Video CooKing: Towards the Synthesis of Multimedia Cooking Recipes

Keisuke Doman[1], Cheng Ying Kuai[1,⋆],
Tomokazu Takahashi[2], Ichiro Ide[1,3], and Hiroshi Murase[1]

[1] Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
{kdoman,kuai,ide,murase}@murase.m.is.nagoya-u.ac.jp
[2] Faculty of Economics and Information, Gifu Shotoku Gakuen University
1-38 Naka-Uzura, Gifu 500-8288, Japan
ttakahashi@gifu.shotoku.ac.jp
[3] Institute for Informatics, University of Amsterdam
Science Park 904, 1098 XH Amsterdam, The Netherlands
I.Ide@uva.nl

Abstract. In this paper, we propose the concept of synthesizing a multimedia cooking recipe from a text recipe and a database composed of video clips depicting cooking operations. A multimedia cooking recipe is a cooking recipe where each cooking operation is associated with a corresponding video clip depicting it, aimed to facilitate the understanding of cooking operations. In order to synthesize such a multimedia cooking recipe from an arbitrary text-based cooking recipe, a large number of video clips describing various cooking operations should be prepared in the database. Thus, we propose a method to build a database composed of video clips depicting cooking operations, which detects and classifies cooking operations in the cook shows, and tags them with cooking operations in a corresponding cooking recipe text. We also introduce a prototype multimedia cooking recipe interface named "Video CooKing" to demonstrate our concept.

Keywords: cooking recipe, visualization, cooking video, cooking operation, motion analysis, automatic tagging.

1 Introduction

Recently, the number of cooking recipe texts posted on the Web is increasing. For example, "Cookpad"[1] is a recipe-based social networking service where users can post original recipes and also report results and comments. It is so popular that it is said that one fourth of Japanese women in their thirties accesses this service. However, most of the cooking recipes on the Web are text-based and do not have enough explanations about cooking terms, especially cooking operations.

⋆ Currently at Brother Industries, Ltd., Japan.
[1] COOKPAD Inc., "COOKPAD," http://cookpad.com/

K.-T. Lee et al. (Eds.): MMM 2011, Part II, LNCS 6524, pp. 135–145, 2011.

Although some cooking recipes may have image-based explanations, they are not always sufficient for the understanding of some cooking operations.

Therefore, in this paper, we propose the concept of synthesizing a multimedia cooking recipe from a text-based recipe, and also introduce a prototype interface, named "Video CooKing". As shown in Fig. 1, a multimedia cooking recipe is a cooking recipe where each cooking operation is associated with a corresponding video clip depicting it. Compared with an existing text-based cooking recipe, the multimedia cooking recipe makes each cooking operation more understandable with a visual explanation. In fact, a cooking assistance software for a portable game machine "Nintendo DS", that provides users with multimedia explanations of cooking recipes, is already commercially available[2]. However, to create such a software, visual explanations should be prepared manually in advance. We consider that it is unrealistic to manually prepare numerous video clips depicting cooking operations on various ingredients.

To solve this problem, we are aiming at obtaining such video clips from cook shows automatically. A cook show contains video clips depicting each cooking operation in its corresponding cooking recipe, and is broadcast with closed-captions. We obtain a set of tagged video clips depicting cooking operations from many cook shows, and consequently build a database with them. Once such a database is built, we can apply the method proposed in this paper to cooking recipes without corresponding cook shows, for example cooking recipes on the Web.

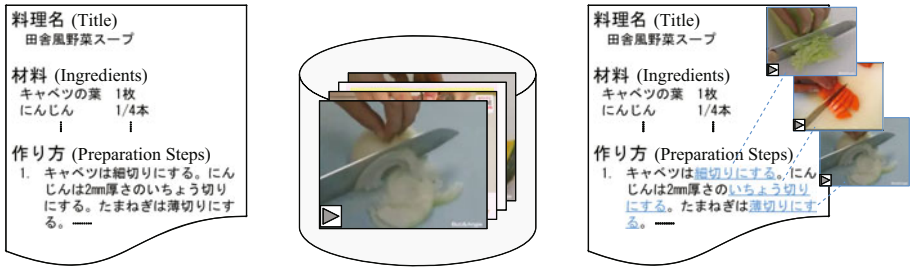
Meanwhile, Hamada et al. proposed the "Cooking Navi" system that analyzes the dependency structure in a cooking recipe text [2], aligns each step to a video segment obtained from a corresponding cook show [3], and also presents the steps along the dependency structure with the aligned video while a user cooks [4]. However, this system can generate a multimedia cooking recipe only when the cooking recipe has a corresponding cook show.

This paper is organized as follows. The next section illustrates the method for associating a cooking recipe text with its corresponding cook show to build a database composed of video clips depicting cooking operations. Section 3 introduces a prototype multimedia cooking recipe interface named "Video CooKing". The paper ends with a summary and discussion of future works in Section 4.

2 Composing a Database of Video Clips Depicting Cooking Operations

As shown in Fig. 2, the flow of associating a cooking recipe text with a cook show is composed of mainly three processes: 1) text processing, 2) image processing and 3) integration. The text processing part analyzes the cooking recipe text and the closed-captions (CC) to extract tags (a pair of an "ingredient" and a

[2] Nintendo Co., Ltd., "It talks! DS Cooking Navi (in Japanese)," http://www.nintendo.co.jp/ds/a4vj/



(a) Text-based cooking recipe (b) Video clips of cooking operations (c) Multimedia cooking recipe

Fig. 1. Synthesis of a multimedia cooking recipe ((a) + (b) → (c))

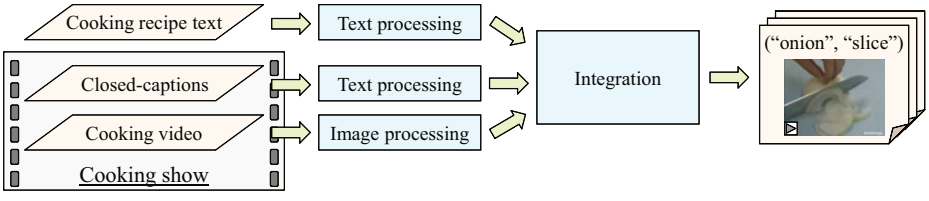


Fig. 2. Process flow for tagging a video clip depicting cooking operations from a cook show

“cooking operation”). The image processing part extracts video clips depicting cooking operations from the cook show, and then classifies them. The integration part associates the tags with the video clips depicting cooking operations.

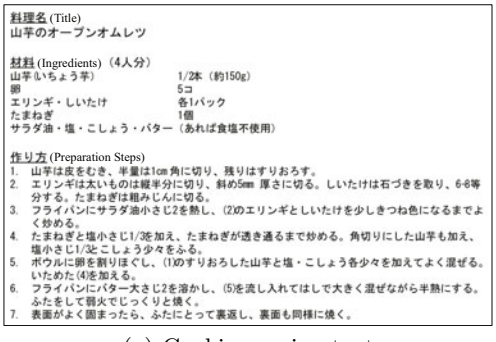
Although the following explanation is based on the processing of Japanese cooking recipe text, it should be possible to be applied to other languages by a similar approach. Accordingly, some language-specific details are omitted in the following explanations for simplicity.

2.1 Text Processing: Extracting Cooking Operations and Corresponding Ingredients

In the text processing part, pairs of an “ingredient” and a “cooking operation” are extracted as tags for video clips depicting cooking operations from a cooking recipe text and CCs. The structure of a cooking recipe text and CCs is shown in Fig. 3. Generally, a cooking recipe is composed of “Ingredients” where all ingredients are listed, and “Preparation Steps” where cooking procedures are described as a numbered list. Meanwhile, CCs contain the transcript of the speech and the timing of its utterance in the audio track. The details of the process are described below.

Analyzing the cooking recipe text and the CC. First, morphological analysis is applied to each sentence in "Ingredients" and "Preparation Steps" of the input cooking recipe text. Next, nouns (ingredients) that commonly appear in both "Ingredients" and "Preparation Steps", together with verbs (cooking operations) are extracted from "Preparation Steps", respectively. Then, considering the grammar, each cooking operation is associated with its target ingredients in "Preparation Steps". As a result, pairs of an "ingredient" and a "cooking operation" are obtained from the cooking recipe text.

Similarly, pairs of an "ingredient" and a "cooking operation" are obtained from the CC, too.



料理名 (Title)
山芋のオーブンオムレツ

材料 (Ingredients) (4人分)

山芋（いちょう芋）	1/2本（約150g）
卵	5コ
エリンギ・しいたけ	各1パック
たまねぎ	1個
サラダ油・塩・こしょう・バター（あれば食塩不使用）	

作り方 (Preparation Steps)
1. 山芋は皮をむき、半量は1cm角に切り、残りはすりおろす。
2. エリンギは太いものは縦半分に切り、斜め5mm 厚さに切る。しいたけは石づきを取り、6-8等分する。たまねぎは粗みじんに切る。
3. フライパンにサラダ油小さじ2を熱し、(2)のエリンギとしいたけを少しきつね色になるまでよく炒める。
4. たまねぎと塩小さじ1/3を加え、たまねぎが透き通るまで炒める。角切りにした山芋も加え、塩小さじ1/3とこしょう少々をふる。
5. ボウルに卵を割りほぐし、(1)のすりおろした山芋と塩・こしょう各少々を加えてよく混ぜる。いためた(4)を加える。
6. フライパンにバター大さじ2を溶かし、(5)を流し入れてはしで大きく混ぜながら半熟にする。ふたをして弱火でじっくりと焼く。
7. 表面がよく固まったら、ふたにとって裏返し、裏面も同様に焼く。

(a) Cooking recipe text

Time	Speech transcript
[09:30:27]	しゅん「おいもで旬おかず」２日目です.
[09:30:35]	よろしくお願いします.
[09:30:41]	長芋さつまいもとあまりフレンチに使わ
[09:30:49]	気になりますね。まず料理を見てください.
[09:30:59]	山芋をすりおろして卵に入れた角切りを
[09:31:07]	秋らしくきのこもたっぷり入れました.
[09:31:11]	山芋とオムレツの取り合わせ。
[09:31:15]	発想の原点はお好み焼きです.
[09:31:25]	そしてあと２つあります.
[09:31:31]	いつもすりおろして使うから
[09:31:37]	シャキシャキとした歯応えを食べて頂きたく
[09:31:43]	畑のキャビア「とんぶり」なんですよ.
[09:31:47]	かもそしてもう１つ. 鴨肉ですね. そうです.
[09:31:55]	ソースはさつま芋をオレンジで煮たもので
[09:32:01]	四角い物はさつま芋.
[09:32:07]	どんな味か楽しみです. 早速１品目.
[09:32:13]	変わった形をしてますね.
[09:32:17]	山芋にも粘りのある物と無い物があって

(b) Closed-captions

Fig. 3. Structure of a cooking recipe text and closed-captions

2.2 Image Processing: Classifying the Video Clips Depicting Cooking Operations from the Cook Show

In general, as shown in Fig. 4, a "face shot" and a "hand shot" appear alternately in a cook show. The upper body of a person is captured in the face shot, whereas the hand of a person is zoomed-up in the hand shot. It is generally considered that the hand shot is more important for cooking assistance, since the closeup of the current state or a cooking operation is captured in the shot [6]. For this reason, we focus on the hand shot and classify scenes in them into the following three categories considering motion.

- "Repetitious motions": A scene containing motions that repeat several times. It is further classified into the following two categories:
  - "Converged": A scene where the periodic changes of pixel values are observed in a specific area of a frame. (ex. cut)
  - "Distributed": A scene where periodic changes of pixel values are observed in a wide area of a frame. (ex. fry, mix)

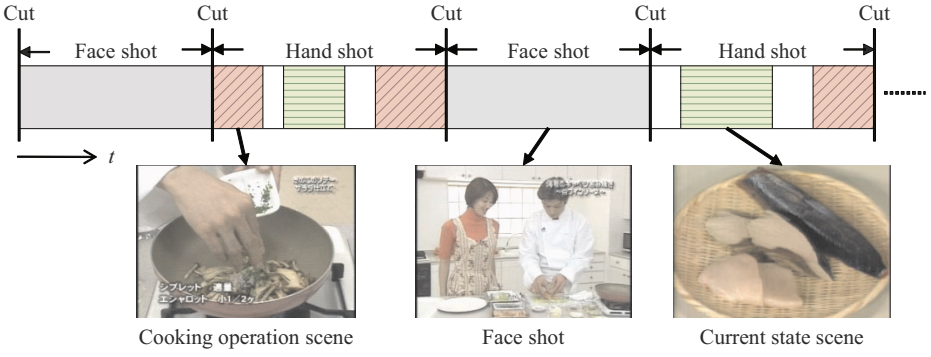


Fig. 4. General structure of a cook show

– "Current state": A scene containing no dynamic motion. (ex. stew, boil, "ingredient (noun)")
– "Other motions": Other than the above. (ex. serve, season)

The details of the classification of the scenes are described next.

Classifying the hand shots. First, the input cook show is segmented into shots, and then only hand shots are extracted from them by the method proposed by Miura et al. [6]. Next, for each segment composed of several continuous frames in a hand shot, an eigenspace is constructed. Then each frame in the segment is projected onto the eigenspace. As the motion feature for the classification, we focus on the trajectory drawn in each eigenspace (Fig. 5 (center column)), especially that drawn on the first eigenaxis (Fig. 5 (right column)). Then, the trajectory of each segment is classified with the following conditions:

$$\begin{cases} \text{"Repetitious motions"} & \text{if} \quad m \geq \theta_m \\ \text{"Current state"} & \text{if} \quad m < \theta_m \ \text{ and } \ \Delta_r \leq \theta_{\Delta_r} \\ \text{"Other motions"} & \text{otherwise} \end{cases} \tag{1}$$

where m is the number of peaks, Δr is the difference of the minimum and the maximum values of the trajectory, and θ_m and $\theta_{\Delta r}$ are the thresholds of m and Δr, respectively. Here, the peak is defined as a point that meets the following conditions:

$$\begin{cases} g(t) - g(t+1) \geq \theta_1, \\ g(t) - g(t-1) \geq \theta_1, \\ g(t+1) - g(t+2) \geq \theta_2, \\ g(t-1) - g(t-2) \geq \theta_2 \end{cases} \tag{2}$$

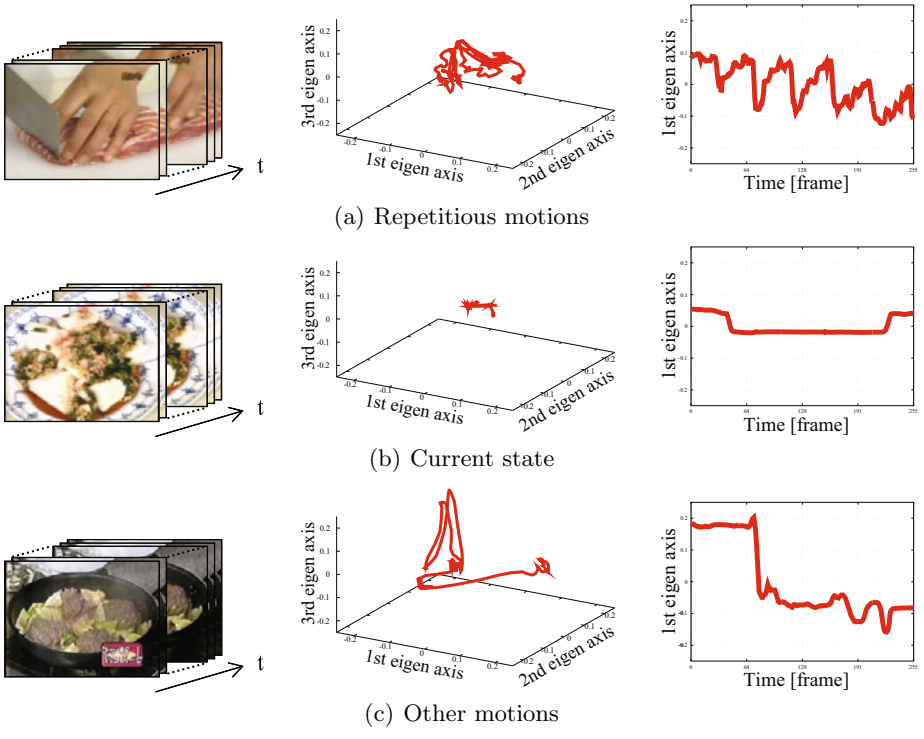


Fig. 5. Example of the analysis results for each motion category. In each motion category, the left column is an input video clip, the center column is its trajectory on the eigenspace, the right column is the trajectory projected onto the first eigenaxis.

where $g(t)$ is the value on the first eigenaxis at time t, θ_1 and θ_2 are the thresholds of the peak strengths. Finally, we regard a series of the segments classified into the same motion category as a scene.

Classifying the repetitious motions. The second-stage of the classification is performed only for "repetitious motions". First, frequency analysis is applied to continuous frames in an input scene, and then, after segmenting the frame into blocks, the temporal change of pixel values is calculated per block. Next, the number of repetitions in each block are counted, where an explicit peak at a certain frequency exists. Example of the results of the repetition count is shown in Fig. 6. Next, regarding each block as a sample point, eigenvalues λ_1, λ_2 are calculated by applying PCA (Principal Component Analysis) to the distribution of repetition counts in a frame. As shown in Fig. 7, there is a clear difference between "converged" and "distributed" in the distribution of the repetition counts.

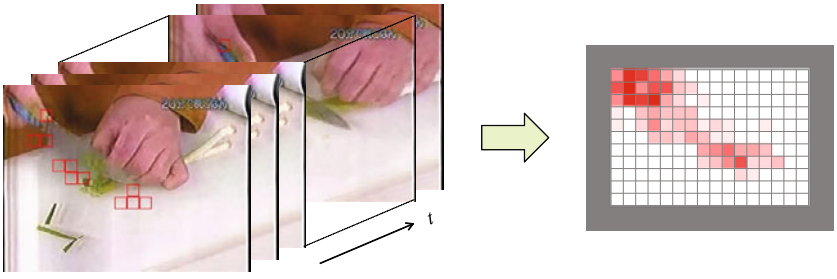
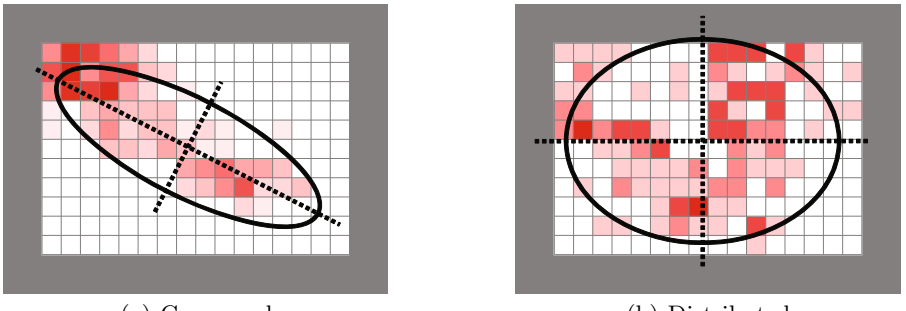


Fig. 6. Example of the frequency analyses. The deeper the color, the larger the count.



(a) Converged (b) Distributed

Fig. 7. Example of the results of PCA applied to the distribution of repetition counts in a frame. The dashed lines represent the first and the second eigenaxes. There is a clear difference between the distribution of repetition counts in "Converged" and "Distributed".

Focusing on this, the input scene is classified further into two motion categories by the following condition:

$$\begin{cases} \text{“Converged”} & \text{if} \quad (\lambda_1 - \lambda_2) \geq \theta_\lambda \\ \text{“Distributed”} & \text{otherwise} \end{cases} \tag{3}$$

where θ_λ is the threshold of the variance on each axis.

2.3 Integration: Tagging the Video Clips Depicting Cooking Operations

Each video clip depicting a cooking operation classified according to the process described in 2.2 is associated with a tag (pair of an "ingredient" and a "cooking operation") obtained according to the process described in 2.1 as follows: First,

the pairs of an "ingredient" and a "cooking operation" that commonly appear in both cooking recipe text and CC are extracted by matching them. Next, each pair is tagged to a video clip depicting a cooking operation according to the time stamp in the CC. Here, only tags that correspond to the motion category of the video clip depicting a cooking operation are selectively tagged.

By applying the method to many cook shows, we will obtain a set of tagged video clips depicting cooking operations, and consequently build a database composed of them.

2.4 Experiment

We evaluated the association performance of the method described above with the following experiment.

Experimental conditions. Eight cooking recipes and corresponding cook shows[3] (320 × 240 pixels, 30 fps and in total 75 min.) were used. A Japanese morphological analyzer MeCab[4] was used to obtain the parts-of-speech of the terms. Cut detection and hand shot extraction from the cook shows were performed manually for this experiment, and the extracted hand shots without corresponging tags were excluded from the classification targets. We manually labeled each scene for the ground-truth. The size of the blocks and the window width for the classifications of "repetitious motions" were set to 16×16 pixels and 256 frames (about 8 seconds), respectively according to the result of a preliminary experiment. We evaluated the tagging accuracy based on two counting rules: 1) a tagging is judged as correct if a pair of an "ingredient" and a "cooking operation" is correctly tagged to a scene (pair-count) or 2) if only a "cooking operation" is correctly tagged (solo-count).

Experimental results. 129 scenes were obtained as the result of the classification of the hand shots, and 135 pairs of an "ingredient" and a "cooking operation" were tagged to them. The tagging accuracy was 52.6% based on the pair-count rule, 68.1% based on the the solo-count rule.

Discussions. In the text processing part, there were many mis-matchings of tags caused by different expressions of a similar cooking operation. We consider that we can cope with this problem by using a thesaurus of cooking terms. In the image processing part, we consider that there are two important points in order to obtain higher classification accuracy: 1) treatment of a scene with a large camerawork which yields a large motion, and 2) choice of the best thresholds for accurate classification of "repetitious motions" and "other motions". In the

[3] NHK Educational Corp., "Today's cooking for everyone (in Japanese)," http://www.kyounoryouri.jp/.

[4] Kyoto University, "Japanese morphological analyzer MeCab," http://mecab.sourceforge.net/

integration part, most of the mis-taggings were caused by the existance of an operation that is not mainly-focused but captured in the frame (Fig. 8(a)), or operations or states that do not involve motion (Fig. 8(b)). As for the former (Fig. 8(a)), although the tagging result was ("water", "boil"), the camerawork was focusing on dissolving scorch in the boiled water using a spoon. As for the latter (Fig. 8(b)), although the tagging result was ("egg", "become hard"), we judged the tagging result as incorrect since such a cooking operation could not be observed as a motion.

However, in view of the complexity of associating a cooking recipe text with a video clip depicting a cooking operation, we consider that the experimental results are sufficient for our purpose at the current stage, since we consider that the user may select what s/he needs from the video clips presented in an interface.

(a) Tagging result: ("water", "boil")

(b) Tagging result: ("egg", "become hard")

Fig. 8. Examples of mis-tagging

3 Video CooKing: A Prototype Multimedia Cooking Recipe Interface

We implemented a prototype interface "Video CooKing" as shown in Fig. 9 to demonstrate the concept of the multimedia cooking recipe. In the interface, each cooking operation in "Preparation Steps" in the left column is linked with its one or more corresponding video clips depicting cooking operations. When a linked cooking operation is clicked, a list of ingredients that are the targets of the cooking operation is shown in the right column. Each ingredient is linked to a video clip depicting a cooking operation if there is more than one corresponding video clips that exist in the database. Users can play / stop the video clip themselves. We consider that this interface that enables an user to browse a multimedia cooking recipe, facilitates the understanding of cooking operations.

Fig. 9. Video CooKing: a prototype multimedia cooking recipe interface. The original cooking recipe on the Web[6] is shown in the left column, and video clips depicting cooking operations corresponding to each cooking operation in the recipe text is shown in the right column when clicked.

4 Conclusion

In this paper, we proposed the concept of synthesizing a multimedia cooking recipe and also introuced a prototype interface. Experimental results showed the effectiveness of our method in tagging a pair of an "ingredient" and a "cooking operation" to a video clip obtained from a cook show, and consequently, the capability of building a database composed of video clips depicting cooking operations. Future work includes the improvement of the interface and the tagging [5].

Acknowledgments. Parts of this work were supported by the Grants-in-Aid for Scientific Research (21013022) from the Japanese Ministry of Education, Culture, Sports, Science and Technology. The "Media Integration Standard Toolkit"[5] was used for the implementation of the system, and some thumbnail images from the "Video database for evaluating video processing" [1] was used for explanations in this paper.

References

1. Babaguchi, N., Etoh, M., Satoh, S., Adachi, J., Akutsu, A., Ariki, Y., Echigo, T., Shibata, M., Zen, H., Nakamura, Y., Minoh, M., Matsuyama, T.: Video database for evaluating video processing (in Japanese). Tech. Rep. IEICE (PRMU2002-30) (June 2002)
2. Hamada, R., Ide, I., Sakai, S., Tanaka, H.: Structural analysis of preparation steps on supplementary documents of cultural TV programs. In: Proc. Fourth Int. Workshop on Information Retrieval with Asian Languages (IRAL 1999), Taipei, Taiwan, pp. 43–47 (November 1999)
3. Hamada, R., Miura, K., Ide, I., Satoh, S., Sakai, S., Tanaka, H.: Multimedia integration for cooking video indexing. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004. LNCS, vol. 3332, pp. 657–664. Springer, Heidelberg (2004)
4. Hamada, R., Okabe, J., Ide, I., Satoh, S., Sakai, S., Tanaka, H.: Cooking Navi: Assistant for daily cooking in kitchen. In: Proc. Thirteenth ACM Int. Multimedia Conf. (ACMMM 2005), Singapore, pp. 371–374 (November 2005)
5. Ide, I., Shidochi, Y., Nakamura, Y., Deguchi, D., Takahashi, T., Murase, H.: Multimedia supplementation to a cooking recipe text for facilitating its understanding to inexperienced users. In: The Second Workshop on Multimedia for Cooking and Eating Activities (CEA 2010) (December 2010)
6. Miura, K., Hamada, R., Ide, I., Sakai, S., Tanaka, H.: Motion based automatic abstraction of cooking videos. In: Proc. ACM Multimedia 2002 Workshop on Multimedia Information Retrieva, lMIR 2002 (December 2002)

[5] Nagoya University, "Media Integration Standard Toolkit: MIST," http://mist.murase.m.is.nagoya-u.ac.jp/