3—17 Bayesian Shot Detection using Structural Weighting

Seung Hoon Han & In So Kweon*
Dept. of Electrical Eng. & Computer Science
KAIST

Chang-Yeong Kim & Yang-Seck Seo^
Signal Processing Laboratory
SAIT

Abstract

A video stream consists of a number of shots each of which has different boundary types such as cut, fade, and dissolve. Many previous approaches can find the cut boundary without difficulty. However, most of them often produce false alarms for the videos with large motions of camera and objects. In this paper, we demonstrate that the shape of the histogram difference between two successive color images, called the structural information, provides an important cue to distinguish fade and dissolve effects from cut effect. Our shot detection method uses an optimal Bayesian classifier weighted by the structural information to model the gradual transitions such as fades and dissolves. The proposed method has been tested for a few golf video segments and shown good performances in detecting fade and dissolve effects as well as cut.

1. Introduction

The cut boundaries show an abrupt change in image intensity or color, while those of fades or dissolves show gradual transition between frames. There is also a slow change in intensity for sequences with image motion. Many methods have been proposed to distinguish the gradual transition boundary from the intensity change due to image motion. Typical ones include Twin Comparison [1], automatic threshold method [2], model-based method [3, 4], and Bayesian approaches [5, 6].

The most closely related work with the proposed one is the Bayesian approach that has addressed optimal threshold problems using *a priori* knowledge for shot duration and shot activity. While Bayesian classifiers are statistically optimal in determining the cut boundary, they have limitations when the probability density function cannot model the shot activity properly. For example, regular frames, cut and gradual transitions cannot be separated properly because of ambiguity between large image motion and gradual transition effects.

Specifically, Hanjalic & Zhang [6] use visual discontinuity patterns to detect gradual transition effects and fades/dissolves. However, it is difficult to find such visual discontinuity patterns in image sequences with camera/object motion, and to determine if they are true patterns for gradual transition effects (fades/dissolves).

The proposed method, called Structural Bayesian Shot Detection (SBSD), combines a Bayesian formulation for each transition type with the structural model, which is based on the transition pattern shapes of filtered histogram difference. It is shown that the cut and the gradual effect can be modeled as a rectangular and a triangular shape, respectively.

The paper is organized as follows: Section 2 introduces the Bayesian model for the transition types. Section 3 describes the structural modeling. Finally, Section 4 presents experimental results and concluding remarks.

2. Optimal Bayesian Formulation

In [5, 6], two hypotheses, shot boundary or not, are considered. In this paper, the following three hypotheses are defined to obtain shot boundaries using transition pattern shapes:

h_0 : no shot boundary (regular frame)
h_1 : cut boundary
h_2 : gradual boundary

Three decisions are subsequently defined according to posterior probabilities and transition pattern shapes:

d_0 : decide no shot boundary (regular frame)
d_1 : decide cut boundary
d_2 : decide gradual boundary

We also define two feature vectors, frame difference and transition pattern, based on the color histogram difference (HD) between two successive images. Now, we can define an optimal Bayesian formulation as

*Address : 373-1 Kusong-dong, Yusong-ku, Taejon, 305-701, Korea, E-mail: {shhan, iskweon}@kaist.ac.kr

$$P(d_i / X,Y) = \sum_{j=0}^{2} w(d_i \mid h_j, Y) P(h_j \mid X)$$

$$= \sum_{j=0}^{2} w(d_i \mid h_j, Y) p(X \mid h_j) P(h_j) \quad (1)$$

$$Decision = \underset{di=\{d0,d1,d2\}}{\arg\max} \; (P(d_i / X,Y))$$

where X and Y denotes the two feature vectors: frame difference and transition pattern. Posterior probabilities are multiplied by a weighting factor w. The weighting factor w, which helps detect shot boundaries correctly, is defined by

$$w_{ij} = w(d_i \mid h_j, Y) = \begin{cases} S(Y) = w_{ii} & \text{if i = j} \\ w_{ii} \prod_{i \neq j} (1 - w_j) & \text{if i} \neq \text{j} \end{cases} \quad (2)$$

Here, $S(Y)$ is one of the similarity functions between an input pattern Y and the ideal transition pattern of effects. If a hypothesis is equal to a corresponding decision, w ($=w_{ii}$) means a correct intersection between input pattern shape and ideal transition pattern shapes of effects. If not, w_{ij} just contributes to deciding d_i. For example, weighting factor works when two posterior probabilities for regular frame and gradual boundary (fade/dissolve) have similar values. A larger weighting value leads to a corresponding decision. A probability density function for each hypothesis is given by the maximum likelihood estimate (MLE) [5]. A Gaussian function (equation (3)) is fitted to the distributions for cut and gradual transition, and an Erlang function (equation (4)) for gradual transition effects (fades/dissolves) and shot duration.

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2 / 2\sigma^2} \quad (3)$$

$$\varepsilon(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} \quad (4)$$

3. Structural Modeling

In this section, we describe how to extract the transition pattern shapes of shot boundaries. There are difficulties in detecting exact shot boundaries in cases of camera motion, object motion, flash light, and so on. Motion effects among them cause most false alarms and degrade the precision rate. To solve this problem, robust motion estimation and segmentation algorithms are required, but which are computationally inefficient for real-time applications. On the other hand, the shot activity experiencing camera/object motion is more complex than gradual transition effects. Therefore, the degree of freedom (DOF) of gradual transition effects (fades/dissolves) is lower than that of motion effects. It is important to consider transition activity in gradual transition effects to separate them from gradually changed frames due to motion in images.

3.1 Transition pattern

The color histogram differences between adjacent frames are defined as

$$fd[i] = \sum_{k \in (R,G,B)} \sum_{j=1}^{G} \left| H^k{}_i(j) - H^k{}_{i-1}(j) \right| \quad (5)$$

where G represents the number of histogram bins.

In an ideal case, the cut boundary showing an abrupt change in the HD is represented by a Kronecker delta function

$$fd_c[i] = \alpha_i \delta(i - i_c) \quad \text{for cut boundary} \quad (6)$$

Assuming that color or intensity has been changed only due to video edit effects and the brightness variation in an image is uniform, the gradual scene transition such as fade or dissolve has the form of

$$fd_e[i] = \beta_i rect\left(\frac{i - i_e}{T_e}\right) \quad \text{for fade/dissolves}$$

$$rect(x) = \begin{cases} 1, & |x| \leq 1/2 \\ 0, & |x| > 1/2 \end{cases} \quad (7)$$

Clipping $fd[i]$ to the average of frame differences reduces noises and motion in an image. This clipped signal $fd_{clip}[i]$ is input to the next convolution process. By convolving the $fd_{clip}[i]$ by a window whose width is M and the magnitude $1/M$, we obtain

$$fd_{conv}[i] = fd_{clip}[i] * \left(rect(i / M) \Big/ M \right) \quad (8)$$

After this convolution process, the cut boundary has the form of equation (9) and the fade or dissolve boundary has the form of equation (10).

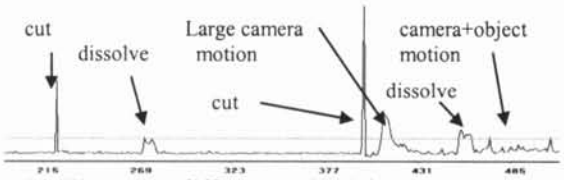
$$fd^{cut}_{conv}[i] = \frac{\alpha_i}{M} rect\left(\frac{i - i_c}{M/2}\right) \quad (9)$$

$$fd^{edit}_{conv}[i] = \frac{\beta_i}{M} tri\left(\frac{i - i_c}{M/2}\right)$$

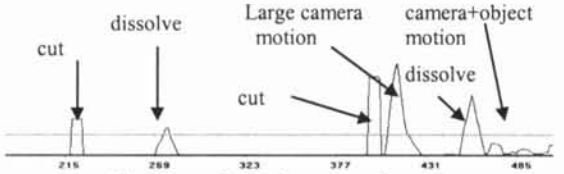
$$tri(x) = \begin{cases} 1 - |x|, & |x| \leq 1 \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

A golf video segment is used to test our transition pattern model. First, we compute the color histogram difference using equation (5) and

an example of the color histogram difference is shown in Figure 1(a).



(a) Histogram differences (HD) between frames



(b) Filtered HD between frames

Figure 1. An experimental result to test the transition pattern model: (a) A color histogram difference; (b) the filtered difference and transition pattern shapes of effects.

In Figure 1(a), we can observe that the sharp peaks corresponding to the cut boundaries and their values are relatively large compared with other boundary values. In dissolves or fades, the differences and the rate of change are small. However, the difference values don't make flat plateau. We can also observe that the difference magnitude due to large camera motion is very similar to that of the cut boundary. These differences due to the large camera motion lead to false alarms in most shot detection systems without motion analysis.

Now, the differences in frames with small camera motion are smaller than those of frames with large camera motion, but are comparable to those of gradually changed frames, such as dissolve frames. These also result in false alarms.

Figure 1(b) shows the filtered frame differences that are generated by an average-clipping operation and a subsequent local window convolution to Figure 1(a). From this filtered frame differences, we can observe that there is a clear distinction between the cut and the dissolve boundary: the rectangular shape for cut-type boundaries and the triangular shapes for dissolve-type boundaries.

Based on these observations, we define the transition patterns of edit-effects as follows: a rectangular shape for the cut and a triangular shape for the gradual scene transition such as fade or dissolve.

3.2 Weighting factor w_{ij}

In this section, we define the weighting factor w_{ij} that is based on the similarity between a pattern feature Y and the ideal transition effect patterns defined in the previous section.

From given filtered histogram differences as shown in Figure 1(b), detecting rectangular and triangular shapes is straightforward. The existence of plateau is checked for the rectangular shape (cut), two triangle sides for the triangular shape (fade or dissolve). Subsequently, their similarities to the ideal transition pattern are calculated by

$$
\begin{aligned}
S_c &= \sum_{m=-FW}^{m=FW} \left(Y(k+m) - avg_{W_c}\{Y(k)\} \right) \quad \text{if } Y(k) \neq 0 \\
S_g &= \sum_{m=-W_l}^{m=0} \left(Y(k+m) - Leftside(k+m) \right) + \\
&\quad \sum_{m=0}^{m=W_r} \left(Y(k+m) - Rightside(k+m) \right) + \left((W_l - W_r)/(W_l + W_r) \right) \\
&\quad \text{if } Y(k) \neq 0
\end{aligned}
\tag{11}
$$

where the third term of S_g enforces the symmetric properties of an isosceles triangle.

The final weighting factor w is given by

$$
w(S) = \frac{1}{1 + 0.5(S/\sigma)^2} \tag{12}
$$

where S means the similarity value defined by equation (11) and σ denotes a scaling factor.

4. Experiments and Results

The performance of the proposed SBSD algorithm is compared with that of the Bayesian Shot Detection (BSD) using some video segments taken from the PGA Championship 2000 and the Showdown at Sherwood.

The golf video segments have various camera effects such as panning, zooming, tilting, and lots of gradual transition effects fades/dissolves. The test data also contain large image motion due to fast golf swing actions as well as camera motions.

Video segments of six different golf games consisting of 12,000 images are used to model probability density functions of the transition patterns. A probability density function for each transition type is fitted experimentally using ML estimate. Frame difference for regular frames has an Erlang function (equation (2)) of which n = 7 and λ = 1.6. The distribution function for cut and gradual transition effect is a Guassian function with average = 50, variance = 150 and average = 15, variance = 16, respectively. A PDF for shot duration has the form of Erlang function with n = 3 and λ = 0.09. Figure 2 shows such probability density distributions.

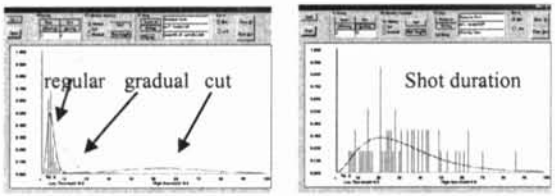


Figure 2. Probability density distributions for each transition type (left) and shot duration (right).

The video segment of the PGA Championship 2000 has provided very challenging shot transitions due to large camera motions for tracking golf balls. The BSD algorithm, without exact motion analysis and object tracking, has produced many false alarms in such a case.

The experimental results are summarized in Table 1. Here, two performance indices such as the recall and precision are defined as

$$\text{Recall} = \frac{Correct}{Correct + Missed} \qquad \text{Precision} = \frac{Correct}{Correct + False\ Alarms}$$

The proposed SBSD method has achieved higher recall and precision rates than the BSD method for both video segments.

Table 1. Comparison between SBSD and BSD

Method	SBSD		BSD	
Sequence (frames)	PGA2K (3609)	Woods/Duval (1609)	PGA2K (3609)	Woods/Duval (1609)
Correct	31	26	28	24
Missed	6	1	9	4
F.A.	5	2	10	3
Recall	0.84	0.96	0.76	0.86
Precision	0.86	0.93	0.74	0.89

Figure 3 shows two pattern similarity probabilities for gradual effects (up) and cut (down) for the testing video segment. There are four shot boundaries (two dissolves and two cuts), which are evident from the triangular and rectangular shapes of the filtered histogram difference. Note that much greater probabilities are present in respective gradual and cut boundaries. Figure 4 shows the detected shots by the SBSD (1st row) and the BSD (2nd row). In case of the BSD (2nd row), the last shot is falsely detected as two shots due to large camera and object motions.

5. Conclusions

In this paper we proposed a robust shot detection method combining Bayesian formulation and structural information. We can obtain structural information by low-pass filtering the color histogram difference between two successive images. A rectangular shape is obtained for the cut effect and a triangular shape for the gradual effect. We have demonstrated the feasibility of the SBSD method to detect the fades and dissolves effects as well as the cut effect for some video segments of real golf games.

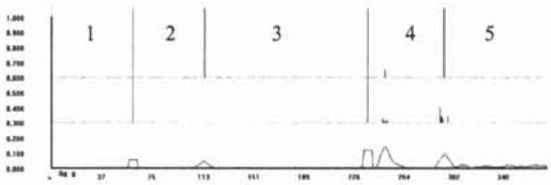


Figure 3. Two pattern similarity probabilities: gradual effects (up) and cut (down).

Figure 4. Detected shots by SBSD (1st row) and BSD (2nd row).

References

1) H. Zhang, A. Kankanhalli, S. Smoliar, and S. Tan, "Automatic partitioning of full motion video", ACM Multimedia systems 1(1), 1993,10-28

2) B. Gunsel and A. Tekalp, "Content-based video abstraction", Proc. IEEE Int. Conf. Image Proc., Chicago, IL, Oct. 1998.

3) A. Hampapur, R. Jain, T.E. Weymouth, "Production model based digital video segmentation", Multimedia Tools and Applications, vol.1, pp9-46, 1995.

4) T. Song, W. Kwon, M.Kim, "On Detection of Gradual Scene changes for parsing of Video data", pp404-423, SPIE: Storage and Retrieval for Image and Video Databases, 1998

5) N. Vasconcelos, A. Lippman, "Statistical Models of Video Structure for Content Analysis and Characterization", IEEE Trans. On Image Processing, Vol.9, No.1, January 2000

6) A. Hanjalic, H. Zhang, "Optimal Shot Boundary Detection based on Robust Statistical Models", IEEE International Conference on Multimedia Computing and Systems (ICMCS'99), Florence, 1999.