Associating Cooking Video Segments with Preparation Steps

Koichi Miura[1], Reiko Hamada[1], Ichiro Ide[2], Shuichi Sakai[1], and Hidehiko Tanaka[1]

[1] The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
{miura,reiko,sakai,tanaka}@mtl.t.u-tokyo.ac.jp
[2] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
ide@nii.ac.jp

Abstract. We are trying to integrate television cooking videos with corresponding cookbooks. The cookbook has the advantage of the capability to easily browse through a cooking procedure, but understanding of actual cooking operations through written explanation is difficult. On the other hand, a video contains visual information that text cannot express sufficiently, but it lacks the ease to randomly browse through the procedures. We expect that their integration in the form of linking preparation steps (text) in a cookbook and video segments should result in complementing the drawbacks in each media. In this work, we propose a method to associate video segments with preparation steps in a supplementary cookbook by combining video structure analysis and text-based keyword matching. The result of an experiment showed high accuracy in association per video segments, i.e. annotating the video.

1 Introduction

1.1 Background

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. Generally, multimedia data consist of image, audio and text. Individual research on analysis of each media has been made, but thorough understanding of multimedia data through single-media processing has shown limitation. To overcome this limitation, integrated processing that mutually supplements the incompleteness of information derived from each media could be a solution. Many attempts have been made to index video by means of multimedia integration, but sufficient accuracy for practical use has not necessarily been achieved since their subjects were too general.

We are trying to integrate television cooking videos with corresponding cookbooks. Such limitation of the target domain should lead to realistic integration

accuracy for practical use using relatively simple technologies. Cooking program is a kind of an instruction video, and in most cases, a supplementary cookbook describing the same recipe is provided. The cookbook has the advantage of the capability to easily browse through a cooking procedure, but understanding of actual cooking operations through written explanation is difficult. On the other hand, a video contains visual information that text cannot express sufficiently, but it lacks the ease to randomly browse through the procedures. We expect that their integration in the form of linking preparation steps (text) in a cookbook and video segments should result in complementing the drawbacks in each media. Moreover, various applications, such as indexing and extracting knowledge, should be possible using the result of the integration.

In this paper, we propose a method to associate video segments and preparation steps in supplementary cookbooks.

1.2 Related Works

Many attempts have been made on integrative multimedia processing for video indexing. In the Informedia project[1], they analyzed mainly news and documentary videos using advanced single media analysis techniques. It achieved significant results, but study on associating videos with external documents is not considered.

As an example of an attempt to associate videos with external documents, there is a work on aligning articles in television news programs and newspapers[2]. They refer to nouns that co-occur in open-captions of news videos and newspaper articles, but contents of the video other than open-caption texts (image and speech transcripts, or closed-captions) are not analyzed.

Another work on synchronization between video and drama scripts[3] also refers to external documents. They extract patterns from each media and synchronize them by DP matching to analyze semantic structures of a video from corresponding documents. The order of scenes in the video is basically synchronous to that in drama scripts, whereas in cooking programs, the order of steps often differ between videos and cookbooks. Moreover, a step may consist of several separate video segments. There may also be omitted steps in the video, or extra video segments. These characteristics make it difficult to employ DP matching for our purpose, thus we have to gather hints from each media and integrate them appropriately.

1.3 System Overview

The outline of the proposed method is shown in Fig. 1. First, the video and the preparation steps in a supplementary cookbook are analyzed. The video is structured by image analysis as described in Section 3, and the preparation steps are analyzed as described in Section 4. Next, the structured video and preparation steps are associated as described in Section 5.

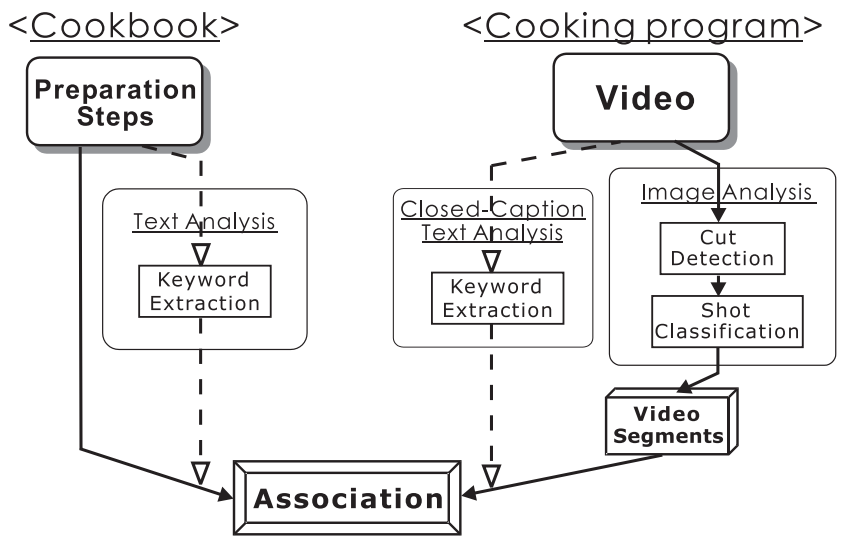

Fig. 1. System overview of associating cooking video with supplementary cookbook.

2 Structural Analysis of Cooking Video

2.1 Image Analysis

Definition of Video Segment First, cut detection is performed to a video stream. Many cooking programs are taken in a studio under good lighting condition, so cut detection is easier than in general video. In this work, we adopt a method that uses DCT clustering[4].

After cut detection, shots are classified. As shown in Fig. 2, shots in cooking videos could be categorized into (a)face shot and (b)hand shot. (a)Face shots are furthermore categorized into (a1)full shot and (a2)bust shot.

Fig. 2. Shot classes in cooking video.

An example of a shot structure of a cooking video is shown in Fig. 3. Here, we observed that over 90% of the head shots of video segments corresponding to beginnings of preparation steps were full shots (a1).

Following this observation, we define a "video segment", which is the minimal unit for the association. Thus, a "video segment" is a sequence of shots that begins with a full shot and ends with a shot before the next full shot, as shown in Fig. 3.

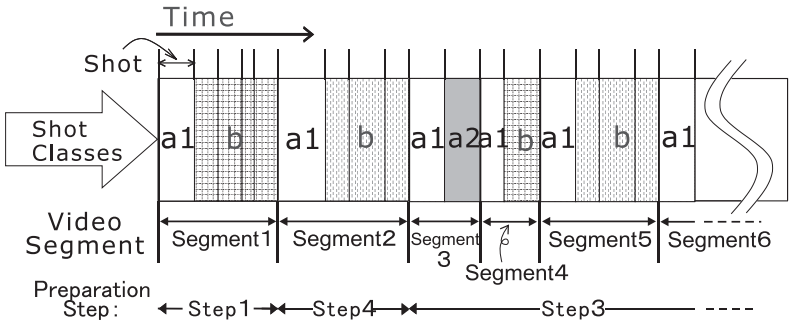

Fig. 3. Example of a shot structure in a cooking video.

Face Shot Detection Detection of face shots is particularly important since they become clues to detect full shots in order to segment the video. Face shots contain significantly large human faces in the images, so they are detected by detecting face regions.

Although various advanced methods to extract face regions exist, in our method, we employ a simple and robust method as follows, since only their existence, locations, and sizes are needed in this work.

1. **Detect skin colored regions:**
   The modified HSV color system[5] is used to detect skin colored regions. V(Value) is used only for excluding dark regions. A certain rectangular region in the H(Hue)-Sm (Modified Saturation) plane is defined as skin color. The distribution of sampled skin-colored regions on the H-Sm plane is shown in Fig. 4. Pixels whose H and Sm drop in the designated rectangular region that circumscribes the distribution, are judged as skin-colored.
2. **Determine the face regions based on certain conditions:**
   The detected skin-colored regions may contain not only faces but also similarly colored objects such as hand, wooden spatula, and table. To exclude them, face regions are determined based on the following conditions. These conditions were defined considering specific features of cooking videos.

– Exclude over-sized or under-sized regions.
– Exclude regions that touch the frame boundary.
– Assume that the sizes of regions are similar when more than one region exist.
– Assume that at least a part of the region is located in the upper-half of the image.

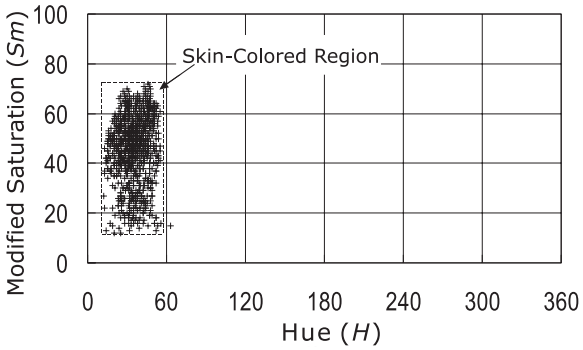

Fig. 4. Skin-colored region on the H-Sm plane[6].

We classified the (a)face shot into (a1)full shot and (a2)bust shot based on the size and the number of face regions.

2.2 Closed-Caption Text Analysis

The contents of audio speech is an important hint for the association. In our method, we use "closed-caption text" that is provided from broadcasting stations as a transcript of speech, instead of performing speech recognition. In case of programs lacking closed-caption texts, speech recognition will be needed in order to obtain transcripts, but sufficient performance will not be expected to be achieved in this case.

Morphological analysis of closed-caption texts is performed in order to extract nouns and verbs. Next, keywords, such as ingredients and verbs, are extracted. Details on keyword extraction are described in section 3 with cookbook analysis.

3 Text Analysis of Cookbook

A recipe in a cookbook consists of a "list of ingredients" and a "preparation steps" part as shown in Fig. 5. "Preparation steps" give explanation on how to cook the "list of ingredients".

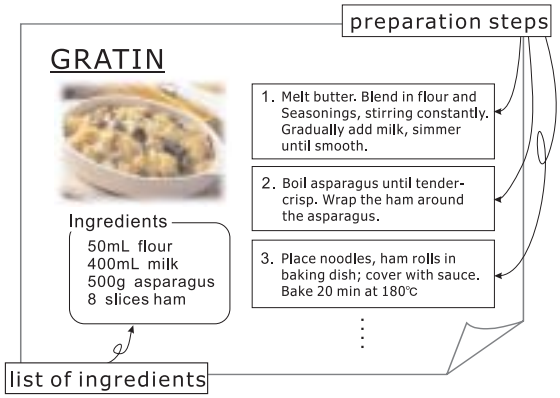

Fig. 5. Example of cooking textbook.

The following procedure is taken to extract keywords from both preparation steps and closed-caption texts.

- Perform morphological analysis.
- Extract ingredient nouns that appear in the "list of ingredients" and all the verbs as keywords.
- Ingredient nouns and all the verbs in a single sentence are regarded as related keywords.

When a step to be analyzed refers to a preceding step as shown in the following example, it is supplemented with all ingredient nouns in the step referred to.

step 1:	Tomato and tuna are cut in square by 2cm, and are mixed in a bowl.
step 2:	Add some olive oil, salt and soy sauce to $\underline{[1]}_{(step\#)}$, and marinate.

Considering verbs, in order to cope with the difference of notations between preparation steps and closed-caption texts, a conceptual dictionary is created and employed. An example of this dictionary is shown in Tab. 1.

4 Association Method

The aim of this work is to associate "video segments" with "preparation steps". In cooking programs, the order of steps are not always synchronous between a video and a cookbook. Moreover, a single step often corresponds to several video segments. On the other hand, there may be omitted steps in the video, or extra video segments.

Table 1. Example of the conceptual dictionary.

words	corresponding cooking motion
slice mince ...	cut
put in pour in ...	add

Considering such irregularities, the following procedure based on the extracted keywords is taken to associate "video segments" with "preparation steps". An example of the association is shown in Fig. 6. The score of a keyword that takes in account of the rareness of the keywords is defined as follows:

$$\frac{1}{M} \times \frac{1}{N} \quad \left(\begin{array}{c} M : \text{Number of steps containing the keyword} \\ N : \text{Number of video segments containing the} \\ \text{keyword in the closed-caption texts} \end{array} \right)$$

1. All keywords in a video segment are compared to the keywords in each preparation step.
2. When both ingredient nouns and related verbs in a segment match those in a preparation step,
   Match with a single step: Associate the video segment with the step.
   Match with several steps: The video segment could belong to several steps. The association is determined referring to the steps that preceding and succeeding video segments are associated to.
3. When no steps match, add the score of the keywords to the steps which has a keyword in common. The video segment is associated with the step with the highest score. If several steps have the highest score, the steps that preceding and succeeding video segments are associated to are referred to.
4. When a video segment could not be associated since the preceding and succeeding video segments are not associated, and thus could not be referred to yet, it will be associated later. The procedure repeats until no more associations could be made.

5 Experiment

5.1 Face Shot Detection

We detected face shots from the first frames of 600 shots (approximately 100 minutes) taken from a Japanese cooking program. The result of face shot detection is shown in Tab. 2. In this experiment, cut detection was manually done to evaluate face shot detection individually, although the result of cut detection applying the DCT clustering method[4] showed over 95% accuracy in a preliminary experiment.

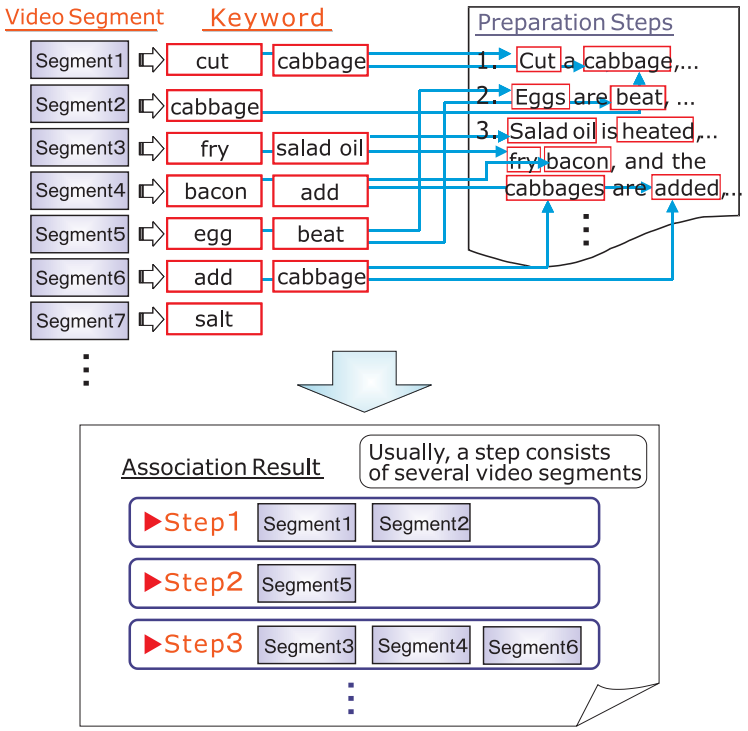

Fig. 6. Example of the association process.

The result of face shot and face region detection is shown in Tab. 2. The number of correctly detected shots is N_C, misdetected shots is N_M, and oversighted shots is N_O. Recall is $N_C/(N_C + N_O)$, and precision is $N_C/(N_C + N_M)$.

The misdetections were mainly due to detecting similarly colored regions such as chicken, meat or wall. The oversights were mainly due to the size of face regions depending on face directions. Although many oversights of face regions are observed in (a1)full shot as shown in Tab. 2, the recall of face shot detection is not much affected, since usually more than two persons exist in a (a1)full shot. Thus, relatively high shot detection rates were achieved despite low recall in face region detection. We consider the result sufficient for actual application to video structuring within the proposed method.

5.2 Associating Video Segments with Preparation Steps

Finally, we made an evaluation experiment that associates video segments with preparation steps. In this experiment, the target cooking videos consist of 20 recipes, with the length of approximately 150 minutes in total, taken from a Japanese cooking program and its supplementary cookbook. Note that shot clas-

Table 2. Result of face shot detection.

Shot class	Correct N_C	Misdetection N_M	Oversight N_O	Recall	Precision
(a1)Full shot	**169**	**24**	**25**	**87%**	**88%**
(face regions)	408	36	162	72%	92%
(a2)Bust shot	**68**	**18**	**20**	**77%**	**79%**
(face regions)	75	22	27	74%	77%

sification was manually done to evaluate the association individually. The result of the association is shown in Tab. 3.

Table 3. Result of the association.

Evaluation	*Total*	*Correct*	*Miss*	*Other*	*Accuracy*
Video Segment $\rightarrow$ *Step*	242	203.5	29.5	9	84.1%
Video Segment $\leftarrow$ *Step*	94	74	20	–	79%
Video Segment $\leftrightarrow$ *Step*	94	59	35	–	62%

"*Video Segment* $\rightarrow$ *Step*" in Tab. 3 shows the association ability per video segment, where a video segment associated with a correct step is regarded as correct. On the other hand, "*Video Segment* $\leftarrow$ *Step*" shows the association ability per preparation step, where a step without lack of associated video segments is regarded as correct. Finally, in "*Video Segment* $\leftrightarrow$ *Step*", a step that has neither much nor little video segments is regarded as correct. Note that when a video segment is associated to two steps, each association is counted as 0.5. *Other* are blocks that were not associated to any step. *Accuracy* is calculated by $Correct/Total$.

As shown in the result of "*Video Segment* $\rightarrow$ *Step*" in Tab. 3, over 80% of video segments were correctly associated with preparation steps. Since the primary aim of the association method is to annotate video segments with preparation steps, this result indicates that the proposed method is effective.

Since the method does not consider temporal order, the association succeeded in most cases when the order of video and preparation steps were different, and also when a step consisted of several separate video segments.

6 Conclusions

This paper proposed and examined a method to associate cooking videos with preparation steps in a supplementary cookbook. A method to associate video segments with preparation steps was proposed, based on a structured cooking video. The experiment showed high accuracy in association per video segments,

i.e. indexing the video. This high accuracy should be considered as the result of using domain specific knowledge.

In the future, closed-caption texts and preparation steps in supplementary cookbooks should be analyzed more precisely in order to improve the association accuracy. Furthermore, we will investigate on a more precise video structure analysis and association method to give videos corresponding to sentences. Various applications, such as indexing, extracting knowledge and constructing a database of cooking operations, should be possible using the result of this work.

Although, some modifications, such as redefining the conceptual dictionary, should be required, the proposed association method may be applied to other educational videos with clear video structures and corresponding textbooks.

Acknowledgments

The sample cooking video images are taken from the "Video Media Database for Evaluation of Video Processing"[7].

References

1. Wactlar, H.D., Hauptmann, A.G., Christel, M.G., Houghton, R.A., Olligschlaeger, A.M.: Complementary video and audio analysis for broadcast news archives. Comm. ACM **45** (2000) 42–47
2. Watanabe, Y., Okada, Y., Tsunoda, T., Nagao, M.: Aligning articles in TV newscasts and newspapers (in Japanese). Journal of JSAI **12** (1997) 921–927
3. Yaginuma, Y., Sakauchi, M.: Content-based retrieval and decomposition of TV drama based on intermedia synchronization. In: First Intl. Conf. on Visual Information Systems. (1996) 165–170
4. Ariki, Y., Saito, Y.: Extraction of TV news articles based on scene cut detection using DCT clustering. In: Proc. Intl. Conf. on Image Processing. (1996) 847–850
5. Matsuhashi, S., Nakamura, O., Minami, T.: Human-face extraction using modified HSV color system and personal identification through facial image based on isodensity maps. In: IEEE Canadian Conf. on Electrical and Computer Engineering '95. (1995) 909–912
6. Ide, I., Yamamoto, K., Tanaka, H.: Automatic video indexing based on shot classification. In: First Intl. Conf. on Advanced Multimedia Content Processing (AMCP '98). (1998) 99–114
7. 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. Report of IEICE, PRMU2002-30 **102** (2002) 69–74