Multimedia Integration for Cooking Video Indexing

Reiko Hamada[1], Koichi Miura[1], Ichiro Ide[2], Shin'ichi Satoh[3], Shuichi Sakai[1], and Hidehiko Tanaka[4]

[1] The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{reiko|miura|sakai}@mtl.t.u-tokyo.ac.jp

[2] Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
ide@is.nagoya-u.ac.jp

[3] National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
satoh@nii.ac.jp

[4] Institute of Information and Security,
2-14-1 Tsuruya-cho, Kanagawa-ku, Yokohama, 221-0835, Japan
tanaka@iisec.ac.jp

Abstract. We have been working on the integration of video with supplementary documents, such as cooking programs. We propose an integration system that performs semantic segmentations of video and text and associates them together. This association is realized using the ordinal restriction of the recipe, cooccurrences of words in the text and the audio in the video, and the relation between the background in a video and words which describe the situation in a text. In this paper, we will introduce the result of an evaluation experiment and show the effectiveness of the proposed integration method. Through our method, many applications should become possible, such as a cooking navigation software.

Keywords: Indexing, Cooking Videos, Association of Video and Text.

1 Introduction

Reflecting the increasing importance of handling multimedia data, many studies are made on indexing to TV broadcast video. Multimedia data consist of image, audio and text, where various studies on analysis of each individual medium have been made. Especially, image processing has been the main medium to handle multimedia data for a long time. But recently, it has started to be considered that image processing alone is insufficient for thorough understanding of multimedia data. From the 1990s, integrated processing that supplements the incompleteness of information from each medium has become a trend [4].

Following this trend, we are trying to integrate TV programs with related documents, taking advantage of the relative easiness of extracting semantic structures from text media. Among various programs, educational programs are considered as appropriate sources, since (1)supplementary documents are available,

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3332, pp. 657–664, 2004.

and (2)the video contains a lot of implicit information that integration could be helpful to thorough understanding of both media. In addition, the demand for cooking videos and their applications are high, since cooking is a daily and important activity.

Therefore we propose an integration system that performs semantic segmentations of video and text and associate them together. This association is realized using the ordinal restriction from the recipe, cooccurrence of words in the text and the audio in the video, and the relation between the background of the video and words which describe the situation in the text.

In this paper, we will introduce the result of an evaluation experiment and show the effectiveness of the proposed integration method. Through our method, many applications should become possible, such as a cooking navigation software.

2 System Overview

In our system, multimedia data is created from a cooking video by matching it to a corresponding part in a recipe text. An example of the matching is shown in Fig. 1.

As shown in Fig. 1, in cooking programs, the order of steps often differs between a video and a textbook. In that case automatic association of a cooking video and its text recipe is a difficult task. In this paper, a solution which combines information derived from multiple media is proposed. The overview of the integration method is shown in Fig. 2.

First, text segmentation is performed by extracting important words from a text using a domain-specific dictionary. The text is divided into semantic segments. This segment is called a "text block". Finally the ordinal structure between text blocks is analyzed.

Meanwhile, shot detection, categorization, and background classification are applied to a video corresponding to the text. Next, shots with a same background type are clustered. This shot cluster is called a "video scene" in this paper.

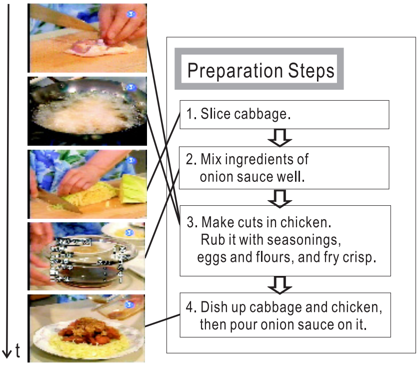

t
Preparation Steps
1. Slice cabbage.
2. Mix ingredients of onion sauce well.
3. Make cuts in chicken. Rub it with seasonings, eggs and flours, and fry crisp.
4. Dish up cabbage and chicken, then pour onion sauce on it.

Fig. 1. Association of a cooking video and a recipe text.

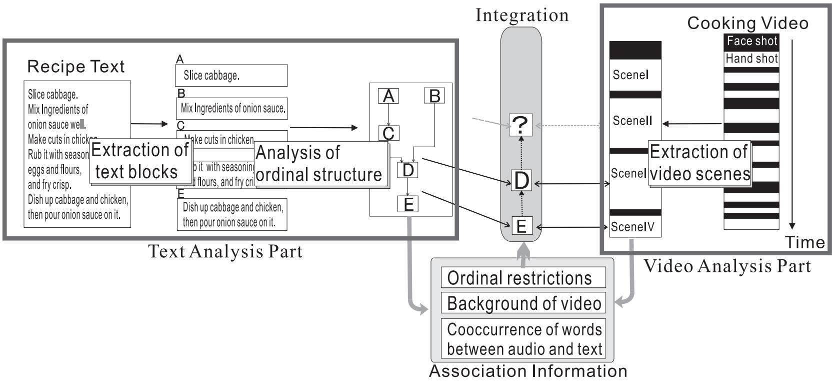


Fig. 2. Overview of the integration method.

Finally, association of each "text block" and "video scene" is performed. At this time, a text block with the highest relevance ratio to a video scene are matched together. The relevance ratio is calculated from the integration of information derived from multiple sources.

Each part of the system is explained in the following sections.

3 Media Analysis

3.1 Extraction of Text Blocks and Analysis of the Ordinal Structure

An overview of the text analysis[1] is shown in the left side of Fig. 2.

First, nouns, verbs and some modifiers are extracted from a cookbook, referring to a domain-specific dictionary. Especially, words which express a cooking condition (ex. "at high heat") are important.

Next, a series of verbs starting with verbs which fulfill the following conditions is extracted as a "text block".

1. Verbs which are associated to the same "ingredient noun".
2. Verbs in a sentence which has "a container noun" + "in".
   (ex. "Bake onions in a frying pan.")

At last, ordinal relations of text blocks are determined. We have already proposed a method to extract the verbal ordinal relation automatically, and have shown its effectiveness[1]. Using this method, ordinal restrictions of text blocks are extracted from the verbal ordinal relations.

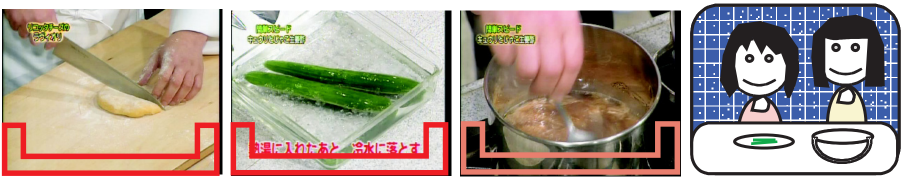
[1] The entire procedure is for Japanese text.

3.2 Extraction of Video Scenes

“Video scenes” are extracted as groups of hand shots that have the same background.

First, cut detection is performed to a video sequence. In our implementation, we adopted a cut detection method using DCT clustering [2]. After the cut detection, the shots are classified into two categories; (1)Hand shot and (2)Face shot, as shown in Fig. 3. Hand and face shots are categorized automatically by face detection as described in our previous publication [3].

Next, hand shots are classified by background color distribution. Backgrounds are categorized into “board”, “table” and “range (=gas stove)” as shown in Fig. 3 (a) ∼ (c), and “others” which are those that can not be categorized into any category.



(a) Board (b) Table (c) Range (gas oven)

(1) Hand shots (2) Face shot

Fig. 3. Shot categories in cooking videos.

In this method, a supervised learning method using multiple cooking programs as training data is performed to extract which part of the images is most probable to represent the “background”. After that, hand shots are clustered using the color information of the “background” part. The “background” part which is extracted by this method is shown in Fig. 3 (a) ∼ (c) as the emphasized block at the bottom. By this process, actual types of the background (ex. table or range) can not be specified, but shots with the same background could be distinguished.

Finally, a continuous series of hand shots with the same background is extracted as a “video scene”.

4 Multimedia Integration

In this section, the integration method which associates a “text block” and a “video scene” is explained.

As shown in Fig. 2, ordinal relations of “text blocks” is structured as an inverted tree. Therefore, when we analyze the restrictions in the order, it is easy to solve them from the latest one to the first one. A text block is associated with a “video scene” with the highest relevance ratio, from the last to the first.

Even for the same combination of a “video scene” and a “text block”, the “relevance ratio” between them may vary according to the order the association was tracked till then. Therefore a text block and a video scene are associated so that it maximizes the total score as a whole.

In this paper, a “relevance ratio” is defined as the total sum of the scores derived from the information listed below.

1. Ordinal restriction of text blocks.
2. Background information of the video scene.
3. The number of the words that cooccuer in the text block and the audio data (closed caption = CC) of the video scene.

In order to adjust the influence of the above three information, scores X_1, X_2 and X_3 which show the degrees of relation are assigned to each of them respectively. The score is distributed to each text block within the assigned score for each information.

The score calculation method to select a text block T_j associated with a video scene $S_{i=I}$ is explained below.

Information 1: Ordinal Restriction of Text Blocks

Here, a score is distributed so that text blocks, which may have more possibility to be associated with a scene $S_{i=I}$ according to the ordinal structure, should have higher score.

The following is an explanation of a score calculation method based on the example in Fig. 4. In Fig. 4, the latest two scenes (IV and III) have already been associated to text blocks, and the next scene (II) is waiting for the association.

The previously associated blocks (D and G) are defined as $T_{i\in\alpha}$. As shown in Fig. 4, we define the nearest blocks which are upper than $T_{j\in\alpha}$ blocks as the first candidates, and the nearest one which are upper than the first candidates as the second candidates. This is because the blocks lower than $T_{j\in\alpha}$ are scarcely possible to be associated to earlier scenes.

Let n_1 be the number of blocks of the first candidates. Then X_1/n_1 is the score of the text of the first candidates. The second candidate blocks get $X_1/n_1 \times n_2$, where n_2 is a suitable ratio $(1 \sim 2)$.

Information 2: Background Information of the Video Scene

At first, background information of each scene can not be used as a hint for the association, because the type of the background is not specified yet. As the association proceeds, text blocks are associated to each video scene, the type of each background class could be inferred.

First, one of the four kinds of attributes in Tab. 1 are given to each word contained in a text block referring to a domain-specific dictionary.

The attribute of a text block is defined as the same as that of the words it contains. Although words with an attribute “o” are usually neglected, when a text block contains only words of attribute “o”, a text block will also have an attribute “o”.

Next, a process is selected among the following A $\sim$ C according to the current state of the association process. (The background class of the target scene $S_{i=I}$ is defined as B_I.)

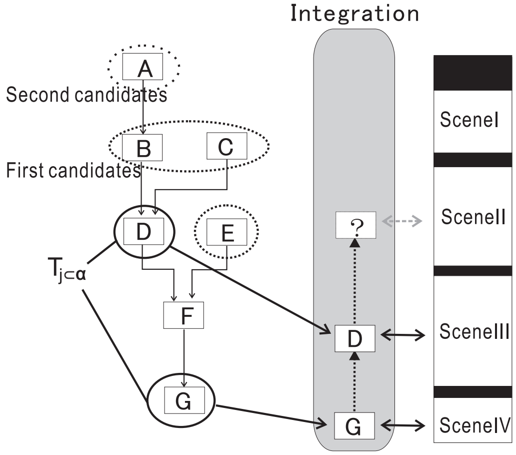


Fig. 4. Extraction of candidate text blocks using ordinal restrictions.

Table 1. Attributes of words related to video backgrounds.

attribute	property	associated background	examples
c	cut	board	cut, slice, knife, cutting board
h	heat	range	bake, heat, frying pan
m	work	table	dish up, mix, bowl, dish
o	others	–	add, chopsticks

A. There has been no scene associated with the text blocks: This is the first condition of the whole association process. A score 0 is given to all text blocks because no hint is available from the background information at this state.

B. Scene with background B_I has been associated with some text block: A score X_{ij} which shows the relevance ratio between a scene S_i and a text block T_j (attribute C_j) is given as in Eq. 1. The number of all text blocks which has been associated to scenes with a background class B_i is defined as n_i, and the number of text blocks with an attribute C_j among them is defined as n_{ij}.

$$X_{ij} = X_2 \times n_{ij}/n_i \tag{1}$$

C. Scenes with backgrounds except B_I is associated to a text block: X_{IJ} is the score between a text block $T_{j=J}$ (attribute $C_{j=J}$) and a video scene S_I. When the number of background types which has not been associated to a text yet (including B_I) is defined as n_B, X_{IJ} is defined as shown in Eq. 2. According to Eq. 2, for example, if many other background types have been associated with an attribute "h", a scene with a background B_I will have a lower relevance ratio with a text block with an attribute "h".

$$X_{IJ} = \begin{cases} \frac{X_2 - \sum_{i \neq I} X_{iJ}}{n_B} & \text{if } \sum_{i \neq I} X_{iJ} < X_2 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Information 3: The Number of the Same Words in a Text Block and Audio Data (Closed Caption) of the Video Scene

High relevance ratio is distributed to a text block T_j when there are many common words between T_j and the closed caption contained in a video scene S_I. Here, the audio is introduced as a hint and the accuracy of the association is expected to improve.

First, effective words are extracted from all the closed captions in a video scene using a domain-specific dictionary. A relevance score X_{IJ} between a scene S_I and a text block T_J is defined as Eq. 3, when the number of common words between T_j and the closed caption in the scene S_i is defined as W_{ij}.

$$X_{IJ} = X_3 \times W_{IJ} / \sum_j W_{Ij} \tag{3}$$

5 Evaluation of the Integration

An evaluation experiment is performed according to the method described in the previous section. Experimental conditions are shown in Tab. 2. Parameters were determined manually through several trials.

The purpose of this experiment was to evaluate the integration method itself. Therefore, shot detection, categorization, and clustering using backgrounds were performed manually. On the text analysis part, extraction of text blocks and analysis of the ordinal structure were performed manually, too.

Video shots and text blocks were associated according to the method described in the previous section, and the result was compared with a ground truth which was created manually.

Table 2. Experimental conditions.

(a) Features of data

cooking program	the number of recipes	duration
"K"	10	1'24"
"O"	10	1'18"
Total	20	2'42"

(b) Parameters

parameter	value
X_1	60
X_2	60
X_3	100

Table 3. The result of evaluation experiment (%).

Used Informations	"K"	"O"	Average
1. Ordinal restriction	20.2	20.5	20.4
2. Background info.	24.6	20.5	22.6
3. Closed caption	60.5	58.9	59.7
All (Proposed method)	**83.3**	**74.1**	**78.8**

The result is shown in Tab. 3. In this table, the results using the Informations 1, 2 and 3 individually are shown to be compared with the result of the proposed method to evaluate the effectiveness of the integration of multimedia information proposed in this paper.

The result shows that the accuracy by the proposed method is much higher than the accuracy using only one information source, and the total average accuracy is about 80%. Through this, the effectiveness of the proposed method is shown.

6 Conclusion

We have been working on integration of video with supplementary documents, such as cooking programs. We propose an integration system that performs semantic segmentations of video and text, and associate them together. This association was realized using the ordinal restriction in a recipe, cooccurrence of words in the text and the audio in the video, and the relation between the background in a video and words which describe the situation in the text.

We introduced the result of an evaluation experiment and showed the effectiveness of the proposed integration method. Through our method, many applications should become possible, such as a cooking navigation software.

Acknowledgement. Part of the work presented in this paper was supported by the Grant-in-Aid for Scientific Researches (14380173) from the Japanese Society for the Promotion of Science.

References

1. Reiko Hamada, Ichiro Ide, Shuichi Sakai, Hidehiko Tanaka: "Structural Analysis of Cooking Preparation Steps in Japanese", Proc. Fifth Intl. Workshop on Information Retrieval with Asian Languages *IRAL2000*, pp.157-164 (Oct. 2000)
2. Y. Ariki, Y. Saito: "Extraction of TV News Articles Based on Scene Cut Detection", *Proc. ICIP'96*, pp.456-460 (1996)
3. Koichi Miura, Reiko Hamada, Ichiro Ide, Shuichi Sakai, Hidehiko Tanaka: "Motion Based Automatic Abstraction of Cooking Videos", Proc. ACM Multimedia 2002 Workshop on Multimedia Information Retrieval, (Dec. 2002)
4. H. D. Wactlar, A. G. Hauptmann, M. G. Christel, R. A. Houghton, A. M. Olligschlaeger: "Complementary Video and Audio Analysis for Broadcast News Archives", *Comm. ACM*, Vol.45, No.2, pp.42-47 (Feb. 2000)