# Automatic Object Model Acquisition and Object Recognition by Integrating Linguistic and Visual Information

Tomohide Shibata
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, 606-8501, Japan
shibata@nlp.kuee.kyoto-u.ac.jp

Norio Kato
Graduate School of Information Science and Technology, University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-8656, Japan
norio@hc.ic.i.u-tokyo.ac.jp

Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, 606-8501, Japan
kuro@i.kyoto-u.ac.jp

## ABSTRACT

In order to make the best use of multimedia contents effectively, the crucial point is the structural analysis of the contents, in which several media processing techniques, including image, audio and text analyses, should be integrated. To understand utterances in videos in accordance with the scene, it is essential to recognize what object appears in the videos. In this paper, we focus on Japanese cooking TV videos, and propose a method for acquiring object models of foods in an unsupervised manner and performing object recognition based on the acquired object models. First, a topic of each video segment is identified based on HMMs to obtain good examples for the object model acquisition. After that, close-up images are extracted from image sequences, and an attention region on the close-up image is determined. Then, an important word is extracted as a keyword from utterances around the close-up image, and is made correspond to the close-up image. By collecting a set of close-up image and keyword from a large amount of videos, object models are acquired. After acquiring the object models, object recognition is performed based on the acquired object models and linguistic information. We conducted experiments on two kinds of cooking TV programs. We acquired the object models of around 100 foods with an accuracy 77.8%. The F measure of object recognition was 0.727.

## Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Applications

*MM'07,* Sepetember 23–28,2007,Augsburg,Bavaria,Germany.

## General Terms

Algorithms

## Keywords

video indexing, object model acquisition, object recognition

## 1. INTRODUCTION

Recent years have seen the rapid increase of multimedia contents with the continuing advance of information technology. In order to make the best use of multimedia contents effectively, the crucial point is the structural analysis of the contents, annotating what each scene contains and explains and how those scenes are related. Such information is surely useful for the video retrieval and summarization. Although the standardization for formats of such information has been actively discussed, such as MPEG7, their automatic annotation is still far from practical use.

Most of traditional approaches to video processing, including video browsing, indexing, classification and summarization, utilized only image analyses, such as cut detection, edge detection and face recognition [11]. In order to overcome the limitation of the approaches relying only on the image information, some studies have recently integrated several media processing techniques, including image, audio and text analyses [14].

Our approach to the automatic indexing is based on the precise understanding of utterances in videos using Natural Language Processing techniques. Since, there is, of course, a definite ceiling to understanding utterances in the videos without referring to visual information, information should be extracted from videos, that is, it is essential to recognize what object appears or what action is performed in the videos. Then, the understanding of utterances has to be performed in accordance with the scene.

It is, however, difficult for the current image processing technique to extract such information in the videos unless detailed object/action models for a specific domain are constructed by hand. Most previous work on object recognition used a lot of sets of an image and some manually-annotated keywords, and learned the correspondence between a region in the image and a keyword using the EM algorithm [1].

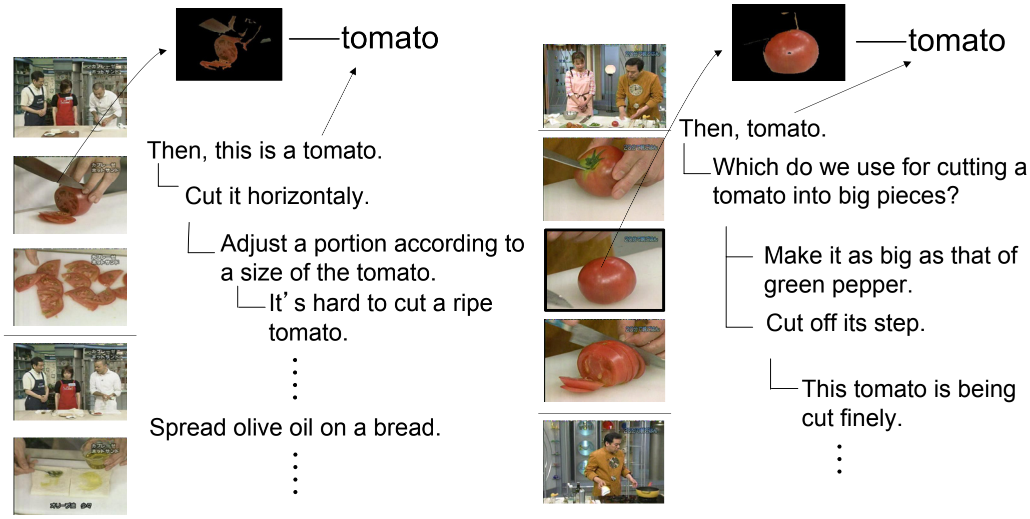

tomato
Then, this is a tomato.
Cut it horizontaly.
Adjust a portion according to a size of the tomato.
It' s hard to cut a ripe tomato.
Spread olive oil on a bread.
tomato
Then, tomato.
Which do we use for cutting a tomato into big pieces?
Make it as big as that of green pepper.
Cut off its step.
This tomato is being cut finely.

**Figure 1: An overview of collecting sets of an attention region and a keyword.**

In such work, however, sets of image and some manually-annotated keywords need to be prepared, which costs too high.

This paper presents a method for acquiring object models given a large amount of videos in a specific domain. Then, object recognition is performed using the acquired object models. Among several types of videos, instruction videos (*how-to* videos) about sports, cooking, D.I.Y., and others are the most valuable since video contents are very suitable for the explanation of actions. In particular, we focus on Japanese cooking TV programs.

In the case of instruction videos, there are concrete references to each object, and thus the color/shape of each object can be learned by collecting the concrete references. In the left part in Figure 1, from the utterance "This is a tomato." when the close-up image of "tomato" appears, we can easily imagine that "tomato" appears in this image. In this close-up image, the region that draws the most attention is extracted, and by making the region correspond to the word "tomato", we can learn that the color of "tomato" is red. Since there may be an analysis error when the color/shape of each object is learned from one set of image and keyword, such learning is performed from a large amount of sets, which leads to stable learning.

As the first step in the object model acquisition, this paper presents a method for acquiring the color information (e.g., RGB) of foods. Hereafter, we call an image in which a food appears in close-up **close-up image**, a region that draws the most attention in the image **attention region**, and a word that is the most important among utterances **keyword**. Since objects change their shape/color along with the progress of cooking, in order to obtain good examples for the object acquisition, a topic of video segments is automatically identified, and then sets of an image and a keyword are collected only from segments whose topic is identified as *preparation*. After that, close-up images are extracted from image sequences by edge detection, and in the extracted close-up image, an attention region is determined considering several features, such as the area of a region and the center of gravity of a region. A keyword is extracted from utterances when the close-up image appears. This is performed by considering an utterance type and discourse structure, which are obtained by the linguistic analysis we previously proposed [13].

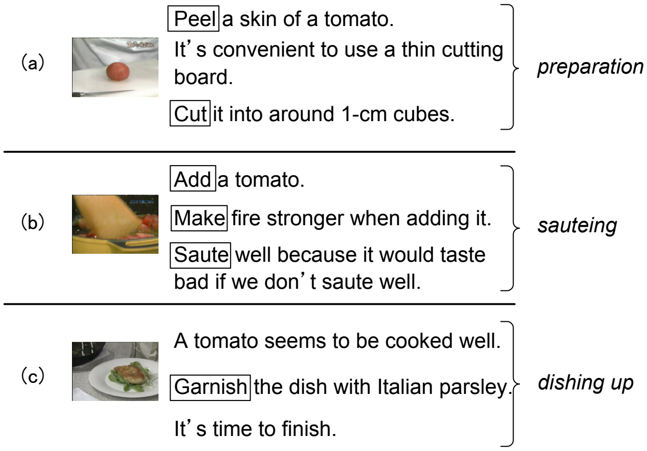

(a)
Peel a skin of a tomato.
It' s convenient to use a thin cutting board.
Cut it into around 1-cm cubes.
preparation
(b)
Add a tomato.
Make fire stronger when adding it.
Saute well because it would taste bad if we don' t saute well.
sauteing
(c)
A tomato seems to be cooked well.
Garnish the dish with Italian parsley.
It' s time to finish.
dishing up

**Figure 2: Examples of images and their topics (The words surrounded by a rectangle represents extracted utterances utilized for the topic identification.).**

After acquiring the color information of each object, object recognition is performed based on the acquired object models and the word importance, which are determined considering the utterance type and the discourse structure.

## 2. TOPIC IDENTIFICATION BASED ON HMMS

In the case of cooking, objects (i.e. ingredient) change their shape/color along with the progress of cooking. Consequently, good examples for the object acquisition cannot be collected from video segments whose topic is *sauteing* or *dishing up*. Therefore, we can expect that the accuracy of

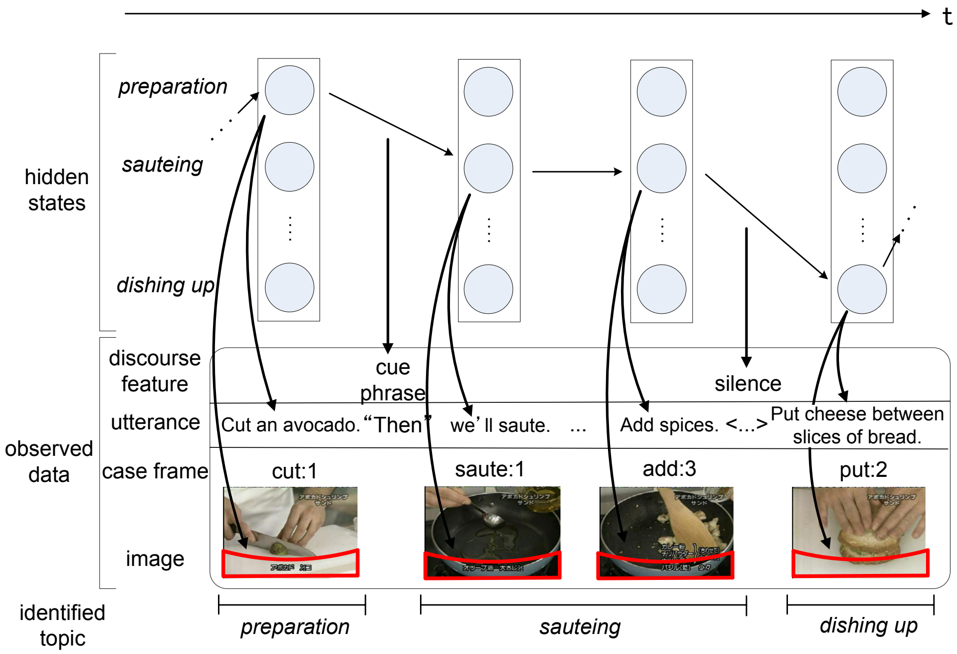


**Figure 3: Topic identification with Hidden Markov Models.**

the object model acquisition would be improved by collecting sets only from segments whose topic is *preparation*. In the example shown in Figure 2, while image (a), whose topic is identified as *preparation*, is utilized for the object model acquisition, image (b), whose topic is identified as *sauteing*, and image (c), whose topic is identified as *dishing up*, are discarded.

This paper identifies an topic of each segment in an unsupervised manner by the method we previously proposed [12], as illustrated in Figure 3. HMMs are employed for topic identification, wherein a state corresponds to a topic, like *preparation* and *frying*, and various features, including visual and audio information as well as linguistic information (instructor's utterances), are observed. This study considers a clause as a unit of analysis and the following eight topics as a set of states: *preparation*, *sauteing*, *frying*, *baking*, *simmering*, *boiling*, *dishing up*, *steaming*. We utilize visual and audio information to achieve robust topic identification. As for visual information, we can utilize background color distribution of the image. As for audio information, silence can be utilized as a clue to a topic shift.

## 2.1 Features for Topic Identification

In this section, features utilized for the topic identification are described. They consist of three modalities: linguistic, visual and audio modality.

### 2.1.1 Linguistic Features

Closed captions of Japanese cooking TV programs are used as a source for extracting linguistic features. We first process them with the Japanese morphological analyzer, JUMAN [8], and make syntactic/case analysis with the Japanese analyzer, KNP [7]. Then, we perform the following processes to extract linguistic features, including case frames, cue phrases, noun chaining, and verb chaining.

**Action extraction in the form of case frame**

**Table 1: Utterance type classification. (An underlined phrase represents a pattern for recognizing utterance type.)**

**[action declaration]**
ex. Then, we 'll cook a steak.
**[individual action]**
ex. Cut off a step of this eggplant.
**[food state]**
ex. There is no water in the carrot.
**[note]**
ex. Don't cut this core off.
**[substitution]**
ex. You may use a leek.
**[food/tool presentation]**
ex. Today, we use this handy mixer.
**[small talk]**
ex. Hello.

Instructor's utterances can be divided into various types such as actions, tips, and even small talk. Among them, actions, such as cut and peel, are dominant and supposed to be useful for the topic identification and others can be noise. Therefore, considering a clause as a basic unit, utterances referring to an action are extracted in the form of case frame, which is assigned by case analysis. Extracting them in the form of case frame is for generalization and word sense disambiguation. For example, "salt *wo* ireru (add salt)" and "sugar *wo* pan *ni* ireru (add sugar into a pan)" are assigned to case frame ireru:1 (add) and "knife *wo* ireru (carve with a knife)" is assigned to case frame ireru:2 (carve).

To extract utterances referring to actions, we classify utterances[1] into several types listed in Table 1[2]. Input sen-

[1] In this paper, [ ] means an utterance type.
[2] Actions are supposed to have two levels: [action declaration] means a declaration of beginning a series of actions and

**Table 2: Examples of the automatically constructed case frame.**

| Verb | Case marker | Examples |
|---|---|---|
| *kiru*:1<br>(cut) | *ga* | <agent> |
| | *wo* | pork, carrot, vegetable, ··· |
| | *ni* | rectangle, diamonds, ··· |
| *kiru*:2<br>(drain) | *ga* | <agent> |
| | *wo* | damp ··· |
| | *no* | eggplant, bean curd, ··· |
| *ireru*:1<br>(add) | *ga* | <agent> |
| | *wo* | salt, oil, vegetable, ··· |
| | *ni* | pan, bowl, ··· |
| *ireru*:2<br>(carve) | *ga* | <agent> |
| | *wo* | knife ··· |
| | *ni* | fish ··· |

(*ga*: nominative, *wo*: accusative, *ni*: dative)

tences are first segmented into clauses and their utterance type is recognized. Utterance types can be recognized by clause-end patterns[3]. As for [individual action] and [food state], considering the portability of the system, we use general rules regarding intransitive verbs or adjective + "become" as [food state], and others as [individual action]. After recognizing utterance types, we extract utterances whose utterance type is recognized as action ([action declaration] or [individual action]).

In general, a verb has multiple meanings/usages. For example, Japanese verb "ireru" has multiple usages, "salt *wo* ireru (add salt)" and "knife *wo* ireru (carve with a knife)," which appear in different topics. Considering this point, we do not extract a surface form of verb but a case frame, which is assigned by case analysis. Case frames are automatically constructed from Web cooking texts (12 million sentences) by clustering similar verb usages [6]. Examples of the automatically constructed case frame are shown in Table 2.

**Cue phrases**

As Grosz and Sidner [4] pointed out, cue phrases such as *now* and *well* serve to indicate a topic change. We use approximately 20 domain-independent cue phrases, such as "then" and "next".

**Noun Chaining**

When two continuous actions are performed to the same ingredient, their topics are often identical. For example, because "grate" and "raise" are performed to the same ingredient "turnip," the topics (in this instance, *preparation*) in the two utterances are identical.

(1) a. We'll grate a turnip.
b. Raise this turnip on this basket.

**Verb Chaining**

When a verb of a clause is identical with that of the previous clause, they are likely to have the same topic. We utilize the fact that the adjoining two clauses contain an identical verbs or not as an observed feature.

(2) a. Add some red peppers.
b. Add chicken wings.

### 2.1.2 Image Features

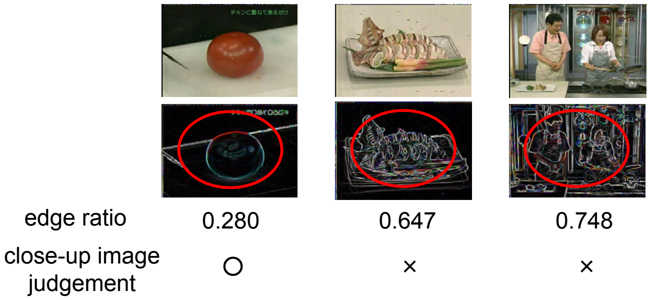


**Figure 4: Close-up image judgement based on the edge detection.**

It is difficult for the current image processing technique to extract what object appears or what action is performing in video unless detailed object/action models for a specific domain are constructed by hand. Therefore, referring to [5], we utilize color distribution at the bottom of the image, which is comparatively easy to exploit. This information can capture the tendencies that, for example, *frying* and *boiling* are usually performed on a gas range and *preparation* and *dishing up* are usually performed on a cutting board, which can be an aid to topic identification. As shown in Figure 3, we utilize the mass point of RGB in the bottom of the image at the beginning of each clause, which is a basic unit of linguistic analyses.

### 2.1.3 Audio Features

As Galley et al. [3] pointed out, a longer silence often appears when a topic changes, which we can utilize as a clue to topic change. In this study, silence is automatically extracted by finding duration below a certain amplitude level that lasts more than one second.

## 2.2 Parameters Estimation

HMMs are employed for topic identification, where a hidden state corresponds to a topic and various features described in the previous section are observed. In this model, considering a case frame as a basic unit, a case frame and background image are observed from a state, and discourse features indicating to topic shift/persistence (cue phrases, noun/verb chaining and silence) are observed when a state transits. HMM parameters, including initial state distribution, state transition probability, and observation probability, are estimated using the Baum-Welch algorithm. Once the HMM parameters are trained, the topic identification is performed using the standard Viterbi algorithm.

# 3. OBJECT MODEL ACQUISITION

## 3.1 Close-up Image Extraction and Attention Region Determination

Close-up images are extracted only from segments whose topic is identified as *preparation*, and then in the close-up image, an attention region is determined. In this study, images are extracted from a video in every second. To extract a close-up image in which a food appears for a certain period of time, videos are divided into shots beforehand, which are considered a basic unit for image processing. A shot boundary is detected using the color histogram difference [9]. If

[individual action] means an action that is the finest one.

[3] We prepare approximately 500 patterns.

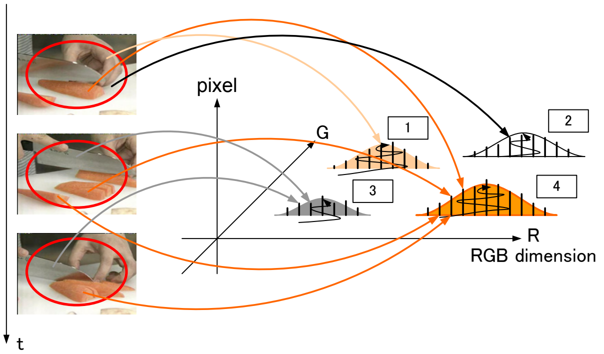


**Figure 5: Mapping the pixels to the RGB dimension and searching for the maximum point with the hill-climbing method.**

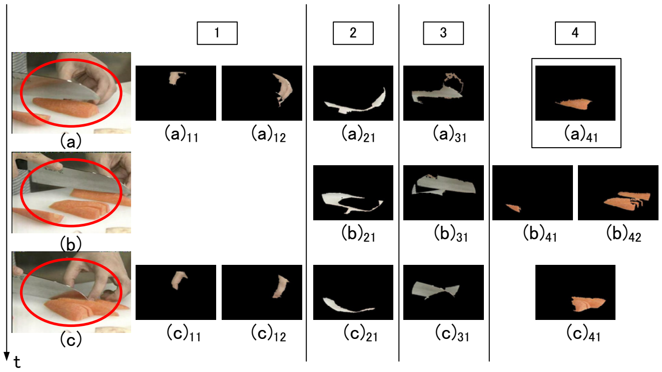


**Figure 6: Region segmentation with labeling.**

the color histogram difference exceeds a certain threshold, this point is regarded as a cut point. Furthermore, at the cut point, a face detector by neural networks is applied in order to exclude face shots in the object model acquisition. If a face is detected at a cut point (the first frame of a shot), the shot is removed from the following image processing.

### 3.1.1 Close-up Image Extraction by Edge Detection

Close-up images are extracted by edge detection. We use the Sobel gradient operators to detect edges, and calculate *edge ratio*, the number of pixels where an edge is detected divided by all the pixels. Note that this ratio is calculated only in the ellipse as shown in Figure 4 since a food rarely appears in the edge of the frame. The image whose edge ratio is lower than a threshold ($Th_{edge} = 0.5$) is extracted as a close-up image. By this treatment, both the image in which a food does not appear in close-up and the image in which a face cannot be detected by the method described in the previous subsection can be removed.

### 3.1.2 Region Segmentation

In the extracted close-up image, to determine an attention region, region segmentation is performed as follows (Figure 5, 6):

1. Considering a shot as a basic unit, on close-up images, pixels in an ellipse are mapped to the RGB dimension. The ellipse is equal to the one shown in the previous subsection.
2. The image is smoothed by using the median filter with a 3 * 3 mask.
3. Local maximum points are found with the hill-climbing search. In the example shown in Figure 5, four local maximum points are found.
4. In the original images, by performing pixel labeling for each maximum point, region segmentation is performed. In the example shown in Figure 6, four regions are obtained: a hand region, a cutting board region, a knife region, and a carrot region.

### 3.1.3 Attention Region Determination

An attention region tends to be larger one, to be near the center of the image, to be one in which pixels are dense around the center of gravity of the region, and etc. For each region an evaluation score is calculated based on features of the region, and then the region that has the highest score is considered an attention region. Here, we have to be careful not to consider a hand or a knife region to be an attention region, even if such region is near the center of the image.

The region that satisfies with one of the following conditions is not considered an attention region:

- Opticalflow

  A region in which a quick motion is detected tends not to be a food but a hand or a knife. Therefore, an opticalflow value in the region is calculated, and the region in which the average opticalflow value in a shot is greater than a threshold is not considered an attention region. This study adopts the block matching algorithm to detect an opticalflow value. In the example shown in Figure 6, since the average opticalflow value is greater than a threshold, region 3 is excluded.
- Ratio of pixels in the upper half of the image

  A food region is rarely located only in the upper half of the image. If a ratio of pixels in the upper half of the image is greater than a threshold ($Th_{upperratio} = 0.95$), this region is excluded.
- Hand region

  So as not to extract a hand region as an attention region, a region whose color is judged as flesh color is not extracted as an attention region. By converting the representative RGB of the region to a modified HSV colorspace by the method proposed by Matsuhashi et al. [10], the flesh color judgement is performed. By our exploratory experimental, if the color $(H, S)$ of a region is in $20 \le H \le 35, 40 \le S \le 65$, the region is judged as flesh color region. However, if we simply exclude regions whose color is flesh, a region of food whose color is flesh such as burdock and bonito can be excluded. Therefore, considering a hand is often located in the upper side of the frame, a region whose color is flesh and its upper ratio is greater than a threshold ($Th_{upperratiohand} = 0.8$) is excluded. In the example shown in Figure 6, regions $(a)_{11}$, $(a)_{12}$, and $(c)_{11}$ are excluded.
- Ratio of border tangent to the ellipse

  If a ratio of border tangent to the ellipse is greater than a threshold ($Th_{boundingratio} = 0.35$), the region

Table 3: Features for each region.

| | |
|---|---|
| $S$: | area |
| $G_{dis}$: | average distance between a pixel and the center of gravity |
| $C_{dis}$: | average distance between a pixel and the center of the frame |
| $Circularity$: | an indicator how a region is equal to a circle. This value is given by $\frac{4\pi S}{l^2}$ using a boundary length $l$ and $S$. |
| $Rectangularity$ : | an indicator how a region is equal to a rectangle. This value is given by $\frac{S}{S_{rec}}$ using $S$ and a bounding rectangle area $S_{rec}$. |

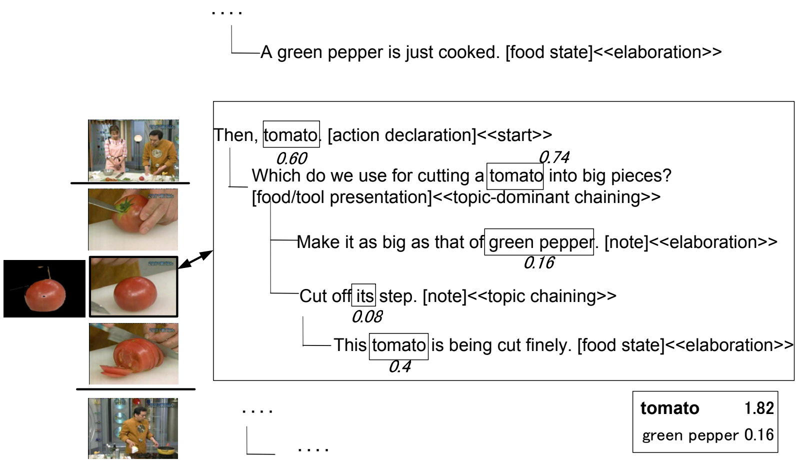

....
A green pepper is just cooked. [food state]<<elaboration>>
Then, tomato. [action declaration]<<start>>
0.60
0.74
Which do we use for cutting a tomato into big pieces?
[food/tool presentation]<<topic-dominant chaining>>
Make it as big as that of green pepper. [note]<<elaboration>>
0.16
Cut off its step. [note]<<topic chaining>>
0.08
This tomato is being cut finely. [food state]<<elaboration>>
0.4
....
....
tomato 1.82
green pepper 0.16

Figure 7: Keyword extraction based on discourse structure analysis.

is excluded. This is because such region is not likely to be a food. In the example shown in Figure 6, region $(c)_{21}$ is excluded.

In order to determine an attention region among several regions that does not satisfy with any above conditions, the features listed in Table 3 are calculated for each region, and then the following score is calculated for each region:

$$0.1 \cdot S + 0.3 \cdot max(Circularity, Rectangularity) - 0.3 \cdot C_{dis} - 0.5 \cdot G_{dis}. \quad (1)$$

The coefficients in this formula are determined so that a region that is larger, is more similar to circle/rectangle, is nearer the center of the image, and is one in which pixels are denser around the center of gravity of the region is extracted as an attention region.

When multiple regions belonging to the same maximum point are extracted in an image, the region that has the maximum score is considered a representative region. In image $(b)$ in Figure 5, two regions belonging to maximum point 4 are obtained, and region $(b)_{42}$, which has greater score, is extracted as a representative region.

After that, the scores of each region are added on in a shot for each maximum point, and the region that has the highest score is extracted as an attention region. In the example shown in Figure 5, the region belonging to maximum point 4 is extracted as an attention region. The image whose edge ratio is the lowest in a shot is considered the representative frame, and the attention region of that frame is adopted as the learning data for acquiring an object model. In Figure 5, region $(a)_{41}$ of image $(a)$ whose edge ratio is the lowest in the shot is extracted as an attention region.

## 3.2 Keyword Extraction

This section describes a method for extracting an important word as a keyword from instructor's utterances when the close-up image appears, and making it correspond with the close-up image. Basically, in an utterance that is close to a close-up image, there is reference to a food that appears in the image. However, as shown in the following examples, there is sometimes reference to the food on which no action is currently performed.

(3) You may use a leek.

(4) It's almost the same size as the onion.

Therefore, a word that is the most important in a certain range should be extracted as a keyword. To set this certain range, discourse structure of closed caption texts is analyzed by the method we previously proposed [13], and then segmentation is performed. That is, since discourse structure analysis distinguishes sets of utterances that have strong coherence relations such as reason, elaboration, and those that do not have such relations, the latter can be considered a segmentation point.

By discourse structure analysis, a structure as shown in Figure 7 is obtained. In this figure, a phrase sandwiched by a square bracket represents an utterance type and a phrase sandwiched by a double parentheses represents a coherence

**Table 4: Functions for calculating the word importance.**

| | |
|---|---|
| $f_{utype}(w_i)$: | 1 if utterance type is action, food/tool presentation, or food state, 0.1 if utterance-type is substitution, 0.3 otherwise. |
| $depth(w_i)$: | a depth in a discourse structure. Units found closer to the root of a discourse structure tree are considered to be more important than those found at lower levels in the tree. |
| $f_{clause}(w_i)$: | 1 if $w_i$ is in a main clause, 0.5 if $w_i$ is in a subordinate clause. |
| $f_{topic}(w_i)$: | 1.5 if $w_i$ is marked with a topic marker, 1 otherwise. |
| $f_{anaphora}(w_i)$: | 0.5 if $w_i$ is the anaphora resolution result, 1 otherwise. |
| $f_{time}(F_{w_i}, F_{image})$: | this score is defined using frame number $F_{w_i}$ of an utterance that contains word $w_i$ and frame number $F_{image}$ of a image by $f_{time}(F_{w_i}, F_{image}) = 1 - \frac{\lvert F_{w_i} - F_{image}\rvert}{F_{th}}$. Here, $F_{th}$ is set to 1000. |

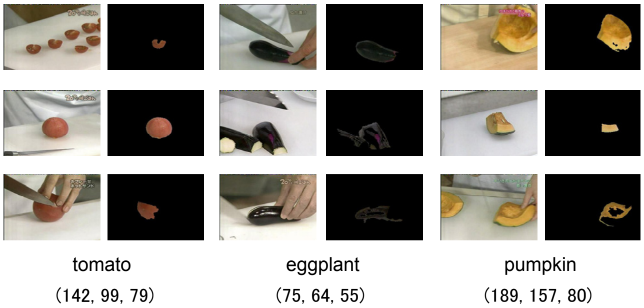


**Figure 8: Examples of collected sets of an original image, an attention region and a keyword and acquired object models.**

relation to the parent sentence. A shot to which a close-up image belongs is matched with the subtree of discourse structure that is closest to the shot. In the subtree of discourse structure the most important word is extracted as a keyword considering several linguistic features, such as utterance type and depth in discourse structure.

For word $w_i$ that has the semantic primitive <food> in a thesaurus, an importance score is calculated by the following equation:

$$Score(w_i) = \sum_{w_i \in Tree} f_{utype}(w_i) \cdot \frac{1}{\sqrt{depth(w_i)}} \cdot f_{clause}(w_i) \cdot f_{topic}(w_i) \cdot f_{anaphora}(w_i) \cdot f_{time}(F_{w_i}, F_{image}) \quad (2)$$

Here, each function is shown in Table 4. In a subtree, the word that has the highest score is extracted as a keyword. For example, in Figure 7, by summing scores for each food in the subtree of the discourse structure, "tomato" obtains 1.82 points and "green pepper" obtains 0.16 points. Therefore, "tomato," which obtains the highest score, is extracted as a keyword.

## 3.3 Object Model Acquisition

Through the processes described in the previous sections, sets of an attention region and a keyword can be collected. Examples of collected sets are shown in Figure 8. In this figure, the left row represents an original image and the right row represents an attention region that is extracted from the image. For each food, RGB histograms of an attention region are added on, and then the average of the most frequent RGB is considered an object model.

# 4. OBJECT RECOGNITION

As illustrated in Figure 9, based on the acquired object models, object recognition is performed according to the following procedure:

1. Considering the shot to which a target image belongs as a basic unit, region segmentation is performed on the images in the shot using the same procedure as the object model acquisition.

2. Although in the object model acquisition the search target is the nearest discourse structure, in object recognition the search targets are the nearest discourse structure, the previous one and the next one.

3. For a region $R_k$ and a word $W_i$ in the discourse structure, the evaluation score $score(R_k)$ of $R_k$ and the importance score $score(W_i)$ of $W_i$ are calculated using the same formula as the case of the object model acquisition. Then, the Euclidean distance $distance(R_k, model(W_i))$ between the representative color of $R_k$ and the object model of $W_i$ is calculated, and then the set of $R_k$ and $W_i$ that has the highest score given by the following formula is found.

   $$\underset{R_k, W_i}{\text{argmax}} \frac{score(R_k) \cdot score(W_i)}{distance(R_k, model(W_i))} \quad (3)$$

   If the highest score is greater than a threshold, the word $W_i$ is adopted as an object recognition result.

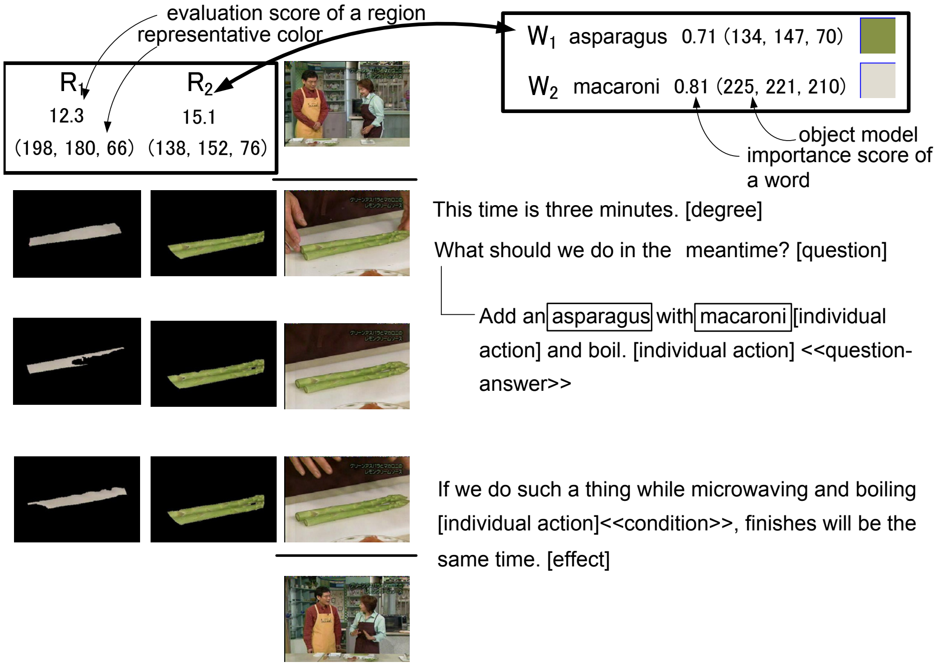
In the example shown in Figure 9, since the set of the word "asparagus" and the region $R_2$ has the highest score, the word "asparagus" is adopted as an object recognition result.

# 5. EXPERIMENTS

To demonstrate the effectiveness of our proposed method, we made experiments of the object model acquisition and object recognition on two kinds of Japanese cooking TV programs: NHK "Today's Cooking" and NTV "Kewpie 3-Min Cooking".

## 5.1 Object Model Acquisition

The object model acquisition was performed from around 200 videos of NHK "Today's Cooking" (around 83 hours) and around 70 videos of NTV "Kewpie 3-Min Cooking" (around 12 hours).


evaluation score of a region
representative color
$R_1$
12.3
(198, 180, 66)
$R_2$
15.1
(138, 152, 76)
$W_1$ asparagus 0.71 (134, 147, 70) 🟩
$W_2$ macaroni 0.81 (225, 221, 210) ⬜
object model
importance score of a word

This time is three minutes. [degree]

What should we do in the meantime? [question]

Add an asparagus with macaroni [individual action] and boil. [individual action] <<question-answer>>

If we do such a thing while microwaving and boiling [individual action]<<condition>>, finishes will be the same time. [effect]

**Figure 9: An overview of object recognition.**

**Table 5: Experimental result of object model acquisition.**

| topic identification | accuracy(%) |
|---|---|
| | 64.8 (70 / 108) |
| √ | **77.8 (84 / 108)** |

**Table 6: Object model acquisition accuracy by the collected set of attention region and keyword.**

| # of collected samples | accuracy |
|---|---|
| more than one | 81.9% (68 / 83) |
| more than two | 87.5% (56 / 64) |
| more than four | 94.6% (35 / 37) |

The accuracy of the object model acquisition is shown in Table 5. The object models of approximately 100 foods were automatically acquired with an accuracy of 0.778. Table 5 also shows that the accuracy was improved by 13.7% by performing the topic identification, which demonstrates the effectiveness of the topic identification in this task.

The object model acquisition accuracy by the collected set of an attention region and a keyword is shown in Table 6. This table shows that the more the collected sets, the more stable the object model learning is. Therefore, we can expect that the accuracy of the object model acquisition will get higher when the number of available videos increases.

Causes of the object model acquisition errors are described. As shown in Figure 10, collected sets of an attention region and a keyword are classified into three types: (a) success, (b) failure of the attention region extraction, and (c) the food (keyword) does not appear in the image. Examples of case (b) and (c) are shown below.

**failure of attention region extraction**

- a food region is merged into a cutting board region

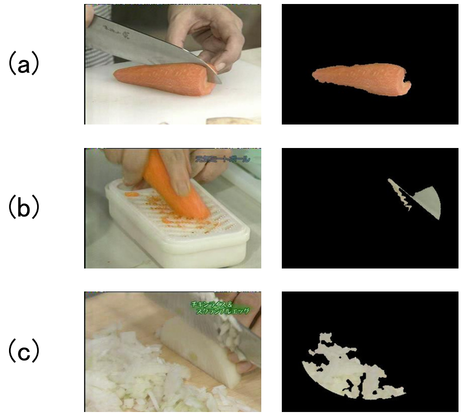


**Figure 10: Examples of collected sets of an attention region and a keyword "carrot": (a) success, (b) attention region extraction failure, (c)the food (keyword) does not appear in the image.**

  In the case of whitish foods such as Chinese cabbage and onion, since a food region was merged to a cutting board or a background region in smoothing, an attention region could not be correctly determined (Figure 11).

- a food region is removed as a moving region

  Although in some examples a moving region such as hand and knife could be correctly removed by calculating the opticalflow value, there were some cases in which a food region was wrongly excluded as a moving region. In Figure 12, a lemon region is excluded as a moving region and a cutting board region was incorrectly judged as an attention region.

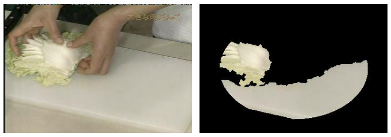


**Figure 11: An example of failure of attention region extraction.**

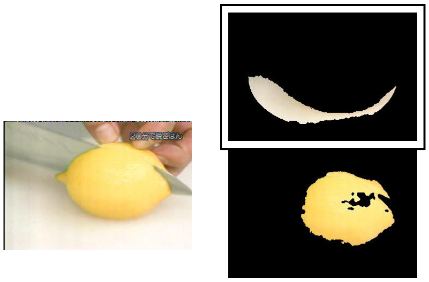


**Figure 12: An example of failure of attention region extraction (The image surrounded by a rectangle represents the attention region that is incorrectly determined.).**

**the food (keyword) does not appear in the image**

- topic identification failure

  Some examples, whose topic is not actually *preparation*, were utilized for the object model acquisition, which is caused by the topic identification failure. In Figure 13, although the topic of this example is actually *boiling*, the topic was incorrectly identified as *preparation*. Consequently, in the case of that an image where the keyword "mushroom" has been extracted from utterances, but in which no mushroom appears, a mushroom region tried to be extracted. As a result, a pan region was incorrectly made corresponded to the keyword "mushroom".

### 5.2 Object Recognition

We conducted experiments of object recognition with the acquired object model on each five videos of "Today's Cooking" and "Kewpie 3-Min Cooking". Correct object recognition results were manually labeled to each shot, and precision, recall, and F measure were evaluated. Note that only the shots before foods are cooked were targets of evaluation.

The accuracy of the proposed object recognition method and the following baselines is shown in Table 7.

**only color information** This baseline considers a food whose object model is the most similar to the representative color of a target image as an object recognition result.

**the food referred the most frequently** This baseline does not utilize color information, and considers the food that is referred the most frequently in the nearest discourse subtree as an object recognition result.

Table 7 shows that the F measure of object recognition was 0.727, which was higher than the baseline that utilizes only color information and that utilizes only linguistic information.

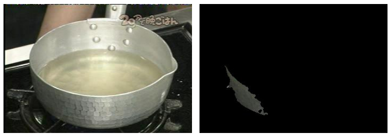


**Figure 13: An example in which the food (mushroom) does not appear in the image.**

Most errors of object recognition were caused by the object model acquisition error. For example, since the object models of "chives" and "eel" were not correctly acquired, they could not be correctly recognized. Since, as described above, the accuracy of the object model acquisition will get higher when the number of available videos increases, we can expect the accuracy of object recognition will also get higher.

Another problem is that this study assumes that each food has one object model. For example, for "leek," since only the object model of "white leek" was acquired, "green leek" cannot be recognized with the acquired model. To handle this problem, multiple object models for each food will be acquired by clustering the color distribution. Similarly, when color of skin is different from flesh, we assume that the object model is correctly acquired if either of them is acquired. Therefore, in the case of "eggplant," for example, since only the color of its skin was acquired, "eggplant" cannot be recognized on the image in which its flesh appears. We are planning to acquire the color of both skin and flesh. Furthermore, we are planning to acquire the information of the way the color changes after it is cooked.

## 6. RELATED WORK

To the best of our knowledge, there is not previous work that attempted to acquire object models from a large amount of videos in an unsupervised manner. Most previous work on object recognition used a lot of sets of image and some manually-annotated keywords, and learned the correspondence between a region in the image and a keyword using the EM algorithm. Duygulu et al. proposed a model of object recognition as machine translation [1]. In their model, recognition is a process of annotating image regions with words. Firstly, images are segmented into regions, which are classified types and keywords supplied with the images, is then learned, using the EM algorithm. This process is analogous with learning a lexicon from an aligned bi-text. Feng et al. explored the use of bootstrapping approach to annotating large image collection [2]. Their idea is to start from a small set of labeled training examples, and successively annotate a large set of unlabeled examples with the co-training approach.

Yanai proposed a generic image classification system with an automatic knowledge acquisition mechanism from the World Wide Web [16]. Their system gathers a large number of images from the Web automatically, and makes use of them as training images for generic image classification. They reported that although classification rate obtained in the experiments for generic real world images is not high and not sufficient for practical use, the experimental results suggest that generic image classification using visual knowledge on the WWW is one of the promising ways for resolving real world image recognition/classification.

In the cooking domain, Takano et al. proposed an object

**Table 7: Experimental result of object recognition.**

| | Precision | Recall | F |
|---|---|---|---|
| proposed | 100 / 132 (75.8%) | 100 / 143 (69.9%) | 0.727 |
| baseline | | | |
| only color information | 75 / 132 (56.8%) | 75 / 143 (52.4%) | 0.545 |
| the food referred the most frequently | 90 / 138 (64.4%) | 90 / 143 (62.9%) | 0.641 |

recognition method [15], which focuses on 11 foods. Their method first collects by hand several images in which each food appears from cooking videos, and obtains the color distribution information of each food. Then, on the target image, the area in which Mahalanobis distance between its color and the color distribution is lower than a threshold is extracted, and the confidence score of each food is calculated based on the extracted area. The food that has the maximum confidence score is considered an object recognition result. They conducted an experiment on segments before foods are cooked, and achieved recall of 74.8% and precision of 78.4%. Our proposed method corresponds to automatically collect several images in which each food appears and detect a food region in the collected image, which are manually done in Takano's work. Although the accuracy of our object recognition method is a little lower than one of Takano's work, considering that our method does not require manual operations and can automatically acquire the object models of approximately 100 foods, we can say that our method achieved the same accomplishment as, or more than, Takano's work.

## 7. CONCLUSIONS

This paper first described a method for automatically acquiring object models from large amounts of video. Then, we presented a method for object recognition using the acquired object models and linguistic information.

We are planning to acquire the color of both skin and flesh, and the information of the way the color of each food changes after it is cooked.

## 8. REFERENCES

[1] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision(ECCV)*, pages 97–112, 2002.

[2] H. Feng and T.-S. Chua. A bootstrapping approach to annotating large image collection. In *ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 55–62, 2003.

[3] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, 7 2003.

[4] B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistic*, 12:175–204, 1986.

[5] R. Hamada, I. Ide, S. Sakai, and H. Tanaka. Associating cooking video with related textbook. In *Proceedings of ACM Multimedia 2000 workshops*, pages 237–241, 2000.

[6] D. Kawahara and S. Kurohashi. Fertilization of case frame dictionary for robust japanese case analysis. In *Proceedings of 19th COLING (COLING02)*, pages 425–431, 2002.

[7] S. Kurohashi and M. Nagao. A syntactic analysis method of long japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4), 1994.

[8] S. Kurohashi, T. Nakamura, Y. Matsumoto, and M. Nagao. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language*, pages 22–28, 1994.

[9] R. Lienhart. Comparison of automatic shot boundary detection algorithms. In *Proceedings of SPIE Conf. on Storage and Retrieval for Image & Video Databases VII*, volume 3656, pages 290–301, 1998.

[10] S. Matsuhashi, O. Nakamura, and T. Minami. Human-face extraction using modified hsv color system and personalidentification through facial image based on isodensity maps. *IEEE CCGEI '95*, 2(2):909–912, 1995.

[11] A. Rosenfeld, D. Doermann, and D. DeMenthon, editors. *VIDEO MINING*. Kluwer Academic Publishers, 2003.

[12] T. Shibata and S. Kurohashi. Unsupervised topic identification by integrating linguistic and visual information based on hidden markov models. In *Proceedings of The Joint 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL2006, poster)*, pages 755–762, 2006.

[13] T. Shibata, M. Tachiki, D. Kawahara, M. Okamoto, S. Kurohashi, and T. Nishida. Structural analysis of instruction utterances using linguistic and visual information. In *Proceedings of Eighth International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES2004)*, pages 393–400, 9 2004.

[14] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, 2006.

[15] M. Takano, K. Miura, R. Hamada, I. Ide, S. Sakai, and H. Tanaka. Object detection from cooking video by restriction from acompanying text information. In *Proceedings of JSAI 17th Annual Convention*, volume 2, pages 255–256, 3 2003. (in Japanese).

[16] K. Yanai. Generic image classification using visual knowledge on the web. In *Proceedings of ACM Multimedia 2003*, pages 67–76, 2003.