A METHOD FOR ESTIMATING CUT-EDIT POINTS IN PERSONAL VIDEOS

Takuya Furukawa †, Hironobu Fujiyoshi† and Akiyuki Nomura‡

† Department of Computer Science, Chubu University
Email: takuya@vision.cs.chubu.ac.jp, hf@cs.chubu.ac.jp
‡sus4 Co., Ltd
Email: aki-nomura@sus4.co.jp

ABSTRACT

We analyze the tendencies in choosing cut-edit points in personal video content edited by users on the Internet, and develop a method to automatically estimate cut-edit points based on the results. When we investigated the relationship among cut-edit points in personal videos using a space-time patch feature(ST-patch feature), we realized that cut-edits were done in frames with a low Continuous Rank-Increase Measure (CRIM) value and a high Motion Correlation (MC) value calculated from the ST-patch feature. Therefore, we propose a method for estimating cut-edit points based on CRIM and MC values. Experimental results indicate that we obtained a recall ratio of 61.3% and a precision ratio of 50.4%.

Keywords— personal video, cut-edit point, camerawork, ST-patch feature

1. INTRODUCTION

Recently, there has been a significant increase in the amount of shared video contents, that is, video content created by individuals and distributed via the Internet for many people to enjoy. Because videos must be edited to create eye-catching content within the volume limitations placed on the video data to be distributed, there is growing demand for support systems that enable users to easily edit video content on the Internet.

As a conventional method of automatic video editing, Kumano $et\ al.$ proposed automatic extraction of baseball highlight scenes [1], and Ozeki $et\ al.$ proposed a method for desktop manipulation video editing [2]. Since these methods are aimed at the video content of TV programs, recorded by professional cameramen, they are not applicable for personal videos created by individual users. Iwaki $et\ al.$ and Kumano $et\ al.$ proposed automatic video editing systems for personal video[3][4]. These systems extract scenes by using multi-modal information such as sound, camerawork, and the motions of humans in the video.

In this study, we analyzed the tendency of choosing cut-edit points in personal video content edited by users on the Internet, and developed a method to automatically estimates cut-edit points based on the results. We evaluated our method by carrying out objective and subjective experiments to demonstrate its effectiveness.

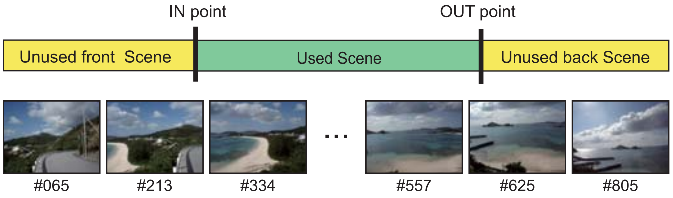


Fig. 1. Cut-edit points

2. PERSONAL VIDEO AND CAMERAWORK CHARACTERISTICS

We manually investigated the existence or nonexistence of camerawork in every scene of edited personal videos to determine where users tend to cut scenes in the videos. Our research was aimed at 1075 personal videos posted on a video-sharing site called “ClipCast”[5].

2.1. Personal video editing on ClipCast

ClipCast has various genres of personal videos uploaded by individual users, such as people, animals, sports, and parties. Users can edit their videos directly on the webpage. Since the editing history is archived, we can analyze the start frames and end frames of cut-edited scenes. In this paper, as shown in Figure 1, we define the following terms: the “IN point” is the start frame, the “OUT point” is the end frame, “Used scene” refers to scenes that are selected for use, and “Unused scene” refers to unwanted scenes. In addition, Unused scenes are classified as “Unused front scene” and “Unused back scene”.

2.2. Camerawork occurrence on Used and Unused scenes

Table 1 lists the results of analyzing the camerawork occurrence in every frame of Used and Unused scenes in the personal videos. As indicated in the table, we can see that the camerawork occurrence in Used scenes is 27.0% which is three times higher than that of the Unused scenes. This is because users are likely to select the scenes that contain some camerawork. We also see that the camerawork of “Follow” is used most often. This means users follow a subject while recording.

978-1-4244-7493-6/10/$26.00 ©2010 IEEE

Table 1. Camerawork occurrence in Used and Unused scenes

	Front	Used	Back
Number of frames	479117	1058145	421057
Camerawork frames	36412	285699	34107
Camerawork occurrence	7.6%	27.0%	8.1%
Follow	30.2%	24.1%	27.4%
Pan left	12.8%	17.8%	14.5%
Pan right	14.9%	19.0%	16.2%
Dolly	13.8%	12.7%	14.0%
Zoom in	8.0%	7.2%	7.8%
Zoom out	9.1%	8.0%	8.9%
Tilt up	5.0%	5.9%	5.4%
Tilt down	6.2%	5.2%	5.8%

2.3. Camerawork occurrence for IN and OUT points

Table 2 lists the results of camerawork occurrence for IN and OUT points. We see that the ratio of camerawork occurrence for IN and OUT points is between the ratios of Used scenes and Unused scenes shown in Table 1. We also see that the ratios of camerawork occurrence at the IN and OUT points are almost the same. Furthermore, the ratio of each type of camerawork also tends to be similar for each point. Therefore, the IN and OUT points are considered to have similar characteristics.

Table 2. Rates of camerawork for IN and OUT points

	IN point	OUT point
Number of frames	1075	1075
Camerawork frames	147	151
Camerawork occurrence	13.7%	14.0%
Follow	32.4%	24.8%
Dolly	17.3%	23.4%
Pan left	13.9%	14.4%
Pan right	13.2%	13.0%
Zoom in	9.8%	8.9%
Zoom out	7.4%	6.7%
Tilt up	3.2%	3.1%
Tilt down	2.8%	5.7%

3. TENDENCIES FOR CUT-EDITING POINTS BY USING ST-PATCH

To obtain more significant tendencies in order to estimate IN and OUT points, we analyzed cut-edit points in terms of motion occurring in the scenes by using the space-time (ST) patch feature. ST-patch features calculated form a ST-patch, which is a local region of images that extends in the time direction, are the spatio-temporal features containing information on both the "appearance" and "motion" simultaneously [6]. We also analyze the Continuous Rank-Increase Measure (CRIM) value of the ST-patch feature in the scenes with camerawork and no camerawork.

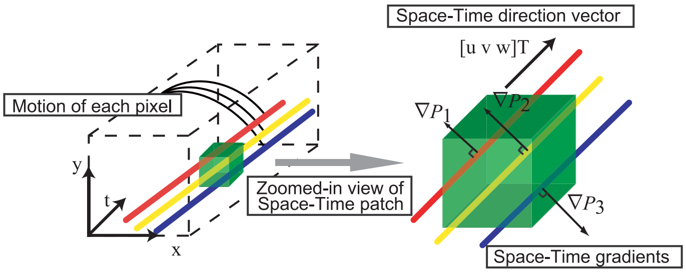

Space-Time direction vector
[u v w]T
Motion of each pixel
∇P1
∇P2
y
t
x
Zoomed-in view of Space-Time patch
∇P3
Space-Time gradients

Fig. 2. Overview of ST-patch

3.1. ST-patch feature

Figure 2 shows an overview of the ST-patch. The three colored lines represent the motion of each pixel, where $[u\ \ v\ \ w]^{\mathrm{T}}$ is a space-time direction vector in the ST-patch, and ∇P_i represents the space-time gradients. ST-patch features are extracted from the x, y, and t axis gradients in the images that extend in the time direction. Let the matrix that is stacked by these space-time gradients from all n pixels within the ST-patch $P(i = 1, \cdots, n)$ denote **G**, and let the matrix that is multiplied by the transpose of the gradient matrix **G** be denoted as **M**.

$$\mathbf{M} = \mathbf{G}^{\mathrm{T}}\mathbf{G} = \begin{bmatrix} \sum P_x^2 & \sum P_x P_y & \sum P_x P_t \\ \sum P_y P_x & \sum P_y^2 & \sum P_y P_t \\ \sum P_t P_x & \sum P_t P_y & \sum P_t^2 \end{bmatrix} \qquad (1)$$

Matrix **M** is an ST-patch feature. This matrix **M** contains information on both the "appearance" and "motion" simultaneously.

3.2. Motion analysis

∇P resides in a 2D plane if there is single uniform motion within the ST-patch. Information about the spatial properties of P is captured in the 2×2 upper-left minor **M** of the matrix **M**.

$$\mathbf{M}^{\diamond} = \begin{bmatrix} \sum P_x^2 & \sum P_x P_y \\ \sum P_y P_x & \sum P_y^2 \end{bmatrix} \qquad (2)$$

For an ST-patch with a single uniform motion, the following rank condition holds: $rank(\mathbf{M}) = rank(\mathbf{M}^{\diamond})$. When an ST-patch which contains more than one motion, the difference in rank cannot be more than 1, because only one column/row is added in the transition from $\mathbf{M}^{\diamond}$ to **M**. Thus, measuring the rank-increase Δr between **M** and its 2×2 upper-left minor $\mathbf{M}^{\diamond}$ reveals whether the ST-patch P contains a single motion or multiple motions:

$$\Delta r = \mathrm{rank}(\mathbf{M}) - \mathrm{rank}(\mathbf{M}^{\diamond}) = \begin{cases} 0 : \mathrm{single\ \ motion} \\ 1 : \mathrm{multiple\ \ motions} \end{cases} \qquad (3)$$

3.2.1. *Continuous Rank-Increase Measure (CRIM)*

The rank-increase Δr cannot measure motion similarity between two different ST-patches. Therefore, the continuous rank-increase Δr is denoted by using eigenvalues of **M** and $\mathbf{M}^{\diamond}$. Let

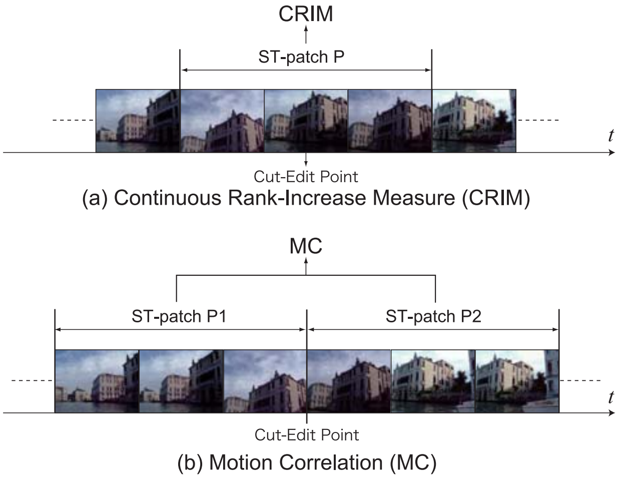

CRIM
ST-patch P
t
Cut-Edit Point
(a) Continuous Rank-Increase Measure (CRIM)
MC
ST-patch P1
ST-patch P2
t
Cut-Edit Point
(b) Motion Correlation (MC)

Fig. 3. ST-patch analysis

$\lambda_1 \geq \lambda_2 \geq \lambda_3$ be the eigenvalues of the 3×3 matrix $\mathbf{M}$. Let $\lambda_1^\diamond \geq \lambda_2^\diamond$ be the eigenvalues of its 2×2 upper-left minor $\mathbf{M}^\diamond$. Then, the continuous rank-increase Δr is denoted by the following equation:

$$\Delta r = \frac{\lambda_2 \cdot \lambda_3}{\lambda_1^\diamond \cdot \lambda_2^\diamond} \qquad (0 \leq \Delta r \leq 1) \tag{4}$$

The case of $\Delta r \approx 0$ indicates consistent motion, and $\Delta r \approx 1$ indicates inconsistent motion.

3.2.2. Motion Correlation (MC)

The motion similarity between two different ST-patches, which are called P1 and P2, is calculated from continuous rank-increase Δr. Let Δr_1 be the continuous rank-increase of P1. Let Δr_2 be the continuous rank-increase of P2. Let Δr_{12} be the continuous rank-increase of an ST-patch combining P1 and P2. Then, the motion similarity is calculated from the following equation:

$$m_{12} = \frac{\min(\Delta r_1, \Delta r_2)}{\Delta r_{12}} \tag{5}$$

The case of $m_{12} \approx 1$ indicates high similarity, and $m_{12} \approx 0$ indicates low similarity.

3.3. ST-patch features on cut-edit points

To capture the global motion in the images, we computed the ST-patch features from the down-sampled images. We computed the CRIM value from an ST-patch on a cut-edit point as shown in Figure 3(a), and the MC value from two ST-patches on the front and back frames around the cut-edit points as shown in Figure 3(b).

Figure 4 plots the histograms of CRIM and MC on the IN and OUT points. We see that CRIM is distributed around the low values, and MC is distributed around the high values. This means there are consistent motions at the IN and OUT points, and little changes in motions in the front and back frames around the IN and OUT points.

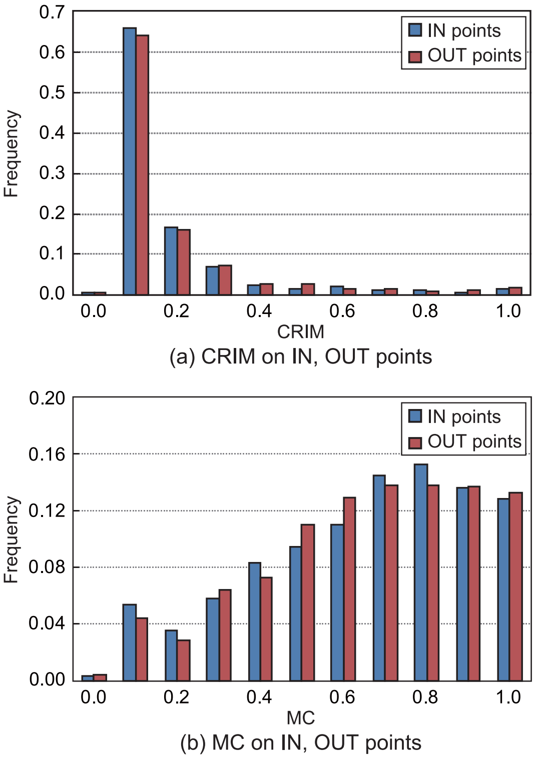


Fig. 4. CRIM and MC on the cut-edit points

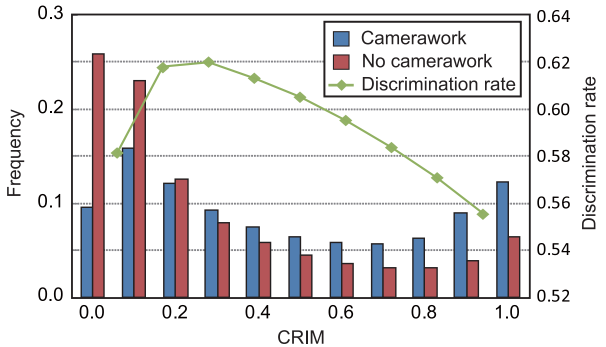


Fig. 5. CRIM values for scenes with and without camerawork

3.4. ST-patch features on camerawork

In section 2.2, we described the tendency we found for users to select scenes in which camerawork occurs. Therefore, we computed the CRIM value in scenes with and without camerawork. Figure 5 shows the histogram of CRIM values in these scenes. We see that around the high values, the CRIM of camerawork scenes has a higher frequency distribution than that of no-camerawork scenes, and around the low values, the CRIM of no-camerawork has a higher frequency distribution than that of camerawork. Thus, when we set the CRIM value for frames with camerawork to be higher than a threshold, and the frames without camerawork to be lower, we obtain a discrimination rate of 61.9% at the threshold of 0.3. Therefore, it is possible to estimate which scenes in a video include a lot of camerawork by using CRIM.

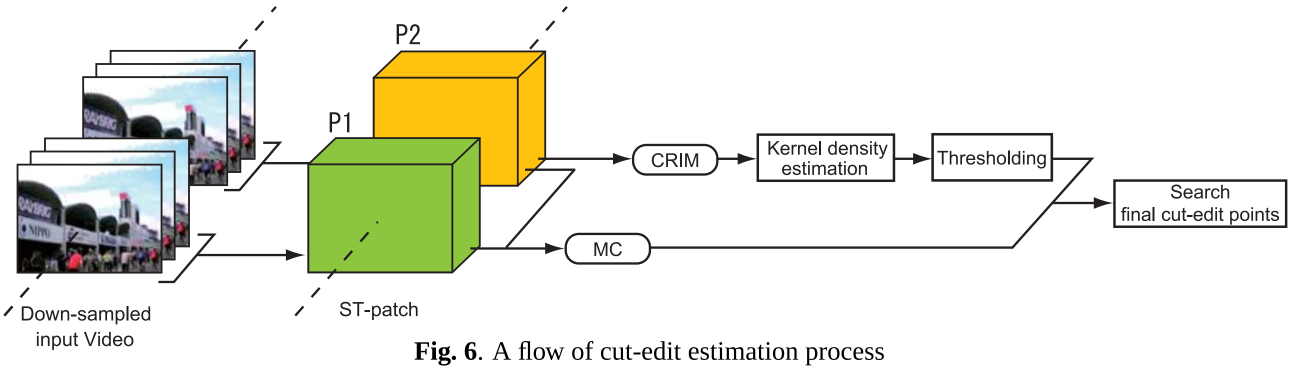


Fig. 6. A flow of cut-edit estimation process

Algorithm1 Algorithm used to estimate the cut-editing points.

```
Step1. Preprocessing:
  ·Input $N$ frames video $F = (f^i|f^0, f^1, \cdots, f^N)$.
  ·Down-sampling and smoothing of input image $f$.
Step2. Extraction of features:
  · Extract features from every frame.
    For $i = 1, \cdots, N$
      - Extract ST-patch feature $\mathbf{M}^i$ from ST-patch $P^i$.
      - Compute CRIM $\Delta r^i$ from ST-patch feature $\mathbf{M}^i$.
      - Compute MC $m_{12}^i$ from two different ST-patches.
Step3. Estimation of cut-candidate points $c$:
  ·Compute $p(x)$ from CRIM values $\Delta r_i (i = 0, \cdots, N)$ by using
   the kernel density estimation.
  ·Estimate the cut-candidate points $c$.
    - cut-candidate points $c$ are estimated as the intersection of $p(x)$
      with a the a threshold of 0.3
Step4. Search for the final cut-edit points $c'$:
  ·Search for the frame $c'$ with the highest value of MC $m_{12}^i$
   from the frames around the candidate point $c$.
      $c' = \underset{i \in \text{ 20 frames around } c}{\text{argmax}} \; m_{12}^i$
```

4. AUTOMATIC ESTIMATION OF THE CUT-EDITING POINTS

From the CRIM and MC results, we obtain the tendencies for IN and OUT frames and frames with and without camerawork. We therefore propose a method for estimating the cut-edit points based on these tendencies. In this section, we describe the flow of estimating cut-edit points and show examples of the estimation results. After that, we explain how we evaluated our method by conducting objective and subjective experiments to demonstrate its effectiveness.

4.1. Algorithm

Figure 6 shows the flow of estimating cut-edit points. **Algorithm 1** is the algorithm used to estimate the cut-edit points. First, input images are down-sampled to 80 × 60 pixels (i.e. this size is quarter of a original image) and smoothed. ST-patch features are computed from these whole images, and CRIM and MC values are computed from the ST-patches. Then, the kernel density estimation is performed on CRIM values using a probability density function. The probability density function is calculated by the following equation:

$$p(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(2\pi h^2)^{\frac{M}{2}}}\exp\left\{-\frac{|x-x_i|^2}{2h^2}\right\} \quad , \qquad (6)$$

where h is the bandwidth and N is the number of frame of input video. Figure 7 shows examples of estimated cut-edit points using kernel density estimation. The kernel density estimation can treat a sparse data distribution as a dense data distribution. Then cut-candidate points are estimated by thresholding for the probability density function $p(x)$. After that, from the 20 frames around the candidate point, the frame with the highest value of MC is determined to be the final cut-edit point. Therefore, scenes with a lot of changes in motion in the video are selected as Used scenes.

4.2. Objective evaluation

We evaluated our method objectively using three retrieval measures [7]: Recall is the proportion of correct retrievals compared to all possible correct retrievals. Precision is the proportion of correct retrievals among all retrieval results. The F-measure summarizes both into one number. Let "T" denote the number of frames of Used scenes edited by users, "S" denote the number of frames of Used scenes estimated by our method, and "C" denote the number of frames of Used scenes found by our method that includes the same scenes as those selected by users:

$$R = \frac{C}{T} \times 100\ [\%] \qquad (7)$$

$$P = \frac{C}{S} \times 100\ [\%] \qquad (8)$$

$$F = \frac{2PR}{P+R}\ [\%] \qquad (9)$$

We evaluated 658 personal videos posted to the ClipCast site by users. Table 3 lists the estimation results for each type of camerawork. We obtained an F-measure of about 60% when the camerawork was "Follow," "Pan left," "Pan right, " and "Dolly." In contrast, the lowest F-measure was obtained when the camerawork was "Fixed."

Table 3. Estimation results of cut-edit points [%]

Camerawork	Recall	Precision	F-measure
Follow	58.2	60.2	59.2
Pan left	64.9	54.2	59.0
Pan right	74.3	53.7	62.3
Dolly	78.1	47.8	59.3
Zoom in	61.6	43.7	51.1
Zoom out	57.9	56.4	57.1
Tilt up	56.6	50.9	53.6
Tilt down	62.2	37.0	46.4
Fixed	38.3	50.0	43.4
Average	61.3	50.4	54.6

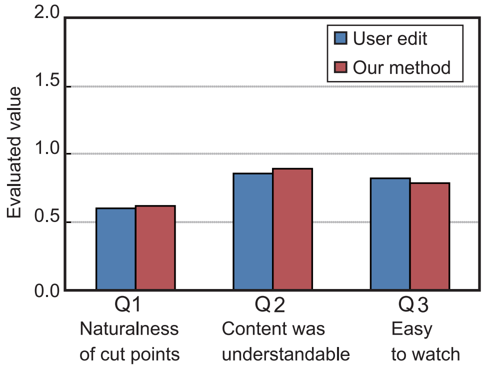


Fig. 8. Results of questions Q1-Q3

4.3. Subjective evaluation

We conducted a questionnaire survey to determine the effectiveness of our method by comparing the videos edited by users with those by our method.

4.3.1. Evaluation procedure

The contents of the questionnaire were as follows:
Q1. Were the cut-edit points natural?
Q2. Was the content of the video easy to understand?
Q3. Was the video easy to watch?
The steps of the questionnaire survey were as follows:
Step1. Subjects watched videos that had not been edited.
Step2. They watched videos edited by users and by our method respectively.
Step3. They answered the questionnaire.
The subjects consisted of 20 adults. The subjects rated the videos on a 5-degree scale (very bad, bad, neither bad nor good, good, very good).

4.3.2. Evaluation results

Figure 8 shows the results of questions Q1-Q3. There was no significant difference between the videos edited by users and those edited by our method. We conducted a Student's t-test (p level less than or equal to 5%), and also found no significant difference between the two methods. The results of the subjective evaluation show that our method can be applied to edit personal videos.

5. CONCLUSION

In this study, we first analyzed the tendency for users to choose cut-edit points in personal videos by investigating the editing history. Then, we obtained tendencies for IN and OUT frames and frames with camerawork and without camerawork by using CRIM and MC. Based on these results, cut-edit points were automatically estimated by using CRIM and MC values. In the objective evaluation, the averages of recall ratio and precision ratio were 61.3% and 50.4%, respectively. In addition, questionnaire results showed that there was no significant difference between the videos edited using our method and those edited by users themselves. Therefore, our method makes it possible to automatically edit personal videos, which simplifies the process and lightens the burden on users. In a future study, we plan to analyze editing tendencies by using the context in the video.

6. REFERENCES

[1] M. Kumano, Y. Ariki, K. Tsukada, S. Hamaguchi, and H. Kiyose, "Automatic extraction of pc scenes based on feature mining for a real time delivery system of baseball highlight scenes," *Proc. IEEE Int. Conf. Multimedia and Expo 2004*, pp. 277–280, 2004.

[2] M. Ozeki and Y. Nakamura, "Evaluation of self editing based on behaviors-for-attention for desktop manipulation videos," *IEEE International Conference on Multimedia and Expo (2006)*, pp. 329–332, 2006.

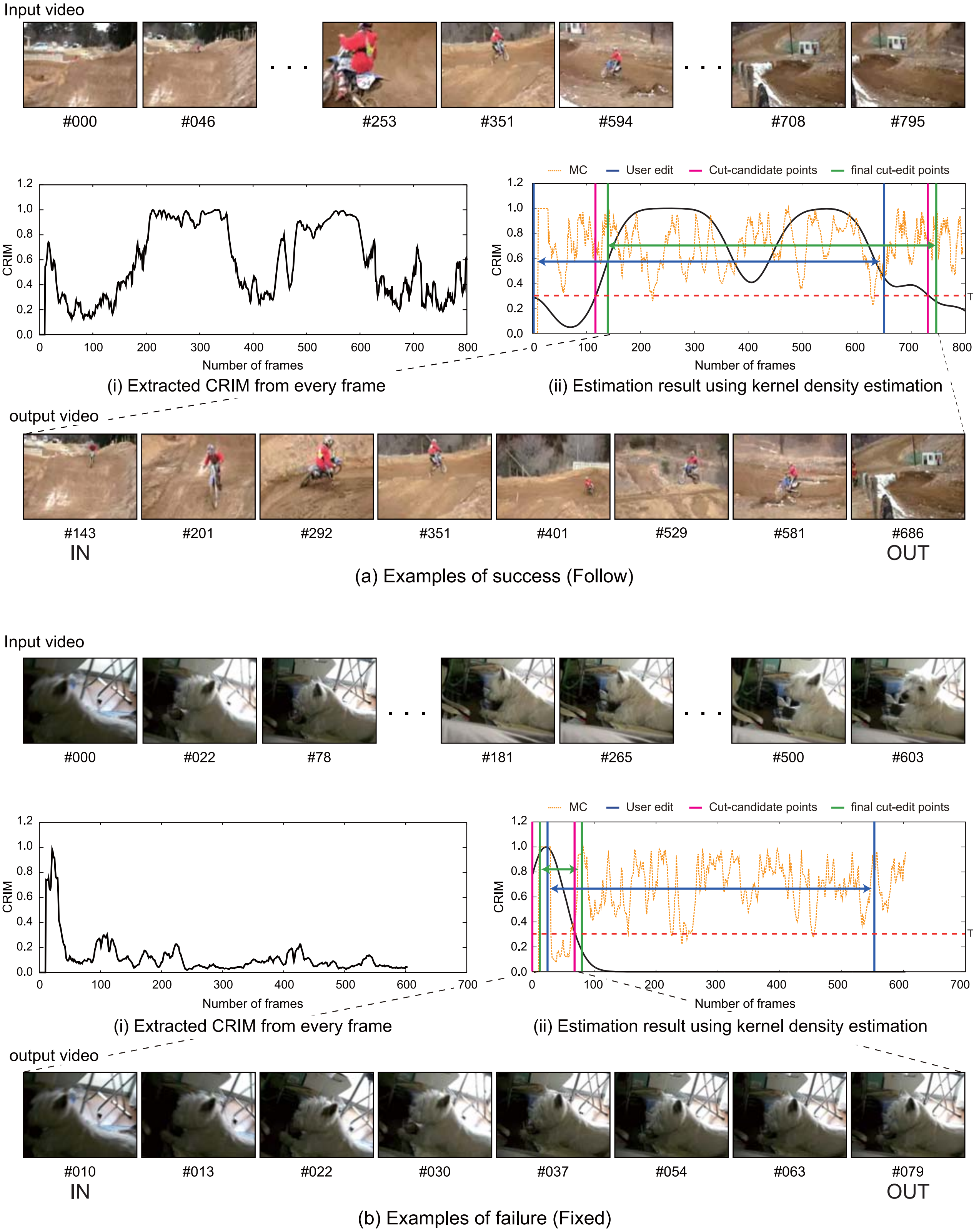
[3] K. Iwaki, M. Nakazawa, and S. Hattori, "An automatic editing system of a personal video stream," *Technical report of IEICE. MVE*, vol. 102, no. 737, pp. 1–4, 2003.

[4] M. Kumano, Y. Ariki, K. Tsukada, and K. Shunto, "Automatic extraction of useful shot sections for a video editing support system based on video grammar," *The Journal of The Institute of Image Information and Television Engineers*, vol. 57, no. 7, pp. 829–839, 2003.

[5] http://clipcast.jp, "Clipcast," .

[6] E. Shechtman and M. Irani, "Space-time behavior based correlation -or- how to tell if two underlying motion fields are similar without computing them?," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), November 2007.*, vol. 29, no. 11, pp. 2045–2056, 2007.

[7] C.J.van Rijsbergen, *Information retrieval*, Burrerworths, London, UK, 1979.



(a) Examples of success (Follow)

(b) Examples of failure (Fixed)

Fig. 7. Example of estimating cut-edit points