Systems and Computers in Japan, Vol. 36, No. 2, 2005
Translated from Denshi Joho Tsushin Gakkai Ronbunshi, Vol. J86-D-II, No. 11, November 2003, pp. 1647–1656

Associating Semantically Structured Cooking Videos with Their Preparation Steps

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

[1]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, 113-0033 Japan

[2]National Institute of Informatics, Tokyo, 101-8430 Japan

SUMMARY

The dramatic growth in the amount of multimedia data in recent years has caused techniques for analyzing those data to become increasingly important. To make use of information contained in multimedia data to the fullest extent, attention has been focused on techniques for processing multiple types of media in an integrated manner. Among the various kinds of videos that exist, the authors selected cooking videos, which are closely related to everyday life, to propose a technique for associating the videos with preparation steps described in related text-based materials referred to simply as textbooks. With educational videos having related textbooks such as cooking program videos, the textbooks are easier to consult than the videos but the videos contain useful visual information that cannot be represented in the textbooks. Therefore, integration of the video and textbook is expected to result in high-order video structuring and indexing. In this paper, the authors first analyze the video structure and define the video blocks that are to be the video units used for association. They also analyze the preparation steps described in the closed captions and textbooks and propose and define an association technique based on keyword extraction. In addition, they show through evaluation experiments that association in terms of video block units or, in other words, video indexing, can be performed with high precision by using the proposed technique. They also show that limiting the subject matter and skillfully incorporating relatively simple elemental techniques enables a practical level of precision to be obtained. © 2005 Wiley Periodicals, Inc. Syst Comp Jpn, 36(2): 51–62, 2005; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/scj.20131

Key words: cooking video; textbook; association; video block; video indexing.

1. Introduction

Advances in telecommunication technology in recent years have contributed to a steady increase in the amount of multimedia data that can be acquired. Multimedia data analysis is becoming an increasingly important technology for organizing this massive amount of data for efficient storage and retrieval.

Multimedia data generally consist of images, audio, and text. Conventionally, research concerning automatic analysis techniques had been done separately for each of these types of media. For example, various research projects concerning image analysis covered such topics as cut detection or similar image detection. However, just like it is difficult to recognize a general object in a single image, it is also difficult to analyze the semantic contents of multimedia data from only the images. Also, with text analysis, high-dimensional semantic contents can be analyzed relatively easily such as in the extraction of important sentences or phrases or the creation of summaries, but gaining an

overall understanding of multimedia data from only text is not simple, and combining the text with video enables the text to be used more effectively. Since the automatic analysis of multimedia data by using each type of medium alone has the kinds of limitations described above, to perform high-precision semantic content analysis, it is necessary to process multiple types of media in an integrated manner.

As part of our research concerning this kind of integrated media processing technique, we included intelligent structuring and indexing specifically targeting cooking videos [1]. With educational videos having related text-based materials such as cooking program videos, the text-based materials, which we will refer to simply as textbooks, are easier to consult than the videos, but the videos contain a lot of useful visual information that cannot be represented in the textbooks. Therefore, integrated processing of these videos and textbooks is expected to provide mutual supplementation of the information or shortcomings of each individual medium. As a result of this supplementation, not only can the implementation of high-order video structuring and indexing be expected, but also the generation of new easy-to-use multimedia data having a form in which the text and video are linked. In addition, applications connected to intelligent cooking support such as video digests or knowledge extraction using analysis results can be considered.

In this paper, we propose a technique for analyzing the structure of cooking videos and associating the videos in terms of individual preparation steps with preparation steps described in related textbooks. By limiting the subject matter to cooking videos, we aim to obtain a practical level of precision that could not previously be obtained by using general techniques, while using existing relatively simple elemental techniques.

The remainder of this paper is structured as follows. Section 2 presents an overview of the technique for associating videos with preparation steps described in related textbooks, and Section 3 proposes a cooking video structure analysis technique for making this association. Section 4 proposes a text analysis technique and Section 5 proposes a technique for associated videos with preparation steps. Section 6 shows evaluation test results for the video analysis technique proposed in Section 3 and the association technique proposed in Section 5. It also discusses the evaluations based on those results. Section 7 presents conclusions summarizing the entire paper.

2. Technique for Associating Videos with Preparation Steps in Textbooks

2.1. Video structure and definition of terms

In this paper, we refer to the set of all data including audio or teletext that is broadcast synchronized with images as video. An image within a video consists of multiple frames, and a collection of graphically continuous frames is called a shot. The boundary (point where the image changes) between shots is called a cut, and a collection of shots that have a further semantic coherence is called a scene. Figure 1 shows this kind of video structure.

2.2. Related research

Research for aligning independent external text with video included a research project [2] that used news video titles (open captions) and structure information about electronic newspaper articles to calculate a degree of similarity, which was used for aligning the newspaper articles with topics within the news videos. The alignment performed in this kind of research enables semantic information obtained from newspaper articles to be used for analyzing news videos. However, the video content analysis performed for alignment only referenced the title text, and specific video contents were not taken into consideration.

Research related to the synchronization of drama video with scenario documents by using DP matching [3] extracted patterns that can be referenced from multiple media and synchronized the media by using DP matching to optimize the alignment of those patterns. Although this research also uses alignment to structure videos and create databases, its property association technique essentially differs from the technique used in our current research because with a drama, the order of the scenes in the video and the order in the scenario documents are basically the same while with a cooking program, the preparation steps in the textbook and video often differ. Therefore, in our current research, we cannot use a one-dimensional synchronization technique that simply follows along a time series as in DP matching. As a result, our technique references keywords within each medium to associate the media so that they match the structure of the video as described below.

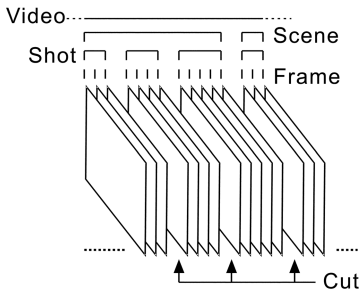


Fig. 1. Graphical structure of a video.

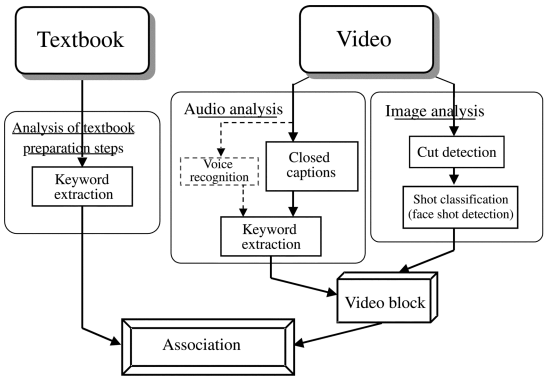


Fig. 2. Technique for associating a cooking video with a related textbook.

2.3. Overview of the proposed technique

Figure 2 shows the conceptual structure of the association technique. First, the video structure and textbook preparation steps are analyzed in parallel. Then, the contents of each are judged in an integrated manner to associate the video with preparation steps.

Video structure analysis proceeds for both images and audio. The images are divided into individual shots according to cut detection, and each shot is classified based on the image contents. Since various types of elemental techniques have already been researched in relation to the image processing required here, we efficiently combined these existing techniques for use in our current research. For the audio, text processing is executed using closed captions,* which show a continuous written presentation of the main audio channel. The text processing consists of morphological analysis of the text followed by the extraction of keywords such as nouns for the names of ingredients and verbs representing cooking actions. The structure of the video is extracted from the results of these analyses for association with preparation steps in the textbook. This is described in detail in Section 3.

Morphological analysis and keyword extraction are also performed for the textbook analysis in a similar manner as was done for the closed captions. This is described in detail in Section 4.

Finally, the results of each analysis are used to associate the video with the preparation steps in the textbook. This is described in Section 5.

*Closed captions are digital text providing a continuous written presentation of the main audio channel in the form of teletext mainly for hearing-impaired viewers. Closed captions are provided by teletext in many TV programs and are also included in live broadcasts in the United States. Closed captions are also provided in an increasing number of programs in Japan, including cooking programs.

3. Structure Analysis of Cooking Videos

The structure of a cooking video is analyzed in order to associate the video with preparation steps in the related textbook. Although analyses are performed for both image and audio data, the video structure is first analyzed according to image analysis to define the video blocks that are the units used when associating the video with preparation steps in the textbook. Keywords to be used for association are also extracted according to analysis of the closed captions.

3.1. Image analysis

3.1.1. Definition of video blocks

Image analysis first detects cuts to divide the video into individual shots. Various cut detection techniques have been investigated including a technique that uses color histograms or color correlograms to detect color tone changes in an image [4]. However, the proposed system employs a technique that uses DCT clustering [5]. Since a cooking video is shot under ideal lighting conditions in a studio, high-precision cut detection can be expected. Also, since this cut detection technique also obtains features of clusters constituting each shot at the same time that it detects cuts, we believe that those features can be used in the future for shot classification, which is performed after cut detection.

Shots are classified after cut detection in order to analyze the video structure. The shots in a cooking video can be broadly classified into face shots (human shots) (a) and hand shots (b) based on the configuration within the image as shown in Fig. 3. Face shots (a) can be further divided into full shots (full body shots) (a1) and bust shots (upper body shots) (a2).

Figure 4 shows an example of the shot configuration in an actual cooking video according to these shot classes. When we focused on the shots immediately following the separation of preparation steps in these shot configurations of cooking videos, we found that more than 90% were face

(a1) Full shot (b1) Bust shot

(a) Face shot (b) Hand shot

Fig. 3. Shot classes in cooking videos.

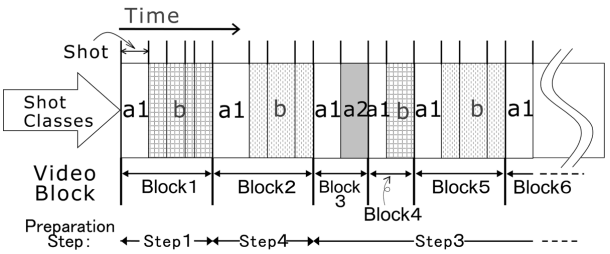


Fig. 4. Example of the shot configuration in a cooking video.

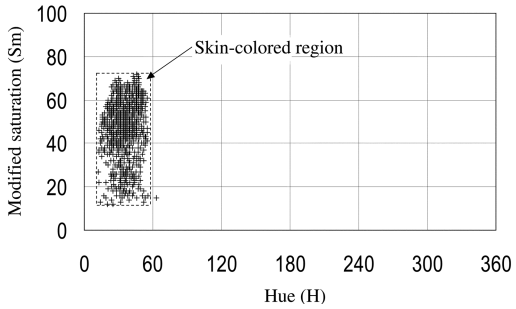


Fig. 5. Skin-colored region distribution in the *H–Sm* plane (based on sampling of actual skin-colored regions).

shots (a) and among these, more than 90% were full shots (a1).*

As a result, we decided to define a "video block" as one unified preparation step and to let these blocks be the smallest units to be used for association in the video. As shown in Fig. 4, a video block is a set of consecutive shots starting with a full shot (a1) and extending until the next full shot (a1) appears.

3.1.2. Detection of face shots

Face shot detection is particularly important in shot classification because it provides clues for dividing a video into video blocks. Since a human face appears in the image, a face shot can be detected by extracting a face region.

There are various techniques for extracting a face region. However, since it is sufficient here to simply know the existence, position, and size of a face region, instead of using more advanced techniques for complex modeling of the positions of the eyes, mouth, or other features, we decided to detect face shots with a practical level of precision by simply and robustly extracting face regions according to a procedure that:

(1) extracts skin-colored regions, and
(2) determines face regions according to fixed conditions from the detected skin-colored regions.

To extract skin-colored regions, we used the modified HSV color system [6] (*H*: hue, *Sm*: modified saturation, *V*: intensity), which is suitable for this purpose.

For the *V* value, only threshold value processing is performed to eliminate dark regions. Skin color is determined by referencing the colors of various actual face regions, setting the rectangular region in the *H–Sm* plane (Fig. 5 [7]) as the skin-colored region, and deciding whether or not the *H* and *Sm* values in each pixel are contained in this region. At this time, a 3×3 median filter is applied to the binarized image to eliminate noise.

Since skin-colored regions are extracted based only on color, other similar colored objects such as hands, wooden rice paddles, or tables also end up being detected, not just human faces. To eliminate these kinds of objects, after the skin-colored region is extracted, face regions are extracted according to the following kinds of conditions.

(1) For the aspect ratio $r = x/y$ of the rectangle circumscribing an extracted skin-colored region, only those regions that satisfy $r_{min} \le r \le r_{max}$ are extracted as face regions

This condition is used because when a human face is filmed, there are generally certain constraints regarding its shape. Also, since this value differs significantly according to the direction the face is pointing, the range of values has a certain degree of margin in the proposed technique, and the values $r_{min} = 0.38$ and $r_{max} = 1.4$ were set empirically.

(2) Regions touching the edge of the screen are removed

This condition is used because in cooking videos, camera work in which a face region touches the edge of the screen rarely appears, and when a skin-colored region touches the edge of the screen, it is often because hands are being filmed during cooking steps.

(3) Regions for which the area is too large (at least 1/12 of the screen) and relatively small regions (less than 30% of the largest area) are removed

This condition is used because in cooking videos, face regions are filmed with a certain size on the screen, and when a region is too large or too small, it is often because a nonface region (such as a wall or table) ended up being detected. Also, in a cooking video, when multiple face regions appear in the same shot filmed in the structure of a studio set, the areas of those faces are of a certain size. In addition to taking into consideration variations in the area of a face due to the direction the face is pointing, relatively

*Results of investigating 137 preparation steps in 38 recipes from numerous cooking programs.

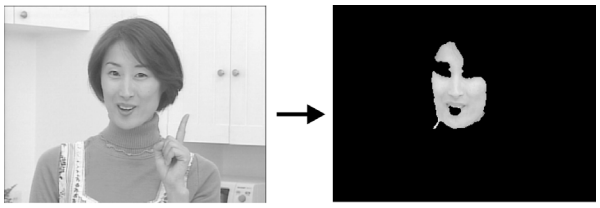


Fig. 6. Example of face region extraction.

small regions are also removed as noise. These kinds of threshold values are set empirically.

(4) At least part of the region is in the upper half of the image

This condition is used because in the shot composition in cooking videos, face regions always appear in the upper half of the image.

Among these conditions, condition (1) uses features of general face images. Although conditions (2) to (4) were derived from properties of cooking videos, they can also be applied to general videos that are filmed in a studio set.

Figure 6 shows an example of face region extraction. Also, full shots (a1) and bust shots (a2) are classified by performing threshold value processing (threshold value: 1/36 of the entire screen) on their areas after the face regions are extracted.

3.2. Voice text analysis

To analyze the audio contents in the current research, we used the closed captions that are provided from broadcasting stations as a continuous written presentation of the main audio channel and analyzed this as voice text instead of performing voice recognition. When closed captions are obtained as text data, they are synchronized with the video based on appearance times, which are recorded together with the text data.

After morphological analysis is executed for the closed caption text that was obtained, words that can be used as association clues such as names of ingredients and verbs related to cooking are extracted as keywords. This analysis is explained in detail in Section 4 together with the analysis of preparation steps described in the related cookbooks.

4. Analysis of Preparation Steps in Related Textbooks

To associate videos with preparation steps described in related textbooks, the text in the preparation steps must

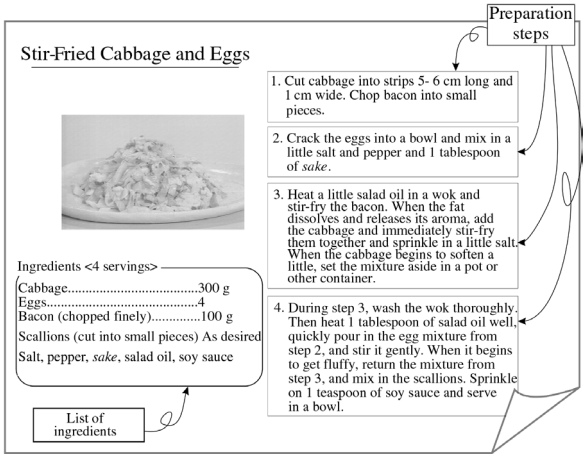


Fig. 7. Example of a cooking textbook.

also be analyzed in parallel with the video structure analysis. For the analysis of the preparation steps, morphological analysis of the text is executed in a similar manner as for the closed captions, followed by the extraction of keywords such as nouns representing names of ingredients and verbs representing cooking actions.

Figure 7 shows an example of a cooking textbook. The textbook consists mainly of "preparation steps" and a "list of ingredients," and the latter providing important clues for extracting ingredient names.

The text analysis for the closed captions and the preparation steps in the textbooks is performed as follows.

- The Japanese morphological analysis system JUMAN [8] is used for the morphological analysis.
- Only ingredient names that match the list of ingredients in the textbook are extracted. A dictionary is used at this time, and even if the presentation differs in terms of hiragana, katakana, or kanji, they are handled as the same ingredient name. (Verbs are handled in a similar manner below.)
- Verbs are extracted from the results of the morphological analysis, and "*kudasaru*," "*dekiru*," and "*suru*" (independent word not accompanying a noun in an s-row changing verb) are treated as verbs that are not related to cooking and are excluded.
- Only the above ingredient names and verbs are considered as keywords.
- Text punctuated by a period and commas or spaces following verbs* is treated as one sentence, and

*In the closed caption text of cooking videos, commas rarely appear and spaces are often used.

ingredient names and verbs that are judged to belong to one sentence are treated as related words.

When preparation steps that are described in textbooks are analyzed, if a previous step number is referenced and "*no* + noun" does not appear immediately following that step number, all ingredient names included in the referenced step are supplied. However, the ingredient names that are supplied are only for a directly referenced step, and even if the referenced step further references a previous step, no ingredient names are supplied for that previous step. Table 1 shows specific examples of references to preparation steps.

Also, in relation to verbs, to deal with differences in expression between the video and textbook, a dictionary like the one shown in Table 2 was created as necessary, and verbs that matched cooking actions in the dictionary were replaced by normalized verbs for analysis. This dictionary includes verbs that are replaced by high-level concepts and verbs that deal with differences in expression. Verbs that represent 31 cooking actions, including those shown in Table 2, were recorded in the dictionary. To use this dictionary as a more practical tool, cooking dictionaries or other recipes must be used to increase the vocabulary it contains.

This text analysis processing uses a property specific to cooking textbooks to simplify processing in the part that uses words in the textbook's "list of ingredients" as keywords. However, this technique can also be applied to text from other domains as long as a dictionary is created that describes keywords that are the subject of each domain.

Table 2. Contents of the verb dictionary

Normalized verbs	Cooking actions
kiru (cut)	*sengiri ni suru* (julienne) *usugiri ni suru* (thinly slice) ...
ireru (enter)	*nagashiireru* (pour in) *kuwaeru* (add) ...
kirime wo ireru (make shallow cuts)	*kireme wo ireru* (score) *kirikomi wo ireru* (put cuts in the side of)
ajimi suru (taste)	*aji wo miru* (taste)
arau (wash)	*arainagasu* (rinse)
yaku (roast, bake, grill, toast)	*yakimasu* (roast, bake, grill, toast) (colloquial speech)

5. Association of Video Blocks with Preparation Steps

When associating videos with preparation steps, video blocks are used as the units for videos and individual steps in the textbook are used as the units for preparation steps.

For a cooking video, the order of the preparation steps in the textbook does not necessarily match the order of the steps in the video. In addition, it is often the case that one preparation step is divided and appears in two or more locations in the video so that no video corresponds to a preparation step or, conversely, video does not correspond to any preparation step. Therefore, the association determines which preparation step a video block is to be associated with based on extracted keywords according to the

Table 1. Example of preparation steps that are referencing previous steps

Step 2:	Enter [1]$_{(\text{step number})}$ in a mortar, add flour and sugar, and grind them together.
→	Since "*no* + noun" does not appear immediately following the referenced step number [1], all ingredient names contained in step 1 are supplied here.
Step 5:	Thinly slice the duck meat$_{(\text{noun})}$ of$_{(\text{possessive} = no)}$ [3]$_{(\text{step number})}$ (step number + *no* + noun construction) and serve it in a bowl.
→	Since "*no* + noun" appears immediately following the referenced step number [3], no ingredient names are supplied in this case.
Step 4:	Stir fry [1]$_{(\text{step number})}$ in a frying pan.
Step 5:	Place the pot$_{(\text{noun})}$ of$_{(\text{possessive} = no)}$ [3]$_{(\text{step number})}$ (step number + *no* + noun construction) over the heat, add [4]$_{(\text{step number})}$, and cook.
→	Since step number [1] is referenced in step 4, the ingredient names contained in step 1 are supplied here.
→	Although step number [3] and step number [4] are referenced in step number 5, since "*no* + noun" appears immediately following the location of step number [3], its ingredient names are not supplied. Also, the only ingredient names supplied at the step number [4] location are the ingredient names contained in step number 4. (The ingredient names for step 1 are not supplied.)

following procedure. Figure 8 shows how this association is executed. Also, a score is set for each keyword according to the following equation. This equation takes into consideration the novelty of an appearing keyword.

$$\frac{1}{M} \times \frac{1}{N}$$

M: Number of preparation steps in which the keyword appears
N: Number of video blocks in which the keyword appears

(1) Take one video block and compare all keywords that are contained in it with keywords that are contained in each preparation step.

(2) If both an ingredient name and the verb related to it among the keywords of a video block match a certain preparation step, do the following:

When there is only one matching preparation step: Associate the video block with that preparation step.

When the keywords match two or more preparation steps: Decide that there is a possibility that the video block belongs to multiple preparation steps. A decision that a video block may belong to multiple preparation steps is only made in this case.

(3) If both an ingredient name and the verb related to it among the keywords of a video block do not match a certain preparation step, add the scores of the keywords to the preparation steps that the keywords match, and associate that video block with the preparation step having the highest score among the scores for individual preparation steps that the video block has. If one preparation step with the highest score cannot be determined, associate the video block with the preparation step having the higher score of the preparation steps associated with the preceding and following video blocks. If this decision also cannot be made, associate the video block with the same preparation step as the one with which the preceding video block was associated.

(4) If a decision has been made that the video block may belong to multiple preparation steps, refer to the preparation steps associated with the preceding and following video blocks and permit the association to be made only with a preceding or following preparation step or a preparation step related to a preceding or following preparation step. In other words, the video block may belong to only one preparation step as a result.

(5) Associations are made sequentially beginning with a video block for which the corresponding preparation step can be determined.

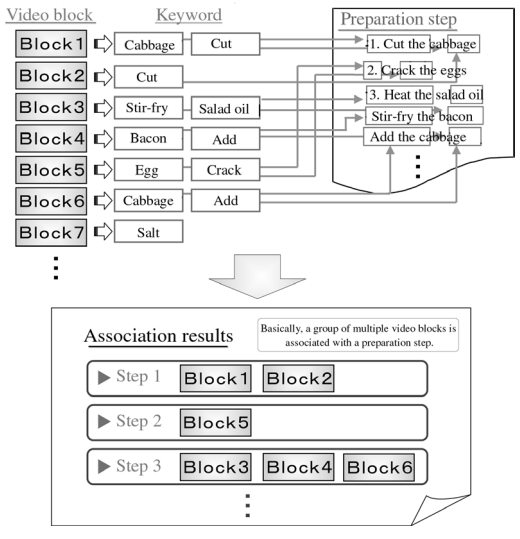


Fig. 8. Example of the association process.

6. Experiments

6.1. Image processing experiments

This section describes experiments for cut detection and face shot detection, which constitute the image processing part of the video structure analysis technique that is performed as preprocessing for association.

6.1.1. Experimental conditions

Table 3 shows the conditions in effect during video capture.

During actual processing, each frame was converted to uncompressed PPM format for use. Also, approximately 100 minutes of video (a total of 600 shots) of a specific cooking program was used for this preprocessing experiment.

6.1.2. Cut detection

Table 4 shows the results of applying the cut detection technique based on DCT clustering. If the number of correctly detected cuts is denoted by N_C, the number of mistakenly detected cuts by N_M, and the number of omissions by N_O, then recall is defined as $N_C/(N_C+N_O)$ and precision is defined by $N_C/(N_C+N_M)$.

From Table 4, it is apparent that cooking video cuts are detected with high precision. Most of the omissions

Table 3. Conditions during video capture

Size	320 (H) × 240 (V)
Number of colors	24 bits
Saved data format	Motion JPEG (compression rate: approximately 1/3)
Sampling rate	10 frames per second

Table 4. Cut detection results

N_C	N_M	N_O	Recall	Precision
568	10	31	94.8%	98.3%

occurred for dissolves, which are cuts in which the image switches while overlapping the preceding and following shots.

These kinds of cuts are also known to be difficult to detect when various other techniques are used. However, in cooking videos, which our current research focuses on as its subject, the shots preceding and following this kind of cut mostly shoot the same subject, and since a scene change almost never occurs at a dissolve cut point, we believe that omissions of this kind of cut are not critical.

6.1.3. Face shot detection

Table 5 shows the face shot detection results. Face shots were detected by assuming that cuts were detected correctly and extracting face regions for the first frame of each of 600 shots. Table 6 shows the face region detection results.

Since face regions were detected according to skin-colored regions, the main cause of mistaken detections was the detection of parts having similar colors such as a wall or a piece of chicken. The main causes of omissions were that the skin-colored regions were too small due to the directions the faces were pointing and that the face regions were not discriminated from larger regions consisting of the faces together with walls having similar colors as the faces.

Although many face region detection omissions occur for full shots (a1) as shown in Table 6, since two or more people appear in many cases for this type of shot, at least one of them was detected, and the shot classification precision was not affected very much as shown in Table 5.

6.2. Association of videos with preparation steps

Finally, we performed experiments for associating videos with preparation steps. For these experiments, we used videos (approximately 150 minutes) for a specific cooking program covering 20 recipes as well as the textbooks corresponding to those videos.

Table 5. Face shot detection results

Type of shot	N_C	N_M	N_O	Recall	Precision
Full shot (a1)	169	24	25	87%	88%
Bust shot (a2)	68	18	20	77%	79%

Table 6. Face region detection results (numbers indicate numbers of faces)

Type of shot	N_C	N_M	N_O	Recall	Precision
Full shot (a1)	480	36	162	72%	92%
Bust shot (a2)	75	22	27	74%	77%

First, we assumed that cut detection and face shot detection, which constitute the image processing part, were performed ideally to evaluate the performance of the association technique independently. Table 7 shows these results.

For "Video → Preparation step" in Table 7, since the association was evaluated based on video blocks, a correct evaluation was when a correct preparation step was associated with the video block. Also, for "Video ← Preparation step," since the association was evaluated based on preparation steps, a correct evaluation was when a video block was always associated with the preparation step without fail. Finally, for "Video ↔ Preparation step," a correct evaluation was when video blocks and preparation steps were associated in both directions with no unassociated video block or preparation step.

Since some video blocks may belong to n preparation steps, a "Correct" value for one preparation step in this case is calculated as $1/n$. Also, a video block corresponding to "Other" in Table 7 is one that is not associated with any preparation step in the textbook such as a video block explaining cooking in general. In this experiment, except when all video blocks are not associated with any preparation steps, this kind of video block is handled as an "Incorrect" video block since it always ends up being associated with some preparation step. The "Success Rate" is defined as (Number of Correct)/(Total number).

At this time, the average numbers of keywords extracted from each video block were 2.3 for ingredient names and 5.1 for verbs, and among these, the numbers that were used for associations were 2.2 for ingredient names and 2.0 for verbs. The average numbers of keywords extracted from each preparation step were 3.8 for ingredient names and 5.6 for verbs, and among these, the numbers that were used for associations were 3.3 for ingredient names and 3.5 for verbs.

Since the proposed technique can associate videos with preparation steps regardless of the text sequence, the associations were created correctly in many cases even when the video flow and the sequence of preparation steps in the textbooks were switched in a complex manner or

Table 7. Association results (independent evaluation of association technique)

Evaluation target	Total number	Correct	Incorrect	Other	Success rate
Video → Preparation step	242	203.5	29.5	9	84%
Video ← Preparation step	94	74	20	-	79%
Video ↔ Preparation step	94	59	35	-	62%

when the video corresponding to one preparation step was divided and appeared in two or more locations.

Next, we used the technique described in Section 3.1 to perform cut detection and face shot detection, which constitute the image processing part, to evaluate the comprehensive performance of the entire system. Table 8 shows these results.

The "Total Number" (number of video blocks) for "Video → Preparation step" differs from the number in Table 7 since the cut detection and face shot detection results contain mistaken detections, and the "Success rate" also decreases mainly due to mistaken detections of video blocks.

Even for this comprehensive experiment, the processing time was a multiple of the video length for cut detection, and otherwise, no processing presented any particular problem in terms of the amount of computations.

6.3. Discussion

First, when the results of the independent evaluation of the association technique (Table 7) are compared with the comprehensive evaluation results (Table 8), it is apparent that the success rate was lower for the comprehensive evaluation than for the independent evaluation. This is because the cut detection and face shot detection results contained mistaken detections. Since cuts were detected with high precision as is apparent from the experimental results shown in Table 4, the main reason the success rate was lower was due to face shot detection.

However, for association based on video blocks ("Video → Preparation step"), the decrease in the success rate was limited to a mere 3%, and the association technique was able to compensate for the mistaken detection of face shots to a certain degree. In other words, we can conclude that the face shot detection technique proposed in Section 3.1.2 exhibited nearly sufficient performance for association based on video blocks. For association based on preparation steps ("Video ← Preparation step"), for which the decrease was 13%, and for association in both directions with no unassociated video block or preparation step ("Video ↔ Preparation step"), for which the decrease was 21%, the head shot detection technique requires further improvement.

Next, when analyzing the cause of failures of the association technique evaluated independently, we found that among video blocks that were mistakenly associated with preparation steps, approximately 30% of these errors occurred because the relevant block contained no prominent keyword and an incorrect preceding or following video block ended up being referenced. Some other causes are given below. However, it generally seems that errors occur because there were insufficient keywords. Therefore, we believe that to increase the success rate, the number of keywords must be increased by creating a dictionary for cooking utensil names to be used as keywords in addition to the ingredient names and verbs.

- Since ingredient names and verbs that belong to the same sentence were treated as related words, unrelated ingredient names and verbs essentially ended up being mistakenly associated.
- Since the specific contents indicated by demonstratives or abstract words such as "vegetables" or "ingredients" were not analyzed, ingredient names could not be selected.

The following means can be considered for dealing with these problems.

- Perform a structure analysis of the text and associate verbs and nouns based on linkage relationships between them.
- Clarify specific contents indicated by demonstratives or abstract words.

Finally, we evaluated the performance of the proposed technique. The proposed technique, which originally aimed to semantically index video, is an association technique based on video blocks. According to the "Video → Preparation step" results in Table 8, the video and preparation steps were associated with a success rate greater than 80%, and it is apparent that the proposed technique is successful for this kind of indexing.

If this kind of indexing is implemented, applications can be considered such as displaying corresponding text preparation steps while the user is viewing the video or creating a video digest by selecting from among the video

Table 8. Association results (comprehensive evaluation of the entire system)

Evaluation target	Total number	Correct	Incorrect	Other	Success rate
Video → Preparation step	222	179.2	31.8	11	81%
Video ← Preparation step	94	62	32	-	66%
Video ↔ Preparation step	94	49	55	-	41%

blocks corresponding to the same preparation step, for example. The performance of the proposed technique can be considered practical for these kinds of applications because a means of reserving better blocks during the selection process can be devised.

The results of association based on preparation steps showed a success rate of approximately 80% when cut detection and shot classification processing were idealized but the success rate was less than 70% for the comprehensive evaluation.

Some applications that can be considered for an association technique based on preparation steps include a cooking support system that presents video corresponding to preparation steps according to the progress of the cooking procedure or that stores video corresponding to cooking steps in a massive database and combines these to generate video corresponding to cooking steps for which no video exists. To use the association technique for a cooking support system, it is sufficient if video blocks always correspond to preparation steps without fail, and a success rate value of approximately 80% can be considered a practical level of precision.

Also, an application that generates video corresponding to preparation steps must associate video blocks and preparation steps in both directions with no unassociated video block or preparation step. However, since the success rates for this case were approximately 60% when cut detection and shot classification processing were idealized and approximately 40% for the comprehensive evaluation, the proposed technique must be improved for this kind of application.

This kind of association based on preparation steps in which video is provided for text can perform its processing in finer units by positioning the proposed technique as the first stage and ultimately creating a more detailed structure for the preparation steps to associate video with individual cooking actions. To achieve this kind of association, structure analysis of the preparation steps in the textbooks [1] and a deeper level of structure analysis of the video are required.

7. Conclusions

In this paper, we proposed a technique for associating cooking videos with preparation steps described in textbooks that are provided with those videos. First, we investigated structure analysis of the videos and proposed and implemented a technique for associating the videos and preparation steps. We also performed evaluation experiments and showed that video could be associated with preparation steps based on video blocks or, in other words, semantic indexing of video could be performed with a high degree of precision by using the proposed technique. We also showed that limiting the subject matter and skillfully incorporating relatively simple elemental techniques enables a practical level of precision to be obtained.

Since the proposed technique is not overly concerned with the text sequence, it is applicable even when the video flow and the sequence of preparation steps in the textbooks are switched in a complex manner or when the video corresponding to one preparation step is divided and appears in two or more locations.

Although the subject matter is limited to cooking videos in this paper, (1) by creating dictionaries that collect together suitable keywords, the proposed technique can also be applied to other kinds of educational TV programs that have external texts or training videos such as for introducing procedures in assembly tasks. Also, even when the structure of these kinds of videos differs from the structure of cooking videos, (2) the proposed technique can be applied by introducing a video structure analysis technique in place of face shot detection.

Some subjects of future research are to make the improvements described in Section 6.3 to further increase the association precision and to increase the precision of individual video structure analysis techniques.

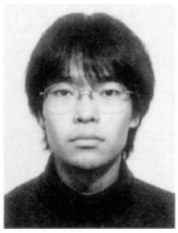
In addition, more detailed video structure analysis and association techniques that can support the assignment of video to text, not just the indexing of video, must be investigated. If these techniques are implemented, not only can semantic indexing of videos in finer units be implemented, but also various applications can be considered such as creating video digests by using association results or generating new easy-to-use multimedia data having a form in which text and video are linked. As the use of computers in the home increases in the future, this kind of research is expected to lead to intelligent cooking support facilities such as video digests used for viewing brief overviews of recipes when selecting recipes or structured videos used as effective educational materials providing exact instructions during actual cooking.

Acknowledgments. The screen images that appear in this paper were selected from within "Rank-Up Cooking" from the Video Database for Evaluating Video Processing that was presented to the public by the Video Data Base Working Group (VDBWG) affiliated with the Pattern Recognition and Media Understanding (PRMU) Technical Group of the Institute of Electronics, Information and Communication Engineers.

REFERENCES

1. Hamada R, Ide I, Sakai S, Tanaka H. Associating cooking video with related textbook. Proc ACM Multimedia 2000 Workshops, p 237–241.
2. Watanabe Y, Okada Y, Tsunoda T, Nagao M. Aligning articles in TV newscasts and newspapers. J Jpn Soc Artif Intell 1997;12:921–927.
3. Yaginuma Y, Sakauchi M. A proposal of a synchronization method between drama image, sound, and scenario document using DP matching. Trans IEICE 1996;J79-D-II:747–755.
4. Smith M, Kanade T. Video skimming and characterization through the combination of image and language understanding. Proc Int Conf Computer Vision, p 61–70, 1998.
5. Iwanari E, Ariki Y. Scene clustering and cut detection using DCT components. Tech Rep IEICE 1994;PRU93-119.
6. Matsuhashi S, Fujimoto K, Nakamura O, Minami T. A proposal of the modified HSV color system suitable for human face extraction. J Inst Telev Eng 1995;49:787–797.
7. Ide I, Yamamoto K, Hamada R, Tanaka H. An automatic video indexing method based on shot classification. Trans IEICE 1999;J82-D-II:1543–1551.
8. Kyoto University Graduate School of Informatics, Department of Intelligence Science and Technology, Speech Media Laboratory. A user-extensible morphological analyzer for Japanese JUMAN Ver. 3.6, 1998.

AUTHORS (from left to right)


Koichi Miura earned his B.S. degree in information and communication engineering from the University of Tokyo in 2001 and completed his master's course in 2003. He holds an M.S. degree (information science and technology). He is engaged in research concerning image analysis and image digests.

Motomu Takano earned his B.S. degree in information and communication engineering from the University of Tokyo in 2003 and is enrolled in his master's course there. His research interest is image analysis.

Reiko Hamada (member) earned her B.S., M.S., and Ph.D. degrees in electronic engineering from the University of Tokyo in 1998, 2000, and 2003 and is currently a research fellow at the Graduate School of Information Science and Technology. Her research interests are natural language processing and multimedia integrated processing. In 2002, she received the 63rd Information Processing Society of Japan National Convention Promotion Award. She is a member of the Information Processing Society of Japan.

AUTHORS (continued) (from left to right)



Ichiro Ide (member) earned his B.E. degree from the University of Tokyo in 1994 and completed his master's course in 1996 and Ph.D. course in electronic engineering in 2000. He became a research associate at the National Institute of Informatics in 2000. Since 2002, he has also been a research associate in mathematical and physical sciences at the Graduate University for Advanced Studies (Sokendai). His research interests are natural language processing and integrated media processing. In 1995, he received the 51st Information Processing Society of Japan National Convention Promotion Award. He is a member of the Japanese Society for Artificial Intelligence, Information Processing Society of Japan, IEEE Computer Society, and Association for Computing Machinery.

Shuichi Sakai (member) earned his B.S. degree from the Department of Information Science at the University of Tokyo in 1981, completed his D.Eng. degree in the Department of Information Engineering in 1986, and became a researcher in the Electrotechnical Laboratory. In 1991–92, he was a visiting scientist at MIT. From 1993 to 1996, he was chief of the Massively Parallel Architecture Laboratory of the Real World Computing Partnership (RWC). From 1996 to 1998, he was associate professor at the Institute of Information Sciences and Electronics at the University of Tsukuba; in 1998, he moved to the Graduate School of Engineering at the University of Tokyo, and has been a professor in the Graduate School of Information Science and Technology since 2001. He is engaged in research on computer systems and applications, especially architecture, parallel processing, scheduling problems, and multimedia. He has received several professional awards including the Information Processing Society's Best Paper Award (1991), Japan IBM Science Award (1991), Ichimura Academic Award (1995), and ICCD Outstanding Paper Award (1995). He is a member of the Information Processing Society of Japan, the Japanese Society for Artificial Intelligence, IEEE, and Association for Computing Machinery.

Hidehiko Tanaka (member) earned his B.E. degree from the University of Tokyo in 1965 and completed his D.Eng. degree in 1970. He became a lecturer in the Engineering Department there in 1970, an associate professor in 1971, and a professor in 1987. Since 2001, he has been a professor and dean in the Graduate School of Information Science and Technology. In 1978–79, he was a visiting professor at the City University of New York. His research interests include computer architecture, parallel processing, natural language processing, media processing, distributed processing, and computer-aided design. He has authored several books including *Non-Neuman Computers*, *Information Communication Systems*, and coauthored *Computer Architecture*, *VLSI Computers I, II,* and *Software Oriented Architecture*. He is a member of the Information Processing Society of Japan, the Japanese Society for Artificial Intelligence, Japan Society for Software Science and Technology, IEEE, and Association for Computing Machinery.