Automatic Creation of Cooking Highlight Videos by Sound Loudness and Optical Flows

Takumi Kitase* and Takafumi Koike**

Graduate School of Computer and Information Sciences, Hosei University,
Kajino 3-7-2, Koganei, Tokyo 184-8584, Japan

* takumi.kitase.4j@stu.hosei.ac.jp, ** takafumi@hosei.ac.jp

Abstract

We present a method for creating a highlight video of cooking as social media. If we can easily prepare cooking videos, it is expected to be used for communication. However, it is difficult to extract entertaining scenes to make highlight video of your cooking because cooking videos also contain so many unnecessary scenes. We use sound and image features to extract entertaining scenes without complexed-processing. First, we extract scene segments with sound larger than a pre-determined threshold. Second, we define two evaluation scores: frame score, which is defined as the standard deviation of the optical flow in each frame divided by the mean of the optical flow, and scene score, which is defined as the mean score of all frame scores in a scene. Finally, we evaluate the scene score of all extracted scenes and generate the resulting highlight video by arranging the scenes with top 20 scores in time order. Experimental results show that our method can extract entertaining scenes related to cooking.

Keywords: Cooking video highlight, sound loudness, optical flow

1. Introduction

Cooking is not only for making dish but also for communication. One common example is posting photos or videos which include dishes they made to social media, such as Twitter and Facebook. The main aim of this activity is to get as many reactions from others about the post. Videos can include more information such as sound and dynamic motions compared to photos. Therefore, posting videos which include their cooking activities can get more reactions from others compared to photos.

Kitase et al. have presented an idea of automatic creation method of cooking highlight which will be post to social media by image feature^(1,2). Posting activity as video has laborious editing tasks. They are supposed to create highlight of their video because social media users do not like watching lengthy posts, making a highlight video from the original video before posting is preferred. They cannot post in real time because of the editing task.

Many video summarization techniques are proposed, and some of them focus on cooking videos. However, most of them mention the accuracy of summarization, not the real-time performance. Hayashi et al. have proposed an automatic cooking video editing system considering information of corresponding cooking recipe⁽³⁾. In this method, redundant scenes in cooking videos such as intervals between operations and waiting time are reduced by referring to the cooking instructions in the corresponding recipe. Doman et al. have presented a method which links videos describing cooking operations to corresponding terms in recipes representing the operation to understand how to do the cooking operation clearly⁽⁴⁾. These methods require corresponding cooking recipes in addition to cooking videos. Short processing time of cooking highlight will facilitate posting on social media. However, these methods need time for preparing cooking recipe and for analyzing descriptions in cooking recipe. For this reason, these methods making it unsuited for posting social media when users post their cooking activity as video.

In this paper, we present a method which automatically creates entertaining highlight videos from cooking videos in quasi real-time. Cooking highlight which is created by our method consists several scene segments, a brief sequence of video frames. This method consists of three steps. First, the method extracts scene segment whose sound has large value

Fig. 1. Process of our method.

more than specified threshold. Next, the method evaluates extracted scene segments and calculate each segment's score by function which defined by optical flow in the video frame. Finally, the method creates highlight video with several highest scored scenes.

2. Automatic Highlight Creating Method for Cooking Video

Fig. 1 shows the process of our method. Cooking video which records individually cooking (for example, cooking in home) is an input of our method. Our method creates cooking highlight video through three steps.

2.1 Policy on Creating Highlight Video

We define two policies for creating cooking highlight video by reference to some Tasty⁽⁵⁾ videos. These policies are “Extracting entertaining scenes” and “Shortening cooking highlight”. We describe explanations of policies below.

(a) Extracting entertaining scenes

The recorded video includes no-events scenes, which are not record any events such as scenes where nobody is present. Cooking highlight should only include entertaining scenes. Our method uses scene segment extraction by sound and scene segment evaluation with optical flow for extracting entertaining scenes.

(b) Shortening cooking highlight

As we described in *Introduction*, lengthy post is not appropriate for social media. Since cooking highlight video should have short duration, we have to choose most entertaining scene segments. Our method uses optical flow based scene segment evaluation for choosing entertaining

Fig. 2. Scene segment extraction by using threshold θ_a . Aligned boxes under the graph are video frames in video. t axis indicates the time of input video data. When our method detects $|v|$ over θ_a , our method starts to extract scene segment for 1 second.

scene segments.

2.2 Extraction of Scene Segment by Sound

Our method extracts scene segment where the value of wave form of sound is high. We use absolute value of wave form of sound in video as a feature of video sound. We assume scenes with loud sound may include some events. We describe the absolute value of wave form of sound as $|v|$. We also define θ_a as a threshold. Our method take $|v|$ which located head of the frame as a representative value of each input video frame. When our method detects $|v|$ which is larger than θ_a in some frame, our method starts to extract scene segment for a second from the frame which includes detected $|v|$ (Fig. 2). While extracting scene segments, our method doesn't check whether $|v|$ is over θ_a or not (see the second extraction in Fig. 2).

2.3 Scene Segment Evaluation with Optical Flow

After extracting scene segments with wave form value of sound, our method evaluates each scene segment by calculating its score. A function which uses the estimated optical flow of each frame calculates this score. In this section, we describe how to estimate optical flow and use estimated optical flow to evaluate the extracted scene segments.

(a) Optical Flow Estimation

A scene which include active motions such as cooking motion will be an entertaining scene. To evaluate movements in the scene segment, we use optical flow as a feature. We use Lucas-Kanade method⁽⁶⁾ to estimate optical flow for each

Fig. 3. Optical flow estimation. Our method uses norm of estimated optical flow where pixels located at multiple of 16 coordinates to evaluate scene segment.

frame. We use norm of estimated optical flow 16-pixel interval from the top left pixel in the frame (Fig. 3) to calculate scene segment's score. On the following, we describe this value as $|s(x,y)|$. $s(x,y)$ indicates estimated optical flow of (x,y) pixel in the frame.

(b) Evaluation Functions

We introduce several functions to evaluate each scene segment's score. Each scene segment has own score by function. Here, we define the score of k th frame in scene segment as

$$F_k = \frac{\sigma(s)}{M}, \quad (1)$$

where $\sigma(s)$ is standard distribution of $|s(x,y)|$ in the frame. M is the average of $|s(x,y)|$ in the frame. Therefore, we define M as

$$M = \sum_{m=1}^X \sum_{l=1}^Y \overline{s(d \times m, d \times l)}. \quad (2)$$

d is coefficient of skipping pixels. According to (a) in 2.3, in fact, coordinates $(d \times m, d \times l)$ always be in multiples of 16. Therefore, $d=16$ in (2), the range of X is no more than 1/16 of frame width, the range of Y is no more than 1/16 of frame height. When some objects are moving, there are some large $|s(x,y)|$ around the area of the objects. Then, $\sigma(s)$ will be much higher than M (Fig. 4). Hence, F_k will be a large value if the magnitude of optical flows in frame k are large.

We define the i th scene segment's score S_i as

$$S_i = \sum_{k=2}^f \overline{F_k}. \quad (3)$$

In this equation, each scene segment has f frames per second. As mentioned in 2.2, each scene segment has f frames since it is one second long. The starting value of k is two because

(a) Graph of M and $\sigma(s)$.

(b) Corresponding scene of (a).

Fig. 4. Sample data of M and $\sigma(s)$ and corresponding scene. The scene shows a cook is cutting vegetables. High $\sigma(s)$ in graph corresponds motion of cutting with knife in the video.

Fig. 5. Creating cooking highlight video in time order. Each box indicates scene segments and a number inside box indicates the order of score in whole scene segments. Underlined numbers indicate scene segment which will be used for cooking highlight video.

the first frame in scene segment usually depicts scenes different from the previous frame in previous scene segment. Therefore, optical flow estimation doesn't work correctly in the first frame and its evaluate function F_1 become invalid.

2.4 Creation of Highlight Video

Our system chooses the scene segments in descending order of score at the final step of our method. The number of

Fig. 6. Experiment environment on a kitchen for shooting cooking video. Dotted area shows an area of shooting by a video camera.

scene segments in highlight video can set manually. Our method aligns chosen segments in time order (Fig. 5).

3. Experiment

3.1 Experimental Condition

We have captured a video of cooking chocolate tarts at a kitchen in house for experiment. The video includes scenes from the start of cooking to the timing of putting chocolate tarts in refrigerator. We show an environment of the kitchen in Fig. 6. A camera put up on a worktop to avoid capturing any movements which are not related to cooking such as person moving far from kitchen.

The duration of the recorded video is 13m03s, audio sampling rate is 48,000Hz. The frame rate of recorded video is 25 fps, the resolution of the recorded video is 640×360 pixels. In the experiment, range of $|v|$ is $0.0 \leq |v| \leq 1.0$ and we set $\theta_a = 0.2$. We consider this value on the basis of distribution of the recorded sound of the video. We set the value of f is 25 in equation (3). This is the frame rate of the recorded video. We set the number of scene segments in cooking highlight is 20.

3.2 Evaluation for Experiment

For the experiment, we define the evaluation items and set variables related to our method. We define evaluation items to evaluate the performance of our method: “performance of scene segments extraction by sound” and “performance of evaluate function”. In “performance of scene segments extraction by sound”, it evaluates our method fulfills the policy of “Extract entertaining scenes” in highlight video as described in 2.1. In “performance of evaluate function”, it evaluates our method can fulfill the

policy of “Shortening cooking highlight” in 2.1. In this experiment, it is evaluated by how many scene segments with motions related to cooking on the worktop in created highlight video.

3.3 Results

Applying the scene segment extraction by sound, our method extracted 45 scene segments. Examples of major scene segments in the extracted video are: cracking chocolate and putting the pieces in a bowl, inserting the bowl into microwave oven to melt chocolate outside of worktop, stirring chocolate in the bowl, placing a dish with tart crusts on the worktop. Fig. 7 shows thumbnails of scene segments extracted by extraction scene segments by sound.

3.4 Discussion

In the step of extraction scene segments by sound, our method extracts many scenes which include motions as for 3.3. However, our method extracted scenes of inserting bowl into microwave outside of the shooting area. Our method should not have extracted this scene because this movement in the scene performed outside of capture area. Another problem is a disability to extract entertaining but quiet scenes. Our method isn’t able to extract a scene of decorating tarts. The scene should be important because the scene indicates the visual change of tart. We assume that our method can’t extract this scene because of small sound. Decorating scene has small sound such as sprinkling sugar compared to other dynamic scenes.

In the step of scene segment evaluation by optical flow, most of all scene segments in created highlight video are dynamic scene segments. By applying evaluation function, our method doesn’t use microwave-scene segment as we mentioned before to cooking highlight. In Fig. 7, the thumbnails of scene segments in the second row are microwave-scene segments are thumbnails in the second row. We assumed the cause of ignoring these segments as lack of motion and following low score evaluation. However, our method doesn’t use the scene segment of finished chocolate tarts on a dish. We assume that is caused by the same reason.

4. Conclusions

We aim to create an entertaining highlight video for cooking, which viewers find very entertaining. We have presented a method for creating a cooking highlight video by sound loudness and optical flow to extract cooking scenes. Our method can extract many dynamic cooking motions

Fig. 7. Thumbnails of the extracted scene segments in 3.3 in time order. Generated highlight video includes numbered thumbnails in numerical order. Others are thumbnails of scene segments which is not used for highlight video.

which are expected to attract viewers and avoid to include no-event scenes which should not be included in highlight. For the future work, we will improve the accuracy of choosing entertaining scenes.

References

- (1) T. Kitase and T. Koike: “Automatic Creation of Cooking Highlight Video by Image Features,” The 78th National Convention of IPSJ, 2016. (in Japanese)
- (2) T. Kitase and T. Koike: “Automatic Creation of Highlight Video for Cooking by Applying Optical Flow Extraction,” VRSJ Research Report, Vol. 21, No. CS-2, pp. 1-6, 2016. (in Japanese)
- (3) Y. Hayashi, K. doman, I. Ide, D. Deguchi, and H. Murase: “Automatic authoring of a domestic cooking video based on the description of cooking instructions,” Proceedings of the 5th International Workshop on Multimedia for cooking & eating activities, pp. 21-26, 2013.
- (4) K. Doman, C. Y. Kuai, T. Takahashi, I. Ide, and H. Murase: “Video CooKing: Towards the synthesis of multimedia cooking recipes,” International Conference on Multimedia Modeling, pp.135-145, 2011.
- (5) BuzzFeed: “Tasty”, <http://bit.ly/tastyYT>, 2016. (browse on May 11, 2017)
- (6) B. D. Lucas and T. Kanade: “An Iterative Image Registration Technique with an Application Stereo Vision,” Proceedings of the 7th International Joint Conference on Artificial Intelligence, Vol. 2, pp. 674-679, 1981.