SVM-BASED SHOT BOUNDARY DETECTION WITH A NOVEL FEATURE

Kazunori MATSUMOTO *Masaki NAITO* *Keiichiro HOASHI* *Fumiaki SUGAYA*

KDDI R&D Laboratories, Inc.
2-1-15 Ohara, Fujimino-shi, Saitama 356-8502, Japan
{matsu, hoashi, naito, fsugaya}@kddilabs.jp

ABSTRACT

This paper describes our new algorithm for shot boundary detection and its evaluation. We adopt a 2-stage data fusion approach with SVM technique to decide whether a boundary exists or not within a given video sequence. This approach is useful to avoid huge feature space problems, even when we adopt many promising features extracted from a video sequence. We also introduce a novel feature to improve detection. The feature consists of two kinds of values extracted from a local frame sequence. One is the image difference between the target frame and that synthesized from the neighbors. The other is the difference between neighbors. This feature can be extracted quickly with a least-square technique. Evaluation of our algorithm is conducted with the TRECVID evaluation framework. Our system obtained a high performance at a shot boundary detection task in TRECVID2005.

1. INTRODUCTION

The accurate segmentation of shots in a video sequence is fundamental and an essential functionality for numerous video retrieval and management tasks. Many researchers have proposed algorithms to perform shot boundary detection based on certain features extracted from video frames, such as pixel differences, edge differences, color histograms, etc. Moreover, comparative surveys have also been carried out [1, 2].

From a learning theory perspective, it is a natural approach to combine such promising features in order to decide whether a boundary exists or not within a given video sequence. But naïve feature combination makes an excessive feature space to handle. In this paper, we adopt a 2-stage data fusion approach with a Support Vector Machine (SVM) technique. The overview of our data fusion approach is as follows: At the first stage, every adopted feature is judged by a specific SVM. This means the number of feature types is equal to the number of SVMs at the 1st stage. And the other SVM at the second stage synthesizes the judgments from the 1st stage. The details of this data fusion approach are described in section 2.

Another major topic of this paper is a novel feature for shot boundary detection. This new feature is introduced to improve segmentation accuracy. The basic shell of our idea is as follows:

Let there be three continuous frame images, of which the middle one is regarded as a target image. Assume a synthesized image from the two non-target images. In our approach we minimize the difference between the target and the synthesized image by adjusting synthesis parameters. If an abrupt shot boundary exists within the continuous three images, the difference will become significant, even after optimal minimization is conducted. Practically, this optimization process is not a burden by using a least-squares technique.

In addition, we consider the difference between non-target images simultaneously. Dealing with these two different types at the same time results in a better performance. There is little difference between the target image and a synthesized image, both in the case of a non-cut sequence and one including a dissolve cut. However, it must be noted that the difference between non-target images is considerable in the case of dissolve although minor in the case of a non-cut sequence. The matrix in Table 1 summarizes the qualitative trend of these two differential values.

We extended this basic idea to apply the frame sequence so that the number of frames was more than three. Details of the proposed feature are described in section 3.

Type of difference \ Type of cut in a sequence	**None**	**Abrupt cut**	**Dissolve cut**
Image difference between the target and synthesized image	small	significant	small
Image difference between non-target images	small	significant	significant

Table 1: The matrix showing which type of cut affects the two types of difference.

The evaluation of a shot boundary detection system based on our proposal is conducted with the TRECVID evaluation framework [3]. In this evaluation, the performance of our new feature is compared with that of other promising features by *recall* and *precision* measures. Subsequently, the overall performance of our data fusion system is compared with that of other systems of TRECVID2005's shot boundary detection task.

2. DATA FUSION WITH MULTIPLE SVMs

The methods of shot boundary detection can be classified into two major types. One is the method to extract features from a compressed domain, while the other involves extracting features from an uncompressed domain. As our concern is to achieve accurate detection, we adopt the data fusion approach to combine the promising features extracted in an uncompressed domain.

Figures 1 and 2 show our 2-stage discriminators with SVMs. Figure 1 is the structure of the discriminator in training mode, and Figure 2 is in prediction mode. "F1" ~ "F6" represent the feature values extracted from a video sequence. A conventional and useful *multiple pair-wise* technique [4] is applied for all these features. Table 2 shows the brief description of these features. "CUT Label," "DSH Label," and "D01L," ~ "D05L" are the label data for training. The values of every label data are assigned frame by frame. The "CUT Label" discriminates whether an abrupt cut occurs just before a relevant frame. "DSH Label" discriminates as to whether the center of a dissolve transition exists at a relevant frame. "D01L" ~ "D05L" discriminates whether the center of a dissolve transition with a specific period exists at a relevant frame. It is not easy for hand-labelers to specify such a dissolve transition. But through the effort of TRECVID annotator, we can obtain accurate training data of dissolve transitions. Figures 3 and 4 show examples of abrupt and dissolve cuts respectively. In figure 4, the span of the dissolve transition is three frames.

Feature ID	Description	# dimension
F1	the number of in-edges and out-edges in divided regions (4 by 4) based on [2]	224
F2	Standard deviation of pixel intensities in divided regions (4 by 4)	224
F3	TRECVID2005 approach by FX PAL[5] with Ohata's color domain, with PAC	192
F4	TRECVID2005 approach by FX PAL[5] with RGB color domain, with PAC	192
F5	Edge change ratio described in [7]	192
F6	Novel feature described in section 4.	210

Table 2: Explanation about adopted features.

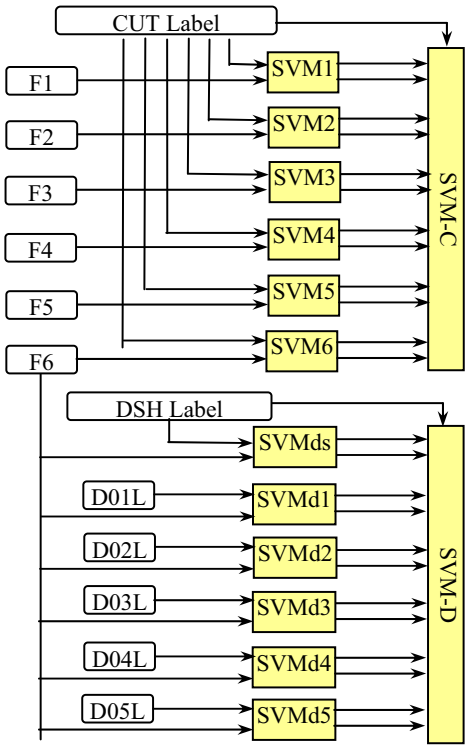


Figure 1: Structure of 2-stages SVMs in training mode.

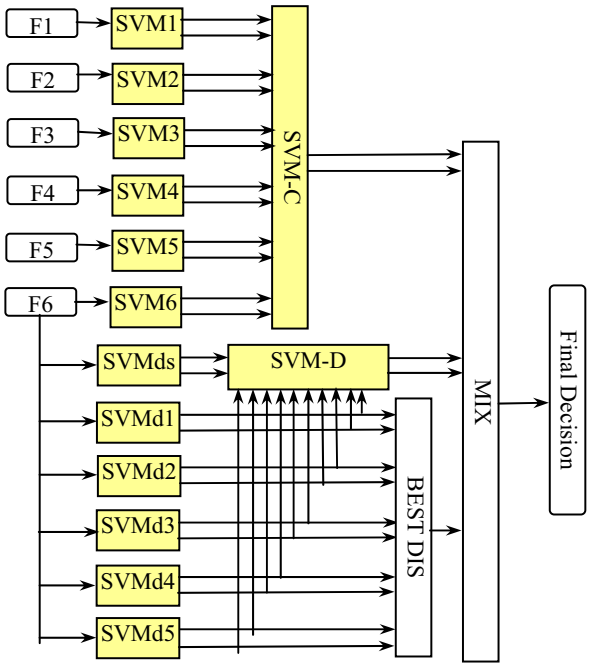
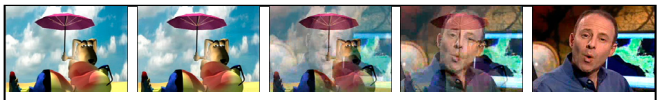


Figure 2: Structure of 2-stages SVMs in a prediction mode.

CUT=0 D01=0 D03=0 DSH=0	CUT=0 D01=0 D03=0 DSH=0	**CUT=1** D01=0 D03=0 DSH=0	CUT=0 D01=0 D03=0 DSH=0	CUT=0 D01=0 D03=0 DSH=0

Figure 3: Example of an abrupt cut and values of labels.



CUT=0 D01=0 D03=0 DSH=0	CUT=0 D01=0 D03=0 DSH=0	CUT=0 D01=0 **D03=1** **DSH=1**	CUT=0 D01=0 D03=0 DSH=0	CUT=0 D01=0 D03=0 DSH=0

Figure 4: Example of a dissolve cut (transition span = 3)

"SVM1" ~ "SVM6" are Support Vector Machines at the 1st stage. Each SVM is designed to detect an abrupt cut based on a specific feature. "SVMds" is designed to detect a dissolve cut with any transition span. "SVMd1" ~ "SVMd5" are designed to detect a dissolve transition with a specific length. For example, "SVM1" discriminates the existence of a dissolve transition whose length is 1. Every SVM outputs two kinds of values: the probability that a specified type of cut is detected and the probability that the same is not detected. "SVM-C" and "SVM-D" are Support Vector Machines at the 2nd stage. "SVM-C" discriminates the existence of an abrupt cut based on the result of the 1st stage, while "SVM-D" also discriminates a dissolve cut.

The functionality of "MIX" on Figure 2 is an arbitration of "SVM-C" and "SVM-D", based on the four probabilistic values. When "SVM-D" detects a dissolve cut and "SVM-C" does not detect an abrupt cut, "BEST DIS" chooses the most probable length of the dissolve transition.

Please note that SVM training may be a resource consuming task, but the computation cost of SVM's prediction is less than that of feature extraction.

3. NOVEL FEATURE

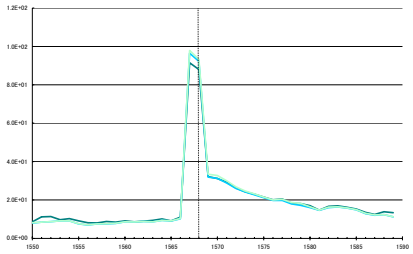
A frame image in a dissolve transition is synthesized from two images, which come from different two video sequences respectively [5]. There are two scaling parameters for the synthesis. We estimate these two optimal scaling parameters frame by frame. The estimation process (least-squares) is as follows:

Let f be a image synthesized from images f_A, f_B.

Let A_R, A_G, A_B be scaling parameters of f_A.

Let B_R, B_G, B_B be scaling parameters of f_B.

(a) The case where the frame distance is 1.



(b) The case where the frame distance is 3.

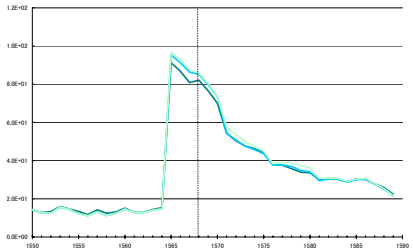
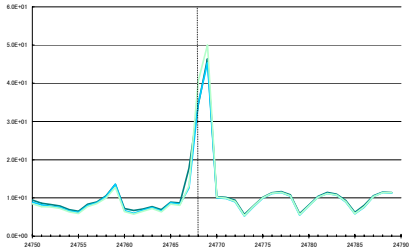


Figure 5: Time series when an abrupt cut occurs.

(a) The case where the frame distance is 1.



(b) The case where the frame distance is 3.

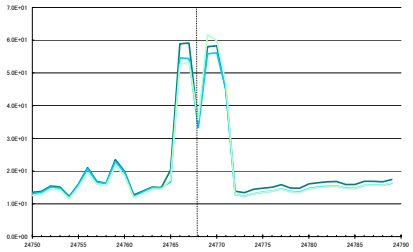


Figure 6: Time series when a dissolve cut occurs.

Let X_i^R, Y_i^R, Z_i^R, X_i^G, Y_i^G, Z_i^G, X_i^B, Y_i^B, Z_i^B be the luminance of pixel i in f_A, f_B,, f. Then:

$$A_R X_1^R + B_R Y_1^R - Z_1^R = \varepsilon_1^R$$
$$A_R X_2^R + B_R Y_2^R - Z_2^R = \varepsilon_2^R$$

$$A_R X_n^R + B_R Y_n^R - Z_n^R = \varepsilon_n^R$$

.

Let $F_R(A_R, B_R) = \Sigma(\varepsilon_i^R)^2$, the estimation problem is to find A_R and B_R which minimize $F_R(A_R, B_R)$.

When minimization is achieved,

$$\frac{\partial}{\partial A} F_R(A_R, B_R) = 0$$

.

Thus the following equation should be solved:

$$\begin{bmatrix} \Sigma(X_i)^2 & \Sigma X_i Y_i \\ \Sigma X_i Y_i & \Sigma(Y_i)^2 \end{bmatrix} \begin{bmatrix} A_n \\ B_R \end{bmatrix} = \begin{bmatrix} \Sigma Y_i Z_i \\ \Sigma X_i Z_i \end{bmatrix}$$

Once optimal synthesis parameters A_R, B_R, A_G, ⋯ are obtained, we can easily calculate the feature values by the definitions.

Figure 5 (a) and (b) show the time series of the above three feature values, where an abrupt cut occurs in the sequence. Figure 5 (a) is the case where the frame distance between f and f_A is 1, which means f and f_A are adjacent. Figure (b) is the case where the distance is 3. Figure 6 (a) and (b) show the time series where a dissolve cut, whose transition span is 1, occurs. Note that the characteristics of cut appear in these figures.

4. EVALUATION

Evaluation is conducted with the TRECVID evaluation framework. The ground truth data of TRCVID2004's shot boundary task (about 6 hours of news video) is used as training data for abrupt cuts discrimination. As the number of short span dissolve cuts is insufficient, we make ground truth data from a subset of TRECVID2005's development data and use it for training of the dissolve cut (a news video of about 6 hours). For the test data, 12 videos of TRECVID2005' shot boundary task.

Table 3 shows each performance of SVM for abrupt cuts. This result shows the advantage of the proposed novel feature and the availability of data fusion.

Figure 7 shows the submitted result of abrupt cut, including a short dissolve transition within a 5 frame span at the shot boundary task in TRECVID2005. The performance of our system is based on the technique described here and the system obtained the highest score. Our investigation of this result notes that our novel feature helps abrupt cuts be found more accurately by not being wrongly labeled as dissolved.

5. CONCLUSION

This paper describes our new algorithm for shot boundary detection and its evaluation. The data fusion approach is useful to manage a huge feature space, even if all promising features are combined together.

	SVM1	SVM 2	SVM 3	SVM 4	SVM 5	SVM6 propose	SVM-C
Rec.	0.861	0.923	0.913	0.920	0.724	**0.920**	**0.945**
Prec.	0.811	0.871	0.922	0.887	0.918	**0.937**	**0.916**

Table 3: Recall and Precision of SVM for abrupt cuts.

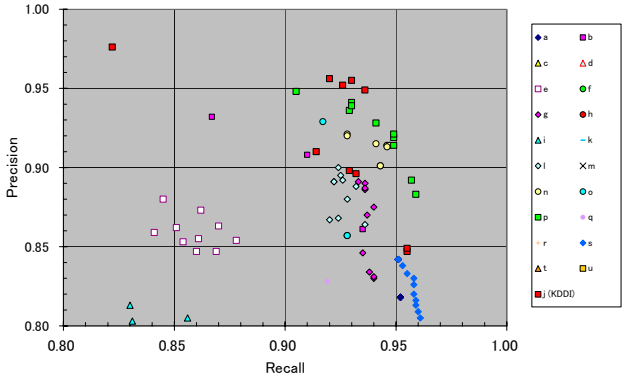


Figure 7: Submitted result of the abrupt cut detection in TRECVID2005 (Red square markers represent our system, which is located in the top right corners).

REFERENCES

[1] Ullas Gargi, Rangachar Kasturi, Susan H. Strayer. "Performance Characterization of Video-Shot-Change Detection Methods," *IEEE Transaction on Circuits and Systems for Video Technology*, Vol. 10, No. 1, February 2000.

[2] Rainer Lienhart. "Comparison of Automatic Shot Boundary Detection Algorithms," *Storage and Retrieval for Still Image and Video Databases* VII 1999, Proc. SPIE 3656-29, Jan. 1999.

[3] NIST, "*Digital Video Retrieval at NIST: TREC Video Retrieval Evaluation*," 2001-2004, http://wwwnlpir.nist.gov/projects/trecvid/.

[4] Amir, A., Berg, M., Chang, S.-F., Hsu, W., Iyengar, G., Lin, C.-Y., Naphade, M., Natsev, A. P., Neti, C., Nock, H., Smith, J. R., Tseng, B., Wu, Y., and Zhang, D. "IBM research TREC-2003 video retrieval system," *TREC Video Retrieval Evaluation (TRECVID 2003)*, Gaithersburg, MD, NIST, 2003

[5] J. Adcock, A. Girgensohn, M. Cooper, T. Liu, L. Wilcox, E. Rieffel, "FXPAL Experiments for TRECVID 2004," *TREC Video Retrieval Evaluation (TRECVID 2004)*, Gaithersburg, MD, NIST, 2004

[6] R. Zabih, J. Miller, and K. Mai. A Feature-Based Algorithm for Detecting and Classifying Scene Breaks. *Proc. ACM Multimedia 95*, San Francisco, CA, pp. 189-200, Nov. 1995.

[7] Lienhart, R. "Reliable Transition Detection In Videos: A Survey and Practitioners Guide," International Journal of Image and Graphics (IJIG), 1(3):469–486, 2001