Associating Video with Related Documents

Reiko HAMADA, Ichiro IDE, Shuichi SAKAI and Hidehiko TANAKA
Graduate School of Electrical Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, JAPAN
TEL: +81-3-5841-7413, FAX: +81-3-5800-6922
E-mail:{ reiko | ide | sakai | tanaka } @mtl.t.u-tokyo.ac.jp

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 issue 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 1990's, 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, cultural 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 supplementary texts.

Many attempts have been made to index video by means of multimedia integration. But sufficient accuracy for practical use is not necessarily achieved since their subjects are 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 method, we also propose an integrative restructuring method of the multimedia data provided both from the video and the supplementary document.

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 according to time stamps.

In the research on aligning articles in TV newscasts and newspapers[2], they use nouns which appear in TV newscasts and newspaper articles that correspond with each other. Once aligning is done, semantic information from the newspaper is 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 same as that in the script, but in cultural 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 Associating Video with Related Documents

2.1 System Overview

The outline of our method is shown in Fig.1. First, we analyze a large amount of documents and extract keywords. Then, the extracted keywords are gathered and classified to create a domain specific dictionary. Next, using the dictionary, the structure of a document is analyzed.

On the video side, cut detection and word spotting are performed to detect semantic scene boundaries. Keywords that appear in a document are specific to each program, so they can be used for word spotting. Finally, considering the semantic scenes of the video and the structure of the document, we associate the video and the supplementary document.

We are currently investigating the possibility of several applications after the association completes.

2.2 Boundary Detection in Video Sequence

The biggest hint in segmenting video into semantic scenes is cut detection. 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].

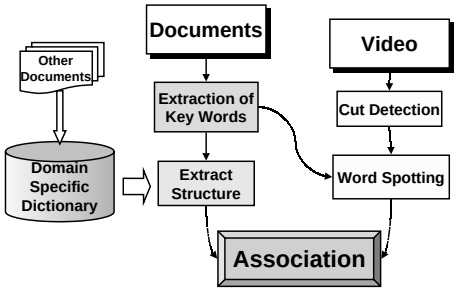

Figure 1: Associating video with related documents.

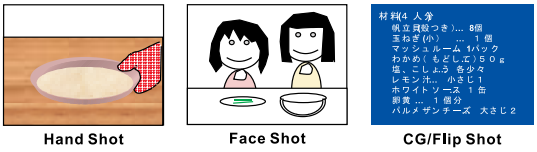

Figure 2: Shot categories.

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.2. The first two shots can be categorized by hand and face detection, and CG/Flip shots are recognized by the duration of still frames.

Hand shots are close-ups of tools and hands cooking something. It should be possible to narrow down subjects by audio and document analysis, and try to recognize specific objects. Telops can also be used when available.

The contents of speech is an important hint to association. It is generally said that either the speaker should be specified or the vocabulary be reduced for achieving enough accuracy in speech recognition. In the case of cooking programs, it is difficult to specify the speakers. But the domain is limited in this case, and keywords can be extracted from the documents, so we can highly reduce the vocabulary for speech recognition and enough accuracy could be expected. In the future, we are planning to reflect the speech analysis result to detect the shot boundary, too.

2.3 Extracting Ordinal Restrictions from Documents

An example of a supplementary document is shown in Fig.3. The document consists of the "Ingredients" part and the "Preparation Steps" part. "Ingredients" can be used to update the dictionary, or analyze words in the "Preparation Steps".

"Preparation Steps" gives explanation on how to cook the "Ingredients". First, we must analyze the "Preparation Steps" to associate with the video. 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 (*e.g.* A material once cooked never returns raw). Therefore, extracting such restrictions from documents is essential for association.

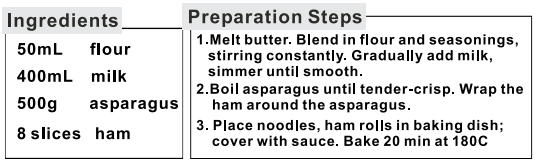

Figure 3: Example of a supplementary document for a cooking program.

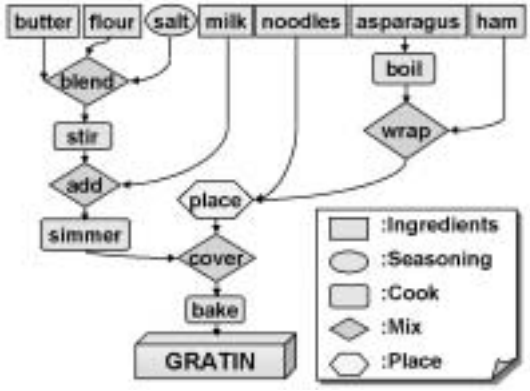

Figure 4: Example of a preparation order graph derived from the "Preparation steps" in Fig.3.

We propose a preparation flow graph to make the restrictions in a document clear, as shown in Fig.4. By this graph, restricted and un-restricted orders could be distinguished clearly (directly linked orders can not be exchanged).

2.3.1 Categorization of Cooking Terms

In our method, we only use nouns and verbs as keywords to make a data flow graph. We categorize nouns and verbs according to Tab.1 and 2. Words that could not be categorized to any of the categories in the Tables are omitted from the analysis.

As shown in Tab.1, we categorize nouns to Ingredients ([Ing]), Seasonings ([Sea]), Receptacles, and Tools. In this experiment, we only use ingredients and seasonings for analysis. In future, we are planning to use receptacles to know which ingredients are put together in a receptacle.

As shown in Tab.2, we categorize the verbs into "Cook", "Blend", "Separate" and "Place". But in this experiment, we categorize these types into two big titles, [Sol] and [Mix]. [Sol] includes "Cook" verbs and [Mix] includes "Blend", "Separate" and "Place" verbs. [Sol] cooks single ingredients, such as 'Bake' and 'Cut'. On the other hand, "Blend" mixes several ingredients, such as 'Add', 'Mix' and 'Sprinkle', "Place" puts ingredients in receptacles, such as a 'baking dish'. If several ingredients are put into a dish, these ingredients are mixed in the dish. So we categorized "Place" verbs in the [Mix] category. In the case of a "Separate" verb, it is clear that originally, these ingredients were one ingredient. So, "Separate" verbs are also categorized in the [Mix] category.

We are planning to make a domain specific noun and verb dictionary, in which words are categorized according to

Table 1: Noun Categories.

Noun	Ingredients	Seasonings	Receptacles	Tools
Symbol	[Ing]	[Sea]	(Currently not used.)	
Example	*Carrot*, *Chiken*	*Salt*, *Pepper*	*Baking Dish*	*Oven*, *Knife*

Table 2: Verb Categories.

Verb	Cook	Blend	Separate	Place
Symbol	[Sol]	[Mix]		
Example	*Bake*, *Cut*	*Add*, *Mix*	*Separate*, *Divide*	*Place*, *Put*

Tab. 1 and 2. In the following experiment, we will discuss our method assuming that an ideal dictionary already exists.

2.3.2 Structural Analysis Experiment

We will introduce the process of structural analysis for making a data flow graph. The whole process is shown in Fig. 5.

1. **Extract nouns and verbs that the document has in common with the dictionary.**
   Extract nouns and verbs in the document and replace with the category in the dictionary, then make "Noun - Verb" sets. Nouns and verbs that exist in the dictionary are extracted and others are ignored. We consider that a verb modifies the nearest noun, satisfying the no-cross condition[5]. We actually applied the method to Japanese documents, so nouns and verbs in the example are in reverse orders to those in English.
2. **Create intermediate states referring to verbs.**
   The sets with [Mix] verbs are connected with any previous sets. New numbers are given to each of the new states in this process.
3. **Connect intermediate states referring to nouns.**
   The intermediate states created in the previous step are connected and data stream is completed in this step. First, if there is [Process #], all the states in that process are connected with [Process #]. Next, other states are connected with the nearest state which has the same [Ing #]. And finally, if there are some states left, they are connected with the nearest states which has the same [Sea #]. If the categorization of "ingredients" and "seasoning" in the dictionary is perfect, we may not need the last rule. But, there may be mistakes in the dictionary, such as mis-categorization of an ingredient as a seasoning or vice versa, since it is difficult to statically categorize some materials as an ingredient or a seasoning.

2.3.3 Automatic Creation of a Domain Specific Dictionary

In the previous paragraph, we assumed that an ideal dictionary already exists. But, creating a dictionary in each domain manually is burdensome. So, we are attempting to automatically create a domain specific dictionary.

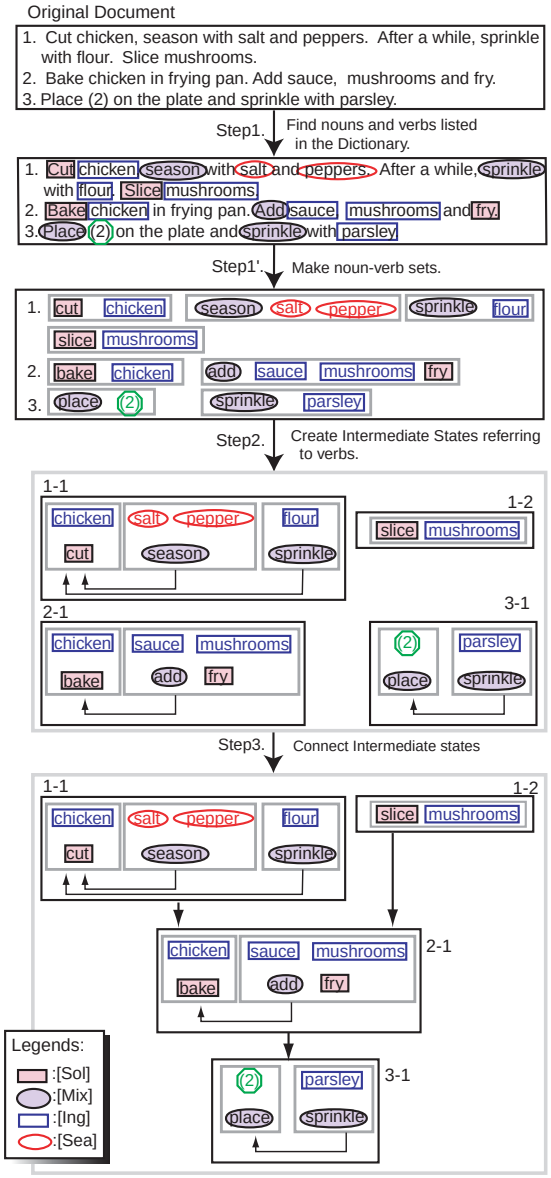

Figure 5: Structuring preparation orders in a cooking textbook.

The text that explains how to make something uses characteristic keywords compared to general sentences. For example, there are few words such as 'cook', 'boil', 'bake' and 'chicken' in newspapers. Then, it is relatively easy to extract such characteristic keywords from these documents statistically. We are now investigating several statistical methods, using frequency of appearance, frequency of co-occurrence, TF-IDF, and so on.

2.4 Associating video with documents

The image of the final outcome of the associated data is shown in Fig.6. Each step, or particular motions in the document are linked to the corresponding part of video. In the future, we will associate the video and the document using the results from both document analysis and image analysis. We will show the result of a preliminary experiment on such association in the next chapter.

Figure 6: Association of video and supplementary document in a cooking program.

3 Preliminary Experiments

3.1 Document Analysis Experiment

In this Section, we will examine the structure analysis by extracting the restrictions from a document. We applied the steps mentioned in the previous Chapter to an actual supplementary document for a cooking program. The documents were gathered from a WWW page. Note that the domain specific dictionary used in the experiment was manually created in this case.

The actual document and the dictionary is shown in Fig.7. The experiment was performed to a Japanese document, and the English sentences in Fig.7 are translations. The result of the cooking flow structural analysis is shown in Fig.8. As shown in Fig.8, the ordinal structure of cooking was extracted correctly. Now, we are planning to perform it on many other documents for evaluation.

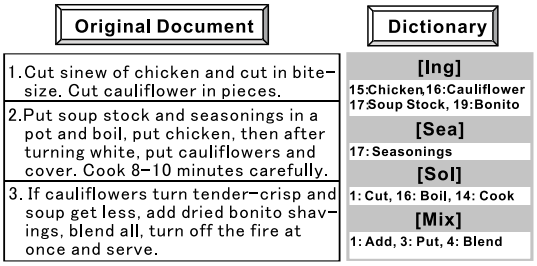

Figure 7: The original document and dictionary in the experiment.

3.2 Associating Experiment

An experiment on associating documents with audio scripts has been performed. We collected a document from a WWW page, and the video from a broadcast program. Audio stream and cut boundary were written down and detected manually. The experiment steps were:

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

We performed this experiment on three programs (10 minutes each). As the result shown in Tab.3, we resulted in classifying shots correctly by 70% in average.

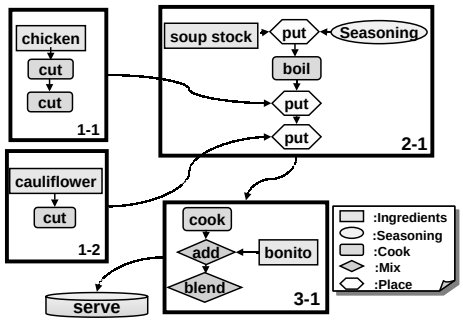

Figure 8: Result of cooking flow structure analysis.

Table 3: Result of shot alignment to preparation steps.

Program#	1	2	3	Sum
Total Step Number	5	8	5	18
Total Shot Number	18	19	17	54
Correctly Aligned Shots	12.5	11	14.5	38
Accuracy	69%	58%	85%	70%

4 Conclusion and Future Work

We proposed an integration method of video with supplementary documents. In our method, we aim for realizing a practical system avoiding 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.

We are currently planning to automatically create a domain specific dictionary, and restructure shot boundaries using the result of speech recognition and improve the association accuracy.

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

References

[1] Hauptmann, A. G. and Witbrock, M. J.; "Informedia News-on-Demand: Using Speech Recognition to Create a Digital Video Library", *CMU Tech. Rep.*, CMU-CS-98-109, Mar 1998.

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

[3] Yaginuma Y., Sakauchi M.; "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] Ariki Y. and Saito Y.; "Extraction of TV News Articles Based on Scene Cut Detection", *ICIP'96*, pp.C456-460, 1996.

[5] Kurohashi S. and Nagao M.; "A Syntactic Analysis Method of Long Japanese Sentences based on Coordinate Structures' Detection", *Journal of Natural Language Processing*, Vol.1, No.1, Mar., 1994(In Japanese).