Detection of Important Segments in Cooking Videos

Reiko HAMADA[†], Shin'ichi SATOH[*], Shuichi SAKAI[‡], and Hidehiko TANAKA[‡]

[†]Graduate School of Engineering, The University of Tokyo
[*]National Institute of Informatics
[‡]Graduate School of Information Science and Technology, The University of Tokyo

Abstract

This paper presents a method to delineate visually important segments from cooking videos, which is expected to be a core technique for cooking video indexing. Visual information is exceptionally important in cooking videos, since it represents essential technical skills owing to properties of instruction type videos. Here, we reveal that ordinary shot-based video indexing approach will not work for cooking videos, since visually important segments are absorbed in a shot. Based on the observation of cooking videos, the method detects repetitious motion segments as visually important segments by using relatively simple and robust techniques. The experimental evaluation shows that the method detects more than 80% of manually selected segments, while 84% of detected segments were correct. We also developed a cooking video abstraction system as a sample application of our method and confirmed that the proposed method is useful for a real application.

1. Introduction

Following the increase of computational power and storage capacity, a large amount of TV broadcast video is transmitted and stored everyday. Recently, practical use of these information has become an important subject. Now, many studies are made on indexing broadcast video and multimedia data. Following this trend, we picked up the cooking video as a target of video indexing by analysis[3].

In many cases, supplementary documents are available along with cooking video. But, unlike other types of video, such as news, visual information is exceptionally important in cooking videos. In news video indexing, for example, text information such as closed captions play a major role to realize news-on-demand system, whereas visual information is used only for topic segmentation in most cases. On the other hand, although cooking videos might also have related text information such as closed captions and corresponding cookbooks, visual information in cooking videos is extremely important because it represents essential technical tips or skills of cooking procedures, due to properties of instruction type videos.

Taking advantage of these visual information, automated cooking video indexing will enable many practical applications, such as, cooking video retrieval system, cooking video abstraction, or construction of multimedia cooking contents through integration with cookbook. In the future, the demand for such indexed cooking video will increase in proportion to the popularity of computers at home. For example, visual recipe on demand, multimedia cooking tips catalogue, or visual cooking courseware will be very effective applications. In addition, applications in kitchen environment, such as smart kitchen or automated cooking system, would be examined using these digitalized cooking recipes.

Typical approaches to video indexing generally separate given videos into segments, *i.e.* shots. Shot boundaries (cuts) are the discontinuous points in an image sequence, so they can be detected in relatively good accuracy. Then, in shot-based video indexing, image, audio and text analyses is applied to the videos to realize content-based access to shots[1, 2].

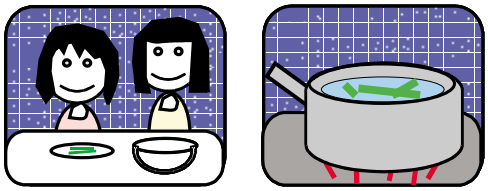
Cooking videos, on the other hand, have peculiar features different from other videos. We will show that general video indexing methods which segment videos into shots as the smallest units for indexing are not suitable for cooking video indexing. We will discuss about features of cooking videos and introduce our method in the following sections.

2. Features of Cooking Videos

In this section, features of cooking videos are revealed and indexing method suitable for cooking videos is dis-

cussed.

As shown in Fig. 1, shots in a cooking video are categorized into (1)Face Shot, and (2)Hand Shot.



(a)Face Shot (b)Hand Shot

Figure 1. Shots categories in cooking video.

In a face shot, almost the whole kitchen is shown. Though a teacher or an assistant explains about cooking procedure, their motion and foods are too small in the picture. So, it is hard to obtain visual knowledge from face shots.

On the other hand, hand shots are close-ups of tools and/or hands while cooking something, and they have rich visual information. However, there are smaller segments in each hand shot, such as, image of preparation steps, cooking motion(s), and appearance of foods. A hand shot contains important segments which are the keys of the recipe, and redundant segments between them as well.

Typical structure of a cooking program is shown in Fig. 2. Face shots and hand shots appear almost alternately. Each hand shot has smaller structure, *i.e.* it contains smaller important segments and relatively redundant segments.

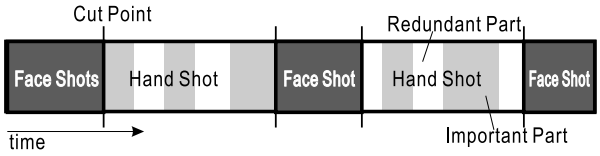


Figure 2. Structure in cooking video.

Due to the structure of cooking videos, shots are too coarse units for indexing; *i.e.* face shot is not visually significant, and hand shot contains both significant segments and redundant segments. Considering this structure, it is crucial to delineate these important segments from hand shots. So, shot-based methods are not suitable for cooking video indexing.

In shot-based video indexing, shot detection is realized by simple signal processing-based techniques, since shot boundaries are relatively abrupt changes in signal. However, important segments in hand shots do not have obvious discontinuities as their boundaries. Although shot-based video indexing is based on signal processing techniques, video content analysis is essential for segmenting and indexing cooking videos.

A cooking program is a kind of an instruction video, and in most cases, there is a supplementary cookbook describing the same recipe as the video. Since users can always refer to the recipe described in textbooks, unimportant steps in the recipe are sometimes omitted in the video. However, cooking video is useful because it can provide visual information that is essential to instruct cooking procedure that text can not express.

So, cooking motions and appearances of foods are particularly important in cooking video. Among them, we have examined cooking motions referring to the actual cooking programs, and we found that most motions in the cooking recipes are repetitious motions. Repetitious motions have many corresponding verbs, namely, “cut”, “toss”, “mix”, “whip”, “grind”, and so on. Besides, there are some verbs that can not be considered as repetitious motions, but are repeated in actual cooking videos in most cases. It is because people tend to repeat important motion for emphasis. Examples of repetitious motions are shown in Fig. 3. Based on this idea, we detect important segments using time periodicity of repetitious motions.

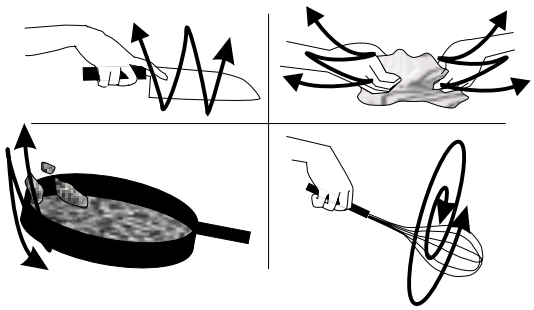


Figure 3. Examples of repetitious motions.

3. Detection of Repetitious Motion

In this section, our method to detect repetitious motions is described.

Many studies on gesture recognition use skin color to detect hand regions. After that, hands are tracked and their motion is recognized[5, 6].

However, in cooking videos, hands are not always in the picture, even in hand shots. Then, we need to detect repetitious motion of tools, such as spatula or chopsticks. Since these tools have various color and shape, it is difficult to identify their graphical features. Accordingly, methods using specific color such as skin color is not suitable for our target.

As shown in Fig. 3, we assume that an object goes back and forth in a small region in the picture of important repetitious motion segments. In our method, we analyze time transition of brightness in small regions using time frequency analysis, and detect repetitious motion by testing existence of periodicity.

First, a frame is divided into small blocks, each of them consist of 3×3 pixels (see Fig. 4). $V_{x,y}(t)$ is the mean brightness of every pixel in a block, where x, y are the spatial coordinates of the block and t is the time coordinate of the frame.

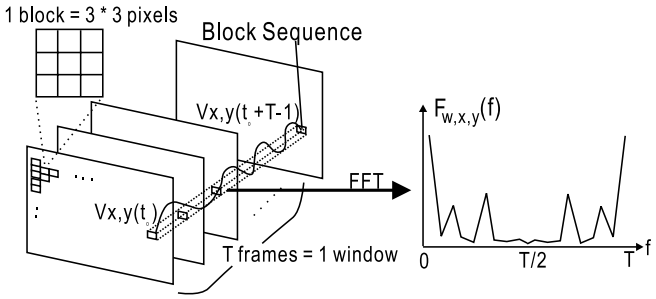


Figure 4. Image division and FFT method application.

As shown in Fig. 4, if x and y are fixed and t varies, $V_{x,y}(t)$ shows the time transition of brightness at x, y. Especially, if the object moves back and forth in the block, $V_{x,y}(t)$ changes periodically. If there is such periodical transition at a certain number of blocks, the method regards that a periodical motion exists in the segment at that time.

We applied FFT(Fast Fourier Transform) to $V_{x,y}(t)$ at every x, y, to examine the periodicity. The size of each window for FFT is $T = 2^n$ frames. We move this window by T_{step} frames and examine whether periodical motion exists or not at each point of time. Equation of FFT is shown as follows:

$$F_{w,x,y}(f) = \left| \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} V_{x,y}(w \times T_{step} + t) \cdot W^{ft} \right|^2 \quad (1)$$

where W^{ft} is $e^{-j\frac{2\pi}{T}}$, and $F_{w,x,y}(f) = F(f)$ is the power at frequency f. w is the window number, which shows the time zone that the FFT was applied. $w \times T_{step}$ is the time offset of the window.

If $V_{x,y}(t)$ is periodical, the graph of FFT result should have a clear peak at a particular f. To extract such a peak, we introduce some statistical criteria. We observed that the target frequency is limited in proportion to the human motions. Suppose that the range of frequency is $f_0 \leq f < f_0 + N$ (see Fig. 5).

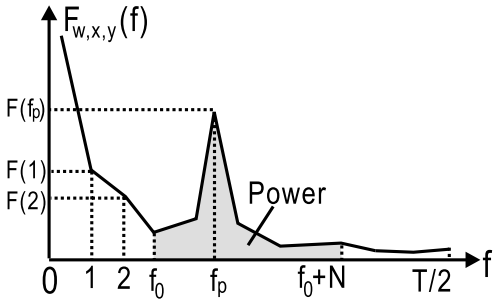


Figure 5. FFT graph.

First, the sum of power within the target frequency is given by:

$$Power_{w,x,y} = \sum_{f=f_0}^{f_0+N} F(f) \quad (2)$$

Next, the maximum value of $F(f)$ is given by Eq. 3. The frequency that gives the maximum is defined as f_p.

$$F(f_p) = \max_{f_0 \leq f < f_0+N} F(f) \quad (3)$$

Now $F(f_p)$ is the maximum but not always the clear peak of the graph. So, we define $F_{peak}(f_p)$, that shows how $F(f_p)$ stands out from the average. This is the ratio of $F(f_p)$ to the average of $F(f)$ excluding $F(f_p)$ (Eq. 4).

$$F_{peak}(w, x, y) = \frac{F(f_p) \times (N-1)}{\sum_{f=f_0, f \neq f_p}^{f_0+N-1} F(f)} \quad (4)$$

If $F(f_p)$ is smaller than the power at low frequencies, namely, $F(1)$ and $F(2)$, it is probably due to slow motion which is not periodical. To exclude such cases, R_1 and R_2 are defined, which are the ratio of $F(f_p)$ to $F(1)$ and $F(2)$ respectively. R_1 and R_2 must be large enough for periodical motion.

When the periodicity of the motion is vague, the total power becomes large but the peak in the graph does not become so clear. So, we define the index of sharpness of the peak as shown in Eq. 5.

$$R_{sharp} = \frac{F(f_p) \times 4}{\sum_{f=f_p-2, f \neq f_p}^{f_p+2} F(f)} \quad (5)$$

In the above, we introduced six parameters: $Power$, F_{peak}, f_p, R_1, R_2 and R_{sharp}. They express the total power and sharpness of the peak in the FFT graph. In this method, existence of periodicity in the motion is examined using these six parameters.

To put it concretely, the repetitious motion is detected in the window w when every 6 parameters exceed thresholds in more than 2 blocks. For example, if repetitious motion

is detected at $w = w_0$, the motion exists within the frame range of $T_{step} \times w_0 \leq t < T_{step} \times w_0 + T$.

If there is only one block that satisfies the above conditions, the segment is not regarded as a repetitious motion because it is very likely due to noise. In addition, pictures are very noisy at the borders of the frame. So we excluded 15 pixels (5 blocks) from each edge from consideration.

4. Evaluation Experiment

4.1. Conditions

We made an evaluation experiment according to the previous section. The target data was cooking video (34 recipes, about 160 minutes in total) from a Japanese cooking program. Properties of the video is shown in Tab. 1.

Table 1. Properties of the target video.

total time	160.1 min
number of recipe	34
file format	mpeg2
size	720 × 480 pixels
frame rate	15 frame/sec.

The method detects repetitious motions that repeat more than two times in a window. Then we empirically determined 32 frames (about 2 seconds) for the window size T. The window is moved by $T_{step} = 16$ frames (about 1 second). The limits of target frequency are $f_0 = 3$, $N = 12$. Parameters on frequency analysis are given in Tab. 2(a).

As mentioned in the previous section, when six parameters shown in Tab. 2(b) exceed the thresholds at more than 2 blocks, the method regards the segment as repetitious motion.

Table 2. Values and thresholds in the evaluation experiment.

(a)Values

$T = 32frame$
$T_{step} = 16frame$
$f_0 = 3$
$N = 12$

(b)Thresholds

$Power > 500$	$F_{peak} > 50$
$f_p > 5$	$R_{sharp} > 3$
$R_1 > 3$	$R_2 > 3$

4.2. Result and Analysis

Correct answers were given by detecting repetitious motion from the target video manually. The result of our method was compared with these correct answers, and the accuracy was evaluated.

The result is shown in Tab. 3. The number of manually detected segments is Ans_H, the number of automatically detected segments is Ans_M and the number of commonly detected segments is Ans_C. Recall is Ans_C/Ans_H, and precision is Ans_C/Ans_M.

Table 3. The result of evaluation experiment.

Ans_H	Ans_M	Ans_C	Recall	Precision
62	59	50	80.6%	84.7%

As shown in Tab. 3, the number of false alarms is small, so our method can detect repetitious motions at more than 84% of precision. On the other hand, relatively large number of correct motions were not detected, and recall is about 80%. Examples of results of our method is shown in Fig. 6.

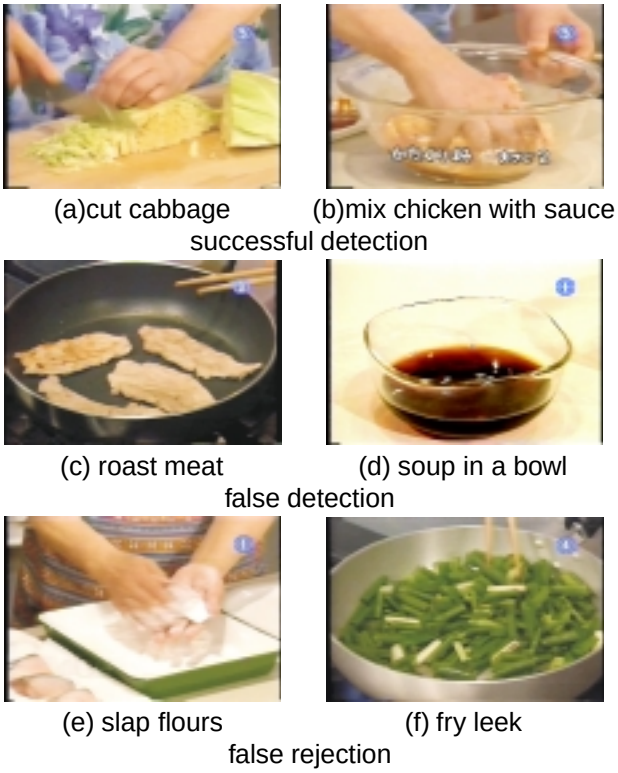


Figure 6. Examples of results: detection of repetitious motions by our method.

First, Fig. 6(a), (b) show the typical examples of successful detection. Each of them is fast and periodical enough, and important motions which represent important tips of cooking procedures. Examples of false alarms are shown

in Fig. 6(c), (d). There are no human motions in Fig. 6(c), (d), but in Fig. 6(c), the periodical motion of chopsticks swaying on the frying pan was detected. In Fig. 6(d), soup in the bowl might have been stirred just before this shot, so soup is swaying periodically.

Fig. 6(e), (f) shows examples of false rejections. In Fig. 6(e), a woman tries to beat flour off a fish. The motion is periodical, but too small. In Fig. 6(f), leek sticks are slowly fried in a frying pan. Not only in this case, slow motions tend to be relatively irregular, so sometimes whether these motions are repetitious motions or not was difficult to detect even for humans. Consequently, the result was unstable when the motion seemed obscure.

It depends on applications whether or not the detected results are useful. Our observation shows that the results seem to fit with perception of humans. To show usefulness of the results, we will provide an actual example of application in the next section.

5. Application of Our Method

We developed a simple application to confirm our method's effectiveness. Generally, abstraction of the video is not so effective because synchronization of audio and image is often lost and the splitted audio is too choppy[4]. However, audio in cooking videos can easily be compensated by recipe texts. In addition, since cooking videos include visually important segments representing tips of cooking procedures as described before, abstract cooking videos can be composed with such segments.

Figure 7. Cooking video abstraction system.

A snapshot of the cooking video abstraction system is shown in Fig. 7.

In this case, the system applies cut detection to given cooking videos first, then hand shots are selected by excluding face shots. In doing this, face regions are detected based on skin color information. Details are omitted due to page limit, but a similar approach can be found in [7].

The system includes the motion segments as well as the first and last segments in the resultant abstract video for a hand shot with motion segments. The middle of the shot is included for a hand shot without motion segment.

We observed that the resultant abstracted cooking videos will surely provide information on important visual tips as well as necessary steps of cooking procedures.

6 Future Works

First, the accuracy of our method should be improved. The current system can hardly detect slow motion, so adaptively stretching analysis window for slow motion segments should be considered. It might be effective to combine our method with object detection and tracking.

Next, important motions which are not repetitious should also be detected for more comprehensive cooking video indexing. Frequency of these motions is relatively small, but sometimes there are non-repetitious motions peculiar to each recipe. These motions have no specific features in common, so it is not easy to distinguish them from unimportant motions such as "put the dish on the table". However, important motions are supposed to be shown in the middle of the picture for a relatively long time. In this case, location and velocity of the center of the motion balance are useful.

Appearances of foods are important information for some motions such as "dish up". In cooking video, sometimes appearances of foods are shown in still images before or after the cooking actions. If those segments are extracted and put together with the important motions, it turns out that almost all important segments are extracted.

The abstraction system introduced in this work can be improved by detecting such important segments. In addition, finer indexing to the video would be realized by integrating the cookbook, in order to achieve a database of digital cooking recipes. By doing this, fine grain cooking video retrieval can be made; *i.e.* retrieving not only each recipe but also each step or motion.

In the future, these results can be applied to the actual cooking environment. For example, we can realize a "smart kitchen," equipped with several sensors to trace the current step within the cooking procedure, which can instruct the user what to do next or how the food will appear.

7. Conclusion

The paper proposes a novel approach for cooking video indexing. We revealed that shot-based methods do not fit to the structure of cooking video, and we decided to employ important segments in a shot for indexing.

In this paper, repetitious motions are regarded as typical important segments, and the automatic detection method for

them was proposed. The effectiveness of our method was shown by the experiments.

We constructed a cooking video abstraction system as an example application to show that our approach is practical for actual applications.

References

[1] Y. Ariki, "Multimedia Technologies for Structuring and Retrieval of TV News," New Generation Computing, Vol.18, No.4, pp.341-358, 2000.

[2] A. G. Hauptmann and H. D. Wactlar, "Indexing and Search of Multimodal Information," Proc. ICASSP'97, April 1997.

[3] R.Hamada, I. Ide, S. Sakai, and H. Tanaka, "Associating Cooking Video with Related Textbook", Proc. ACM Multimedia 2000 Workshops, pp.237-241, Nov. 2000.

[4] M. Christel, M. Smith, C. Taylor, and D. Winkler, "Evolving Video Skims into Useful Multimedia Abstractions," Proc. of ACM CHI'98 Conference on Human Factors in Computing System, April 1998.

[5] T. Nishimura, T. Mukai, and R. Oka, "Spotting Recognition of Gestures Performed by People from a Single Time-varying Image," Proc. of HCI International '97, pp. 33, Aug 1997.

[6] C. Wren, B. Clarkson, and A. Pentland, "Understanding Purposeful Human Motion," IEEE Intl. Conf. on Automatic Face and Gesture Recognition, April 2000.

[7] C. Wren, A. Azarbayajani, T. Darrell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," IEEE Trans. PAMI, Vol. 18, No.7, pp. 780-785, July 1997.