Motion Based Automatic Abstraction of Cooking Videos

Koichi MIURA
Reiko HAMADA
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo, 113-0033, Japan
TEL: +81-3-5841-7413
FAX: +81-3-5800-6922
miura,reiko@mtl.t.u-tokyo.ac.jp

Ichiro IDE
Nat'l Institute of Informatics
2-1-2 Hitotsubashi,
Chiyoda-ku,
Tokyo, 101-8430, Japan
ide@nii.ac.jp

Shuichi SAKAI
Hidehiko TANAKA
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo, 113-0033, Japan
sakai,tanaka@mtl.t.u-tokyo.ac.jp

ABSTRACT

In this paper, we propose a method to abstract cooking videos. We define cooking video abstraction as shrinking videos maintaining the understandability of general cooking procedures visually and intuitively. Cooking motions and appearances of foods are considered as important segments in a cooking video. A method to extract such important segments referring to the intensity of motion in an image is proposed, and effectiveness of the method is shown through evaluation experiments. We also developed a cooking video abstraction system that assembles important segments detected by the proposed method and repetitious motions that is especially important among cooking motions. The resultant abstracted videos were about 1/8 to 1/12 of the original videos in time, maintaining the understandability. The validity of the abstraction method was checked by comparing the abstracted videos with manually abstracted videos provided from a broadcasting station.

1. INTRODUCTION

Following the advance in telecommunication technology, large amount of multimedia data has become available from broadcast video. Multimedia data analysis is becoming important to store and retrieve them efficiently. However, characteristics of videos and purposes of viewers vary among different kinds of videos. Thus, it is necessory to limit the domain of the target video and refer to domain specific knowledge, for high level content analysis.

We chose cooking video as a target, and aim for building a practical system with high accuracy using domain specific knowledge[5, 8]. Cooking is familiar to daily life, and the demand for semantically structured cooking videos should grow in proportion to home (especially kitchen) automation.

In this paper, we propose a method to abstract cooking videos. A cooking program is a kind of an instruction video that people view from a practical point of view. On the other hand, cooking videos include many redundant segments such as chatting, which requires the viewers a certain amount of time to view. Therefore, to select recipes and to actually cook in daily life, a cookbook tends to be easier to browse. However, since videos contain visual information that text cannot express sufficiently, it is more effective to understand the cooking procedures.

Thus, we abstract cooking videos to easily browse through recipes. We define cooking video abstraction as shrinking videos maintaining the understandability of general cooking procedures visually and intuitively.

Various studies have been done on automatic video abstraction for news, documentary, and so on[4, 6]. These kinds of videos are relatively not redundant, and abstracted videos tend to be utilized for selecting a segment in the original videos. Works similar to ours are in sports video abstraction. A work on sports video abstraction[7] refers to an external database corresponding to video contents. This method is effective for videos that could easily refer to external data corresponding directly to their contents, which is difficult in cooking videos.

On the other hand, it is reported that abstracted videos are not so effective since split audio accompanying the abstracted images is too choppy[4]. From this point, audio naturalness is considered in abstracting videos[6, 9], and audio is usually replaced with new narrations in abstracted videos provided from a broadcasting station. However, since motion and preparation steps can be roughly understood solely from visual information in cooking videos, outline of the cooking procedure could be understood to a certain extent without audio. Thus, our abstraction does not consider audio continuity, and abstracted videos are created solely referring to image features.

Considering the above mentioned issues, cooking video could be considered as a good target for abstraction.

2. FEATURES OF COOKING VIDEOS

2.1 Structure of cooking videos

As shown in Fig. 1, shots in cooking videos could be categorized into (a)face shot, and (b)hand shot, which appear almost alternately, as shown in Fig. 2.

In face shots, almost the whole kitchen is shown. Though a teacher or an assistant explains a cooking procedure, their motions and foods are too small to obtain visual informa-

tion. On the other hand, hand shots are close-ups of tools and/or hands while cooking something that have rich visual information. However, each hand shot contains sub-shot segments such as cooking motion and appearance of foods that are important keys to understand the recipe, but also redundant segments between them as well.

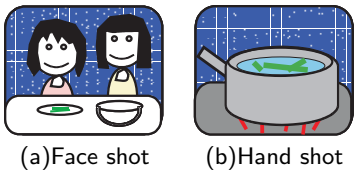


Figure 1: Shot categories in cooking video.

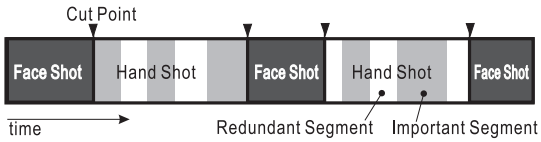


Figure 2: Structure of cooking video.

2.2 Important segments in cooking videos

In abstracting cooking videos which have the structure as shown in Fig. 2, first, face shots are excluded since they have little visual information. Secondly, redundant segments in hand shots are excluded. Here, we consider that (1)important visual information that text cannot express sufficiently and (2)essential information to understand cooking procedures visually, should be included.

Condition (1) appears as two kinds of video segments. One is (a)cooking motion. Tips of cooking motions cannot be easily understood without visual information. Another is (b)appearance of foods, such as color of prepared ingredients and state of a dish. Cooking videos include static segments without motions in order to show the state. When an abstracted video includes these segments, it is considered that it also satisfies condition (2).

In this work, therefore, we extract (a)cooking motion and (b)appearance of foods, to abstract cooking videos. Referring to actual cooking programs, we have examined that these segments have the following motion-related features.

(a) cooking motion: big (intense)
(b) appearance of food: almost none

3. IMPORTANT SEGMENT DETECTION

3.1 Motion based detection

In order to extract important segments considering the motion-related features, we need to detect motion in the image. Here, optical flow is employed to detect the motion since we need to refer to direction, speed, and so on. Among many techniques proposed to detect optical flow, we employed Horn and Schunck's method[2]. This relatively simple method was employed since our purpose requires only rough motion detection.

The following procedure is taken for motion-based important segment detection:

1. Detect cuts and classify shots into (a)face shot, and (b)hand shot.
2. Exclude (a)face shots.
3. Detect optical flows from (b)hand shots.
4. Sum up the length of optical flow vectors detected in all pixels (320×240) in each frame. (=S)
5. Take the average of the sum (S) of every 10 frames in order to reduce mal-effects from noise. (=$\overline{S}$)

Cut detection is performed by applying the DCT clustering method[1], and shots are classified by detecting faces. Faces are detected based on skin color information, shape, and so on, similar to the method described in [3].

Temporal transition of $\overline{S}$ of a part of an actual video is shown in Fig. 3. From this graph, important segments; (a)cooking motion and (b)appearance of foods, are detected.

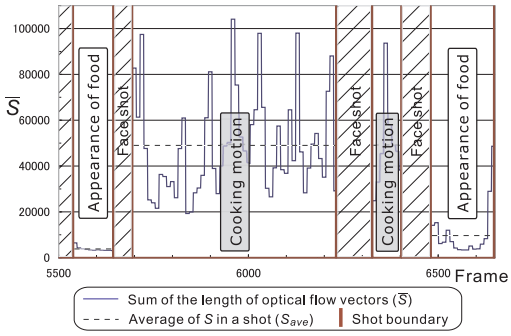


Figure 3: Temporal transition of the sum of the length of optical flow vectors ($\overline{S}$) in each frame.

First, S is the sum of the length of optical flow vectors in a frame, and $\overline{S}$ is the average of S of every 10 frames. S_{ave} is the average of S in a shot. S_{move}, S_{state1}, and S_{state2} are thresholds used for detection.

A segment that satisfies $\overline{S} \geq \alpha S_{ave}$ in a shot that satisfies $S_{ave} > S_{move}$ is judged as cooking motion (α is a constant). This means that a relatively active segment in a shot with a big motion is judged as cooking motion.

As for appearance of foods, a segment that satisfies $\overline{S} < S_{state1}$ for more than T sequential frames, or a segment that satisfies $\overline{S} < S_{state2}$ in a shot that satisfies $S_{ave} < S_{state2}$ is detected. This means that continuous inactive segments and an inactive segment in a shot that has almost no motion are judged as appearance of foods.

3.2 Camera motion exclusion

The above method misdetects also camera motions as cooking motions since only the sum of the length of optical flow vectors is referred to. Therefore, camera motion detection is necessary to exclude the misdetected segments.

Camera motions in cooking videos are categorized into *panning* (a translational motion of a camera) and *zooming* (a zoom up and down motion of a camera). *Panning* is observed when a camera moves from an object to an object, the purpose of which is not to show cooking motions nor appearances of foods. It is the most common camera motion in cooking videos and also the main cause of misdetection since there is usually no important visual information in the image. On the other hand, when *zooming* occurs, an important object usually appears in the center of the image. As a result, it can be neglected.

Thus, we concentrate on detecting and excluding *panning*. The following procedure is taken to detect *panning* simply referring to optical flow vectors:

1. Calculate directions (angles, $0 \leq \theta(i,j) < 2\pi$) of optical flow vectors in all pixels $p(i,j)$ of frame f. Weight

them with the length of vectors $v(i,j)$, and calculate the frequency of vectors by directions. This forms an "angle histogram"($H_f = \{\ h_f(\Theta) \mid 0 \leq \Theta < \pi\ \}$). Here, if the angle of a vector is $\pi \leq \theta(i,j) < 2\pi$, it is weighted as $-v(i,j)$ at $\Theta = \theta(i,j) - \pi$.

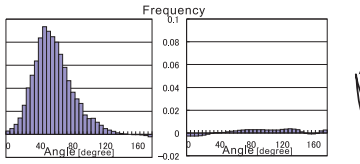
$$h_f(\Theta) = \frac{1}{S} \sum_i \sum_j \delta_\Theta(i,j) \cdot v(i,j) \quad (1)$$

$$\text{where,} \quad \delta_\Theta(i,j) = \begin{cases} 1 & (\,if\ \theta(i,j) = \Theta\,) \\ -1 & (\,if\ \theta(i,j) = \Theta + \pi\,) \\ 0 & (\,otherwise\,) \end{cases}$$

2. Sum up and take the average of the histograms($\overline{H} = \{\ \overline{h}(\Theta) \mid 0 \leq \Theta < \pi\ \}$) as long as the motion could be considered continuous (during frames f_1 to f_2). Continuous motion is detected by referring to $\overline{S}$. Since motions which have opposite directions are canceled mutually, random noise and cooking motions do not appear much in an "angle histogram", and *panning* is detected effectively.

$$\overline{h}(\Theta) = \frac{1}{f_2 - f_1 + 1} \sum_{f=f_1}^{f_2} h_f(\Theta) \quad (2)$$

As the result of this procedure, the angle histogram shows one clear peak in the case of *panning*, as shown in Fig. 4(a), and it shows no clear peak when there is no *panning*, as shown in Fig. 4(b). Taking advantage of this feature, when F_p is the peak value of the angle histogram and F_{th} is a certain threshold, a segment that satisfies $F_p > F_{th}$ with only one peak is considered as *panning*, and excluded from important segments.



(a) *Panning*. (b) Without *panning*.

Figure 4: Angle histogram.

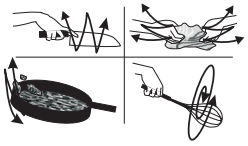


Figure 5: Repetitious motions.

3.3 Repetitious motion detection

Although the above proposed method detects cooking motion and appearance of foods as important segments, more important segment detection is necessary for more effective abstraction. We have examined cooking motions referring to actual cooking videos, and found that repetitious motions as shown in Fig. 5 are more important than others. Therefore, we detect repetitious motions following the method that focus on time periodicity[8] and reflect the result in the abstraction.

3.4 Evaluation: important segment detection

We made an evaluation experiment that detects important segments; cooking motions and appearances of foods. The target cooking videos consist of 6 recipes, with the length of approximately 40 minutes in total, taken from a Japanese cooking program.

In this experiment, shot classification was manually done. Cooking motions and appearances of foods were detected based on the method in **3.1**, and *panning* was excluded based on the method in **3.2**. Thresholds used in this experiment were $S_{move} = S_{state2} = 10,000$, $S_{state1} = 7,000$, $\alpha = 1.0$, $T = 90$ (3 seconds), $F_{th} = 0.025$. They were defined based on preliminary observations.

The result of the experiment is shown in Tab. 1. Correct segments were judged and defined manually. The number of correct answers of automatically detected segments is N_C, misdetection is N_M, and oversights is N_O. Recall is $N_C/(N_C + N_O)$, and precision is $N_C/(N_C + N_M)$.

Table 1: Result of important segment detection.

Important segments	N_C	N_M	N_O	Recall	Precision
Cooking motion	117	10	2	98%	92%
Appearance of foods	39	2	7	85%	95%

As shown in Tab. 1, the proposed simple method is effective to detect important segments. The main cause of misdetections of cooking motions and oversights of appearances of foods was detecting an unimportant motion, such as a motion not related to cooking.

As a result, approximately 40% (7 out of 17) of misdetections of cooking motions previously made without detecting camera motion (*panning*) were eliminated applying the method in **3.2**.

4. VIDEO ABSTRACTION

4.1 Motion based abstraction

We developed an application to abstract cooking videos using the method described in **3**.

In each hand shot, the first segment of repetitious motions is extracted if it includes any. If not, or in segments more than 10 seconds apart from repetitious motions, the first segment of cooking motions detected by the proposed method is extracted. Additionally, the last segment of appearances of foods is extracted. These segments are assembled in chronological order. Note that duration of each segment is set to 2 seconds, and shot classification is manually done.

An example of the resultant abstracted video is shown in Fig. 6. In Fig. 6, repetitious motions((4)∼(7), (11)) have rich visual information that text cannot express sufficiently, such as strength and speed of motion. Next, other cooking motions((2), (9), (10)) enable us to easily understand the cooking procedure visually. Finally, appearances of foods((1), (3), (7), (9), (12)) include important visual information and make cooking steps clear.

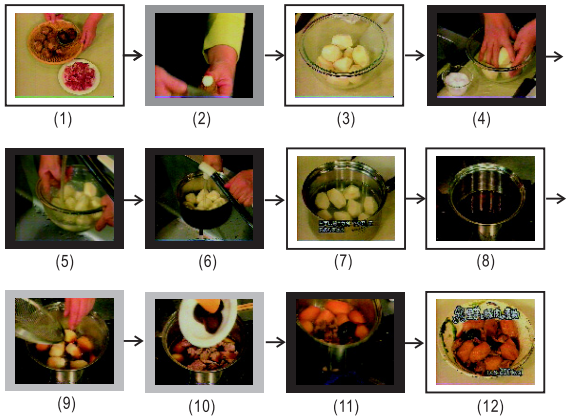


Figure 6: Video segments abstracted from a cooking video (Black frame: Repetitious motion, Gray frame: Cooking motion, White frame: Appearance of foods).

The abstraction method was applied to 9 recipes of 3 television programs from different broadcasting stations. The number of extracted segments for abstraction in each program for abstraction is shown in Tab. 2. The number of repetitious motions is N_R, other cooking motions is N_M, and appearances of foods is N_A.

Table 2: Number of extracted segments for abstraction.

Program	# of recipe	N_R	N_M	N_A	Abstraction rate (in time)
Program 1	4	23	23	18	$\sim 1/11$
Program 2	2	8	23	7	$\sim 1/10$
Program 3	3	8	70	2	$\sim 1/9$
Total	9	39	116	27	$\sim 1/10$

The abstracted videos were approximately 1/8 to 1/12 of the original videos in time. Nonetheless, we observed that the abstracted videos surely provide important visual tips as well as inevitable cooking steps that enable us to easily understand the cooking procedure.

4.2 Evaluation: video abstraction

Program 3 in Tab. 2, accompanies a manually abstracted video at the end of the program as "today's review". Thus, we compared the 3 manually abstracted videos in program 3 with our result.

First, as for the duration of the abstracted videos, manually abstracted ones had a fixed length (40 seconds), and the average of ours was 53 seconds.

Next, contents of both videos were compared. The result of comparison in number of extracted segments is shown in Tab. 3. Since our videos contain no audio, only visual information was compared. The number of segments in manually abstracted videos is Seg_H, segments in our videos is Seg_M, and commonly detected segments is Seg_C. Recall is Seg_C/Seg_H, and precision is Seg_C/Seg_M. Although there were some segments which represent the same visual information in our videos, they were regarded as one segment in the comparison.

Table 3: Comparison in # of extracted segments.

Recipe	Seg_H	Seg_M	Seg_C	Recall	Precision
Recipe 1	12	20	11	92%	55%
Recipe 2	13	24	13	100%	54%
Recipe 3	11	16	10	91%	63%
Total	36	60	34	94%	57%

As shown in Tab. 3, it turns out that segments in manually abstracted videos were mostly extracted by the proposed method with high recall. Actually, segments which cannot be extracted by our method were only 2.

On the other hand, precision was not so high. Although importance of most additional segments by our method which showed appearances of foods and basic cooking motions is little, they were necessary to represent cooking procedures visually. At the same time, some of the contents of these segments were supplemented by newly inserted captions and narrations in manually abstracted videos.

Tab. 4 shows the result of a rough comparison including the contents represented by captions and narrations. From this result, better precision is obtained.

However, manually abstracted videos are "reviews" for viewers who have already watched the original videos. This purpose is slightly different from ours; shrinking videos sufficient to understand general cooking procedures visually and intuitively.

Table 4: Comparison of extracted segments (including caption and narration contents).

Recipe	Seg_H	Seg_M	Seg_C	Recall	Precision
Recipe 1	17	20	16	94%	80%
Recipe 2	18	24	18	100%	75%
Recipe 3	12	16	11	92%	69%
Total	47	60	45	96%	75%

As for the duration of abstracted videos, we will improve it by setting the most suitable time for each segment, adjusting to the user's skill, and so on, as a future work.

5. CONCLUSIONS

This paper proposed a motion based abstraction method for cooking videos. A method to detect important segments referring to the intensity of motion in the image was proposed, and its effectiveness was shown through evaluation experiments. Additionally, we also developed a cooking video abstraction system that assembles important segments detected by the proposed method and repetitious motions. The resultant abstracted videos were about 1/8 to 1/12 of the original videos in time, maintaining the understandability of cooking procedures. The validity of the abstraction method was checked by comparing some automatic abstracted videos with manually abstracted videos provided from the broadcasting station.

In the future, we will develop a more effective abstraction system. For example, if the abstraction rate is adjustable, the application will be more useful. Therefore, we will further investigate on classifying and prioritizing the cooking motions depending on their importance to understand the cooking procedure.

6. REFERENCES

[1] Y. Ariki and Y. Saito. Extraction of TV news articles based on scene cut detection using DCT clustering. In *Proc. Intl. Conf. on Image Processing*, pages 847–850, 1996.

[2] B. K. P. Horn and B. Schunck. Determining optical flow. *Artif. Intel.*, 17:185–203, Aug. 1981.

[3] C. Wren, A. Azarbayajani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. PAMI*, 18(7):780–785, July 1997.

[4] M. Christel, M. Smith, C. Taylor, and D. Winkler. Evolving video skims into useful multimedia abstractions. In *Proc. CHI'98 Conf. Human Factors in Computing Systems*, pages 171–178, 1998.

[5] R. Hamada, I. Ide, S. Sakai, and H. Tanaka. Associating cooking video with related textbook. In *Proc. ACM Multimedia 2000 Workshops*, pages 237–241, Nov. 2001.

[6] R. Lienhart, S. Pfeiffer, and W. Effelsberg. Video abstracting. *Commun. ACM*, 40:55–62, 1997.

[7] N. Babaguchi, Y. Kawai, and T. Kitahashi. Generation of personalized abstract of sports video. In *Proc. ICME2001*, pages 800–803, Aug. 2001.

[8] R. Hamada, S. Satoh, S. Sakai, and H. Tanaka. Detection of important segments in cooking videos. In *Proc. IEEE Workshop on CBAIVL 2001*, pages 118–123, Dec. 2001.

[9] Y. Gong, X. Liu, and W. Hua. Creating motion video summaries with partial audio-visual alignment. In *Proc. ICME2002*, pages 285–288, Aug. 2002.