y \in A_X, A_Y} P_{XY}(x,y) \log P_{XY}(x \mid y) \qquad (4)$$

where $P_{XY}(x/y)$ denotes conditional probability. The conditional entropy $H(Y/X)$ is the uncertainly in Y given knowledge of X. It specifies the amount of information that is gained by measuring a variable and already knowing another one. It is very useful if we want to know if there is a functional relationship between two data sets.

The mutual information (MI) between the random variables X and Y measures the degree of dependence of A and B by measuring the distance between the joint distribution $P_{XY}(x, y)$ and the distribution associated to the case of complete independence $P_X(x)P_Y(y)$, by means of the Kullback-Leibler measure

$$I(X,Y) = -\sum_{x,y \in A_X, A_Y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \qquad (5)$$

The relation between the MI and the joint entropy of random variables X and Y is given by

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \qquad (6)$$

where H(X) and H(Y) are the entropies of X and Y. Figure 3 shows the MI measure for the same sports video sequence as the one in Figure 2. The small values in the figure are the positions of the cuts.

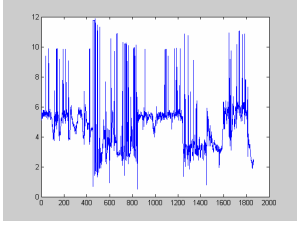


Figure 3: Time series of MI for the same video sequence in Figure 2

2.2 Abrupt Shot Cut Detection

Both histogram and MI can be used for shot boundary detection. A small value of MI or a big value of HDM indicates the existence of a cut between frame f_i and f_{i+1}. But it can be seen from Figure 2 and Figure 3 that when there are great motions in the video both of them could result in false detection. In our approach, HDM / MI is used to as the measure to detect shot cut and is calculated as bellow:

$$HDM / MI_i = \begin{cases} \dfrac{D_H(i) - D_H(i-1)}{MI_i} & \text{if } D_H(i)\text{-}D_H(i\text{-}1)>0 \\ 0 & \text{else} \end{cases} \quad (7)$$

Figure 4 shows the HDM / MI metric for the same sports video sequence as the one in Figure 2 and Figure 3.

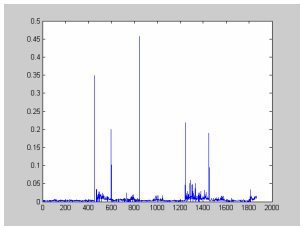


Figure 4: HDM/MI metric for the same video in Figure 1

From the above figures, we can see the new representation could suppress the motion noise greatly and it is an effective feature for abrupt shot boundary detection. And we can use a low threshold to select the candidate abrupt transitions based on this metric.

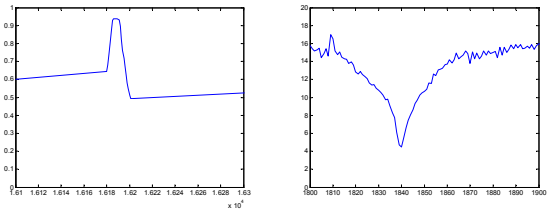
In order to speed up our system we jump 3 frames when processing. After the candidates are selected, we set a window around the candidate transition point to extract some new features. We extend one frames before and after the two candidates frame then we get 6 frames they are $f_1 f_2 f_3 f_4 f_5 f_6$, where f_1 and f_6 are the extend frames. Then we use the f_1 as a prototype and calculate the histogram difference and mutual information between the prototype frame and other frames. Also the adjacent histogram difference and mutual information between the consecutive frames are contained as our feature. We get another five features by realigning the histogram distance between the adjacent frames from top to down. Totally we get 25 features. These 25 features are used as our features for SVM classifier training and predicting.

2.3 Gradual Shot Transition Detection

For gradual transition detection, our approach includes two parts: fade out/in detection and other type of gradual transition.

2.3.1 Fade out/in detection

Fade out/in is a special kind of shot gradual transition, which is different from other types of gradual transition and has its own special feature. In a fade out the last frame of the shot becomes darker and darker and then disappears. Then the frame in fade in becomes brighter and brighter. A fade out/in may include several frames. For fade out/in detection we use two features: image monochrome and joint Entropy between two frames.



a b

Figure 5: Monochrome degree changes in a Fade out/in

The monochrome degree indicates the centrality of color distribution of a frame: if most of pixels in a frame have similar colors, the frame is called a monochrome one. In the fade in/out transition process the monochromes of the frame decrease first, and then increase. The change of the monochrome in a fade out/in is shown in Figure 5 *a*. The pixel values of a monochrome frame concentrate in a narrow scope. In Fade out/in the monochrome frames are black ones. In our approach the monochrome degree is calculated as bellow:

$$MD = \sum_{i=1}^{3} \sum_{j=1}^{\frac{1}{4}bin} h_{ij} / (3 * \sum_{j=1}^{bin} h_{ij}) \quad (8)$$

bin is the dimension of the histogram, i means the *i*th color channel, h_{ij} means the *j*th value of the *i*th histogram. If MD is larger than the predefined threshold, the frame is a monochrome one and it indicate a fade out/in happens.

The Joint Entropy between two images means the information between them. The larger the difference is, the more the information is, and the larger the Joint Entropy is. In Fade out/in, the Joint Entropy has a trough type, as shown in figure 5 *b*. If the joint entropy is lower than the predefined threshold, then a fade out/in happens.

2.3.2 Other type gradual transition detection

For other gradual transition detection, first we define a slide window with width 60 frames and then use two thresholds method [3] to decide if there exists gradual transition. If there is a gradual transition, then we throw the whole window into the classifier. Totally we extract 92 dimension features including histogram difference between adjacent frames, mutual information between adjacent frames, motion vectors, joint entropy and so on. We use the 2005 TRECVID data as our training set to train our model and use the trained model to predict our 2006 result.

Table.1.The detail result for all runs

Run id	All Transitions		Cuts		Graduals		Gradual Frame Accuracy	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
1	0.721(0.868)	0.626(0.753)	0.716(0.917)	0.619(0.793)	0.736	0.646	0.812	0.745
2	0.718(0.870)	0.635(0.769)	0.709(0.913)	0.640(0.824)	0.744	0.622	0.798	0.766
3	0.721(0.867)	0.623(0.748)	0.720(0.922)	0.610(0.782)	0.726	0.656	0.809	0.744
4	0.730(0.872)	0.628(0.751)	0.724(0.922)	0.613(0.781)	0.746	0.670	0.811	0.749
5	0.711(0.856)	0.654(0.788)	0.703(0.907)	0.636(0.819)	0.729	0.703	0.800	0.779
6	0.728(0.880)	0.624(0.754)	0.716(0.917)	0.635(0.813)	0.766	0.594	0.807	0.750
7	0.726(0.873)	0.624(0.751)	0.724(0.926)	0.621(0.795)	0.731	0.631	0.809	0.749
8	0.715(0.856)	0.627(0.751)	0.719(0.919)	0.613(0.782)	0.705	0.667	0.812	0.753
9	0.712(0.837)	0.638(0.781)	0.708(0.907)	0.620(0.799)	0.721	0.687	0.810	0.755

4. EXPERIMENT RESULT

We evaluated our system against the TRECVID 2006 shot boundary test collection. The TRECVID 2006 collection contains 13 broadcast videos from CNN, CCTV, NBC, MSNBC, PHOENIX, LBC, HURRA and NTDTV. The TRECVID 2005 shot boundary test data is used as our training set, totally we trained 9 shot boundary detection models.

The 9 runs of our shot boundary detection result are shown in table 1. Our method achieved a recall of 0.623 and precision of 0.715 for hard cut and achieved a recall of 0.653 and precision of 0.734 for gradual transition in the evaluation. The results of the hard cut are not satisfactory the reason is that TRECVID 2006 data contains 13 videos and in 7 videos the decoder we use has one frame difference with the one used by TRECVID evaluation. This difference has no effect on hard transition with not frame intersections because in the hard cut evaluation for abrupt cut the frame is extended forward and backward by 5 frames. But for short transition the length of which is shorter than 5 frames this is a disaster. Only one example can show this clearly. For example the label in evaluation is pre="523" post="525", the label of our system is pre="524" post="526" and in the evaluation this situation is labeled as insertion. This minor different between two decoders depresses our cut result greatly. If only this situation is corrected for hard cut our result is relatively good which is shown in table 1 in the bracket (the evaluation for gradual transition is not changed). For gradual transition our approach encounter errors when the object or the camera in the video moves fast. Sudden illumination changes also take up a lot of errors.

5. CONCLUSION

In this paper, we summarized the approach adopt in our TRECVID 2006 shot boundary detection task. We explored the histogram and information based feature and SVM based classifier for constructing a unified framework for detecting and identifying different types of shot transitions. Our proposed method generally performed well. It showed some weaknesses under very fast object and camera motion and sudden illumination changes. The biggest contribution of this paper is use simple histogram and information feature in a unified framework to detect both cut and gradual transition with a reasonably high performance. Further, as no complex, specialized video or image processing operation is employed, the method is also highly computationally efficient. Finally, in some condition our method is also sensitive to object and camera motion and illumination change. In the future we may use more motion analysis to compensate this weakness.

6. REFERENCES

[1] T. Kikukawa, S. Kawafuchi, Development of an automatic summary editing system for the audio-visual resources, Transactions on Electronics and Information J75-A (1992) 204-212.

[2] A. Nagasaka, Y. Tanaka, Automatic video indexing and full-video search for object appearances, in Visual Database Systems II (E. Knuth and L.M. Wegner, eds.), pp. 113-127, Elsevier, 1995.

[3] H.J. Zhang, A. Kankanhalli, S.W. Smoliar, Automatic partitioning of full-motion video, Multimedia Systems 1(1) (1993) 10-28.

[4] M. J. Swain, Interactive indexing into image databases, in: Proc. SPIE Conf. Storage and Retrieval in Image and Video Databases, 1993, pp.173-187.

[5] Y. Tonomura, Video handling based on structured information for hypermedia systems, in: Proc. ACM Int. Conf. Multimedia Information Systems, 1991, pp. 333-344.

[6] Ramin Zabih, Justin Miller, and Kevin Mai. Feature-based algorithms for detecting and classifying scene breaks. *Third ACM Conference on Multimedia*, pages 189 – 200, November 1995.

[7] K. Matsumoto et al. Shot Boundary Detection and Low-Level Feature Extraction Experiments for TRECVID 2005, TREC Video Retrieval Evaluation Online Proceedings, TRECVID, 2005.

[8] J. Yuan et al. Tsinghua University at TRECVID 2005, TREC Video Retrieval Evaluation Online Proceedings, TRECVID, 2005.

[9] http://www-nlpir.nist.gov/projects/trecvid/