# Associating Cooking Video with Related Textbook

Reiko HAMADA
Graduate School of Engineering,
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-8656, Japan
reiko@mtl.t.u-tokyo.ac.jp

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

Shuichi SAKAI, and Hidehiko TANAKA
Graduate School of Engineering,
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-8656, Japan
sakai,tanaka@mtl.t.u-tokyo.ac.jp

## ABSTRACT

We have been handling video with supplementary documents, such as cooking programs, and are working on integration of such media. Through the integration, many applications will become possible, for example, reconstruction of multimedia data that supplement the information of each medium, construction of interactive database, or kitchen automation. Until now, we have proposed an integration system that perform integrative analysis of image, audio and text and associate each other. In this paper, we will introduce the latest text analysis result and discuss about future image and audio analysis of the proposed system.

## 1. INTRODUCTION

### 1.1 Background

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, and various research on analysis of each individual medium has been made. Especially, image processing has been the main topic when handling multimedia for a long time. But recently, it has started to be considered that image processing alone is insufficient for thorough understanding of multimedia data. In the 1990s, integrated processing that supplements the incompleteness of information from each medium has become a trend.

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, and (2)the video contains a lot of implicit information that integration could be helpful to thorough understanding of both media.

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 to achieve accuracy from elemental technologies by making use of domain specific characteristics. In our method, we examine and construct a practical system using relatively simple elemental technologies by reflecting the result of one medium's process to another. We will focus on cooking programs, so that we can take advantage of domain specific constraints and knowledge. Through the examination in this specific domain, and the usage of a supplementary document and its analysis, we aim for proposing a novel advanced multimedia integration method.

Using the result of this integration, we also propose an restructuring method of the multimedia data provided both from the video and the supplementary document. In this paper, we will introduce the latest text analysis result and discuss about future image and audio analysis of the proposed system.

### 1.2 Related Works

Many attempts have been made on news video indexing. In the Informedia project [1], many highly developed elemental technologies are used. But the integration strategy is relatively simple, such as merely combining hints from each medium in temporal order.

Though most of multimedia integration researches including [1] do not make use of supplementary documents, several studies aim at associating video with corresponding document like our method. But many of them only analyze each medium separately and then combine them. They usually do not make use of information of each medium for analysis of another medium. For example, in the research on aligning articles in TV newscasts and newspapers [2], they refer to nouns that appear both in newscasts and newspaper articles. Once aligning is done, semantic information from the newspaper will be available for TV newscast analysis.

Another research on synchronization between video images and drama scripts by DP matching [3] uses document information, too. They extract patterns from each medium and synchronize them by DP matching. Similar to the previous one, semantic scenes in a video can be detected by analysis of corresponding documents.

In drama, the order of scenes in the video is basically same in the script, but in programs such as cooking programs, the order of events differ in the video and the document. So, in our task, we must gather hints from each medium and integrate them effectively.

## 2. SYSTEM OVERVIEW

The outline of our method is shown in Fig. 1.

In cooking programs, the order of steps often differ between video and textbook. Nevertheless, there are still some restrictions, such as the time flow of processing materials (A material once processed never returns as it originally was). Therefore, extracting such restrictions from documents is essential for association with video. So, first, in the left side of the Fig. 1, we analyzed a large amount of document to extract keywords and gathered and classified them to create domain specific dictionaries. And using the dictionaries, structural analysis of a document is performed. More detail about structural analysis is described in the Section 3. Keywords extracted in this stage can be used later for image or audio analysis.

On the video side, cut detection is performed to the image sequence, and the video is separated into shots. Shots are too short as semantic scenes, so by image and audio analysis, shots are recombined as semantic scenes. Since keywords that appear in each program are limited, they can be used to restrict image analysis targets or to limit vocabulary for word spotting.

Finally, considering the semantic scenes of the video and the structure of the document, the video and the supplementary document are associated. Detailed plans are described in the Section 4.

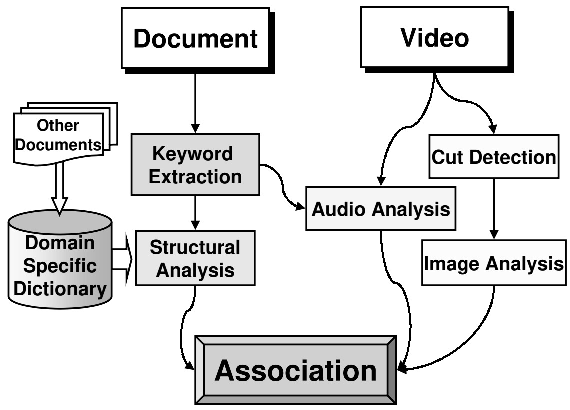


**Figure 1: Associating video with related documents.**

## 3. EXTRACTING ORDINAL RESTRICTIONS FROM DOCUMENTS

The outline of the structural analysis is shown in Fig. 2. We have proposed a method to create a process flow graph from a text to make the restrictions clear. By this graph, restricted and un-restricted orders could be distinguished clearly (directly linked orders can not be changed), and the structure of a cooking process becomes very clear.

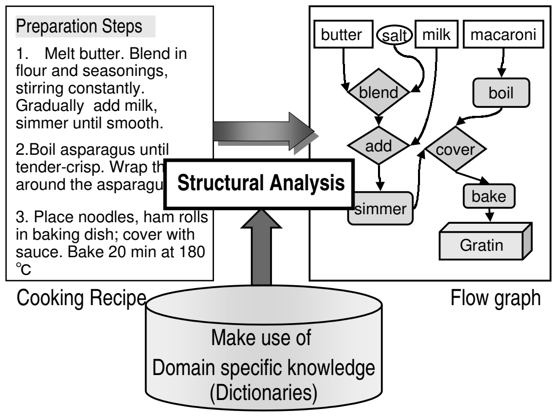


**Figure 2: Outline of structural analysis.**

**Table 1: Categorization of nouns and verbs.**

(a) Nouns

| Category | Example | Vocabulary |
|---|---|---|
| Ingredient | *carrot, chicken* | 1,636 |
| Seasoning | *salt, pepper* | 7 |
| Receptacle | *dish, plate* | 56 |
| Tool | *oven, knife* | 132 |
| Pronominal noun | *this, it* | 10 |
| Action | *cut in slices* | 89 |
| Total | | 1,930 |

(b) Verbs

| Category | Example | Vocabulary |
|---|---|---|
| Single | *roast, cut* | 231 |
| Mix | *add, sprinkle* | 70 |
| Separate | *divide, peel* | 44 |
| Place | *put, place* | 36 |
| Polysemy | *spread, return* | 6 |
| Employment | *make* | 2 |
| Total | | 389 |

To realize this, first, we created domain specific dictionaries by statistically gathering keywords from about 880 actual cooking documents gathered from a single WWW cooking program site, and categorizing and manually correcting them. Next, structural analysis on cooking preparation steps is done referring to the category of each word in the dictionary.

We created noun and verb dictionaries with cooking terms as special knowledge on the target domain, and a general keyword dictionary as general knowledge necessary for basic

Table 2: Categorization of general keywords.

| Category | Example | Vocabulary |
|---|---|---|
| Addition | still more, addition to | 2 |
| Condition | if, in the case of | 14 |
| Time | just before, after, when | 10 |
| Negation | not | 1 |
| Particle | Japanese particles | 27 |
| Conjunction | *ta, da, te, de*(in Japanese) | 4 |
| Total | | 58 |

analysis. Categories and vocabulary of the noun dictionary are shown in Tab. 1 (a), verb dictionary are shown in Tab. 1 (b), general keyword dictionary are shown in Tab. 2. Using the categorized words in the dictionaries, the next four processes are performed to analyze the structure.

**Process 1: Extract words and their categories by matching with the dictionaries**
Words in the dictionaries and step numbers are extracted and the categories in the dictionaries are tagged.

**Process 2: Make noun-verb sets**
Noun-verb sets are made considering that a verb modifies the nearest noun, satisfying the no-cross condition [5]. Whether each ingredient is added or removed is determined by a particle and the category of the verb.

**Process 3: Make blocks by connecting sets**
If the first verb of a set does not satisfy the case frame, it is connected to the previous set and form a block. In blocks, several ingredients are mixed and cooked, and each block becomes a new intermediate state.

**Process 4: Connect blocks**
Each block is connected to the nearest block which has common ingredients or step numbers or some keywords.

An example of the result is shown in Fig. 3. At the center of Fig. 3, cooking actions are arranged vertically in order. In the left and the right side of them, there are ingredients and seasonings as data, and each of those with an arrow pointing toward the center of the graph indicates it is an input data, and an opposite arrow indicates it is an output data. Since only ingredients and seasonings in the graph are the targets of structural analysis, extra information are just added in the graph as supplementary information.

An evaluation experiment on the proposed method was done on randomly selected 32 recipes (135 steps), and 82% of all steps were analyzed correctly. This result showed the effectiveness of the method [7].

# 4. ASSOCIATING VIDEO WITH DOCUMENTS

## 4.1 Video Analysis and Association

Structure of the video is needed to be analyzed to associate with the document analyzed as in the previous section. As

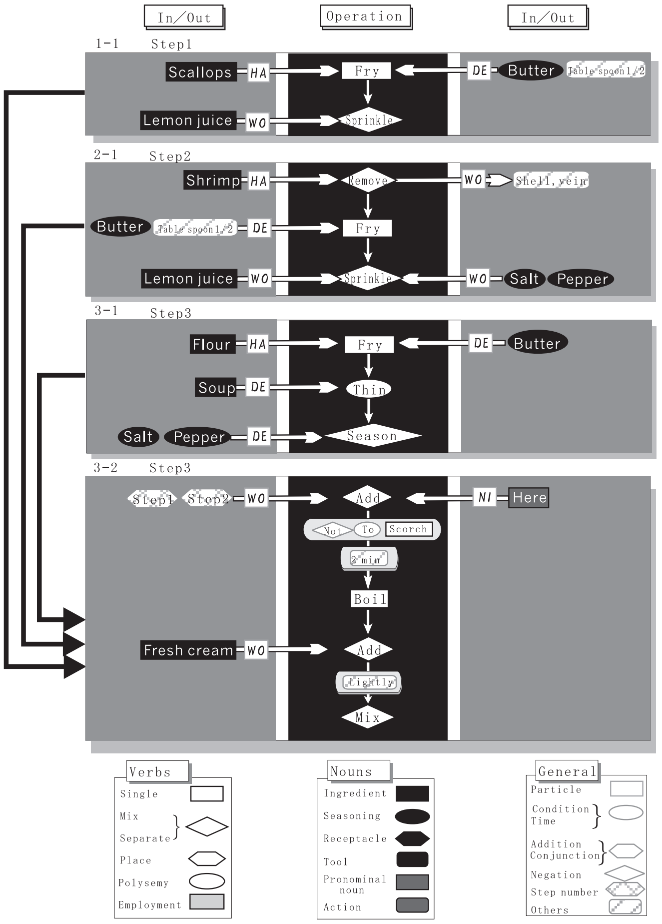


**Figure 3: Result of structural analysis of cooking procedures.**

mentioned in the Section 2, the finest segment in a video sequence is a shot. But, shots are too short to be considered as a semantic scene. In a cooking video, generally a semantic scene corresponds to a step in the supplementary cookbook, or a cooking action such as 'cut' or 'boil' something. In most case, a boundary of such semantic scene is one of the shot boundaries. So, we will detect such important boundary among many cuts and reconstruct semantic scenes by making use of image and audio analysis.

**Image Analysis:**

First, cut detection will be performed to a video sequence. Many cooking programs are taken in a studio under good lighting condition, so cut detection is easier than general video. In our research, we adopt a cut detection method using DCT clustering [4].

After cut detection, we classify the shots into three categories; (1)Hand shot, (2)Face shot and (3)CG/Flip shot, as shown in Fig. 4. Hand and face shots will be categorized by hand and face detection, and CG/Flip shots will be recognized by the duration of still frames.

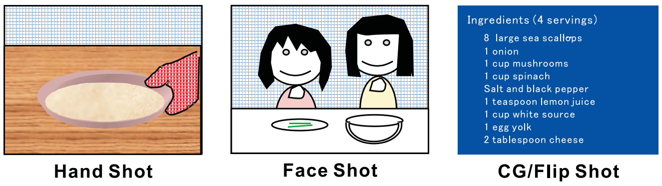


**Figure 4: Shot categories.**

As shown in Fig. 4, CG/Flip shots contain ingredients and quantity list. In most cases, they appear in the end of the program, so we can use these shots to recognize the structure of the video. In case an announcer reads aloud the list, that speech will be useful for audio recognition.

In Face shots, a scene of a kitchen and the upper half of the body of a teacher (and an assistant) is on the screen. Though each objects, such as tools or ingredients are too small to analyze, faces of a teacher or an assistant is analyzable.

On the other hand, Hand shots are close-ups of tools and hands cooking something. From a document that corresponds to the video, possible ingredients or tools in the screen can be extracted. If a database of information about colors or textures of them are available, image analysis can be done easier by performing to narrow downed objects. Furthermore, we are planning to recognize scenes by making use of motion of hands. In the field of gesture recognition, many attempts have been made to recognize delicate motion of hands. These methods can be applied to hands in front of general backgrounds (in many cooking video, the background is kitchen), and rough motion analysis will be made. Adding to that, simple analysis of background (whether the hand is on a sink or a kitchen range) would be useful to infer the scene. Telops can also be used when available.

**Audio Analysis:**

The contents of speech is an important hint for association. It is generally said that either the speaker should be specified or the vocabulary be limited to achieve enough accuracy in speech recognition. In the case of cooking programs, first, since they are usually progressed by conversations, there are always several people on the screen, and next, the speakers change periodically, so it is difficult to specify them. Nevertheless, since the domain is limited and keywords can be extracted from the document, we can limit the vocabulary for speech recognition so that enough accuracy could be expected.

Considering this advantage, we are planning a word spotting system that, in the first stage, spot words using general voice templates for keywords extracted from documents. In the next stage, new templates are recreated from the target audio stream using the result of the first stage, and then the second spotting is done to pick up the words that were overlooked in the first stage. Now we are performing simple preliminary experiments on this method to examine the possibility of the method.

**Association:**

Using the result of image and audio analysis, the structure of video stream will be analyzed. An unit of structure will be a step or an action. The keywords in the audio stream or the result of analyzing action and object in the image will be used to associate with the document. The image of the final outcome of the associated data is shown in Fig. 5. Each step, or particular motions in the document are linked to the corresponding part of video. In the future, the result of association will become useful applications linked with electric appliances in the kitchen.

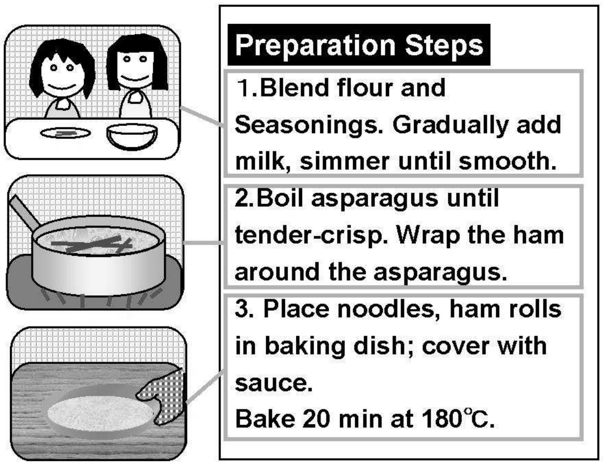


**Figure 5: Association of video and supplementary document in a cooking program.**

## 4.2 Preliminary Experiment

To examine the possibility, an experiment on associating documents with audio scripts has been performed. We collected the video from a broadcast program, and cooking documents from a related WWW page. Audio stream and cut boundary were written down and detected manually. The experiment was performed in the following manner:

1. Extract co-occurring keywords both in text and in video. Keywords are only “Noun” and “Verb”.
2. The shot that has most common words with a specific step belongs to the step.

We performed this experiment to three programs (10 minutes each). We resulted in classifying shots correctly by 70% in average as shown in Tab. 3.

# 5. CONCLUSION AND FUTURE WORK

We proposed an integration method of video with supplementary documents. In the proposed method, we aimed for realizing a practical system avoiding the use of complex and difficult elemental technologies, by reflecting the result of document analysis to audio analysis, and the result of audio analysis to image analysis. We introduced the result of preliminary experiments to show the validity of the method.

**Table 3: Result of shot alignment to preparation steps.**

| Program# | 1 | 2 | 3 | Total |
|---|---|---|---|---|
| Total Step Number | 5 | 8 | 5 | 18 |
| Total Shot Number | 18 | 19 | 17 | 54 |
| Correctly Aligned Shots | 13 | 11 | 14 | 38 |
| Accuracy | 72% | 58% | 82% | 70% |

We are currently planning to develop the image analysis part and make use of the structural analysis of documents, in order to realize a practical system.

In the future, we are considering many application, such as a database that can retrieve recipe, specific action, material, or automatic video editing from a document.

## 6. REFERENCES


[1] 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.

[2] Y. Watanabe, Y. Okada, T. Tsunoda, M. Nagao: "Aligning Articles in TV Newscasts and Newspapers (in Japanese)", *Journal of JSAI*, Vol.12, No.6, Nov. 1997.

[3] Y. Yaginuma, M. Sakauchi: "Content-Based Retrieval and Decomposition of TV Drama Based on Intermedia Synchronization", *First International Conference on Visual Information Systems*, pp.165-170, Feb. 1996.

[4] Y. Ariki, Y. Saito: "Extraction of TV News Articles Based on Scene Cut Detection", *ICIP'96*, pp.456-460, 1996.

[5] S. Kurohashi, M. Nagao: "A Syntactic Analysis Method of Long Japanese Sentences Based on Coordinate Structures' Detection (in Japanese)", *Journal of Natural Language Processing*, Vol.1, No.1, pp.3-20, Mar. 1994.

[6] H. Furuya, H. Nakai, S. Yamazaki Ed. (in Japanese): "The Synthetic Japanese Text for Junior High School -revised and enlarged edition-", *Sho-shin-sha*.

[7] R. Hamada, I. Ide, S. Sakai, H. Tanaka: "Structural Analysis of Preparation Steps on Supplementary Documents of Cultural TV Programs", *Proc. Forth Intl. Workshop on Information Retrieval with Asian Languages*, pp.43-47, Nov. 1999.

[8] Language Media Lab, Grad. School of Informatics, Kyoto Univ.: "Japanese Morphological Analysis System JUMAN version 3.6", Nov 1998.