QUEVICO: A Framework for Video-based Interactive Media

Hidekatsu IZUNO, Yuichi NAKAMURA, Yuichi OHTA
IEMS, University of Tsukuba
1-1-1 Tennodai, Tsukuba 305-8573, JAPAN
{izuno, yuichi}@image.esys.tsukuba.ac.jp

Abstract

QUEVICO is a question-based video composition scheme in which video indexing and editing is designed from the viewpoint of "question and answer", and in which multi-view videos can be effectively used. Based on the tagset in this framework, we can structurize a video in a suitable way for retrieving a video portion relevant to the question. By editing and arranging the obtained portions, a smart answer will be given to the user. This paper introduces the basic idea of QUEVICO, its tagset, answering process, and our prototype system.

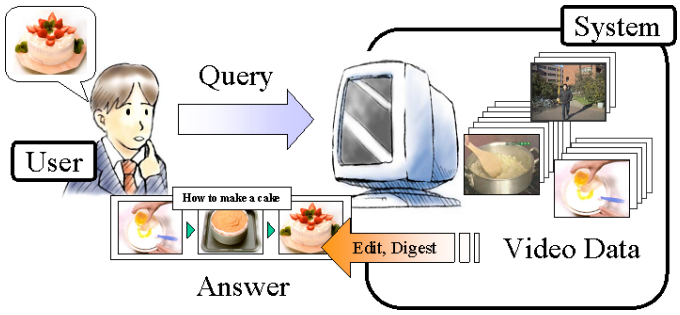

Figure 1: Video-based interactive media

1 Introduction

The aim of this research is to create a video-based interactive media that gives comprehensible answers to a question. While many works have been reported on intelligent help systems or question-answering systems that can communicate in natural languages, we often need explanations more than a text or a speech. If a person asks us to teach how to cook a sashimi, we strongly need visual explanation, *e.g.*, a picture of a raw fish, a demonstration for cutting a fish, and so on. In this sense, a video clip of an actual cooking is worth a thousand words. However, finding relevant video portions and editing those portions into a comprehensible explanation is a difficult task that requires intelligent video content management, and it has not been fully realized.

For this purpose, we propose a novel framework *QUEVICO*[1] that is designed for realizing intelligent video-based teaching materials. This framework has two important features:

- Video indexing and editing is designed from the viewpoint of "question and answer" in work.
- Multi-view[2] videos without editing are effectively used for answering questions through online editing.

In the following sections, we will present the basic idea of QUEVICO and interactive video-based media, the composition of data, and the mechanism for answering questions.

2 Framework for Video-based Interactive Media

2.1 Answering with Videos

Figure 1 shows the basic idea of video-based interactive media. The system stores video data for explaining important works, and the user may ask various questions, for example, "Tell me how to make a sashimi", "How long should I bake it?", and so on. The system answers questions by choosing relevant video portions, by choosing appropriate views, and by arranging or editing them.

One important advantage of using videos is the richness of information. Videos can give different kinds of information simultaneously. For example, "How much should I cut it?" may mean "How long ...?", "With which kitchen knife ...?", "When ...?", and so on. For answering this question in natural language, it is necessary to estimate the category of requested information and to compose sentences given as an answer. This may require precise understanding of the user's intention, or thorough search, in the stored knowledge, for all possible answers.

On the other hand, a video that captured the cutting action can give all together the information on "how long", "how much", "with which tools", and so on. What we have to do is *to know which portion of a video is the relevant answer, or which portion of a video potentially has the information the user can draw an answer*.

A video, however, does not hold complete information of the scene. A cameraman or a director carefully chooses a camera position, a view field, and carefully edits the obtained video. A director often edits out portions that he/she do not want to show. This process determines what information is kept in the video and what can be easily grasped at a glance. Therefore, when we explain something with a video, we need *to use a video taken with an appropriate setting and camerawork*. In this sense, it is desirable that we have multi-view videos without editing.

With the above conditions, the videos can be good resources that reduce difficult for answering questions.

2.2 QUEVICO: A Framework for Video-based Interactive Media

We designed QUEVICO, that is, "QUEstion-based VIdeo COmposition". This is a novel framework for realizing video-based interactive media, which has the following two important features:

- Video indexing and editing is designed from the viewpoint

[1] In Japanese myths, QUEVICO (or KUEBIKO) is a god of knowledge, whose figure is a scarecrow and who is a guardian of agriculture.

[2] It is often called as "multi-angled". Multiple cameras shoot at the same scene with different setting, *e.g.*, different position, different view field, and so on. Videos taken by them are stored in a synchronized format.

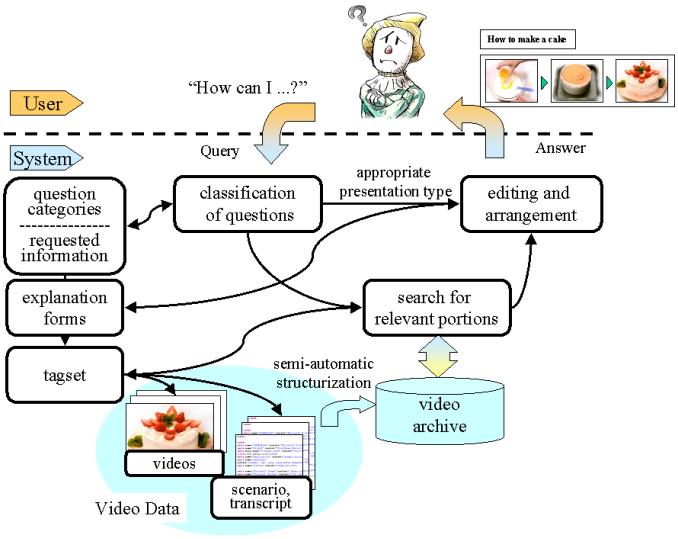

Figure 2: The outline of processing

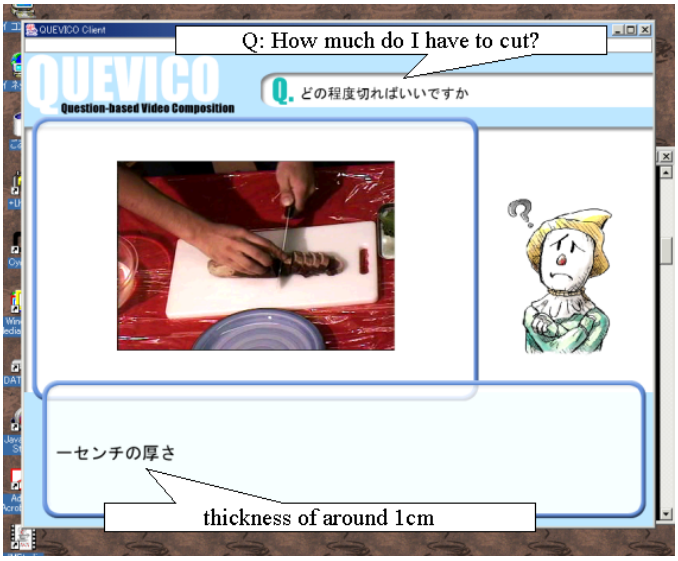

Figure 3: An example of answering a question

of "question" and "answer". A variety of questions were considered, and a XML tagset for marking-up each portion that potentially gives an answer to those questions was determined. The answer is chosen by considering "what information is requested by a question" and "which is the best method to show relevant data for the requested information".

- Multi-view videos without editing are effectively used. When we deal with edited videos such as TV programs, they are insufficient since essential information is often edited out. By dealing with multi-view videos without editing, we simplify the problem of selecting and editing video portions.

Figure 2 shows the outline of our interactive media based on QUEVICO. The video data are stored and marked-up by the tagset of QUEVICO. They are taken by multiple cameras and stored without editing. Although the tagging is currently a manual process that requires considerable cost and time, we have been developing semi-automatic method by integrating image processing and natural language processing[9].

Through the interaction between a user and the system, the system estimates which information should be given to the user. In this portion, we are currently using a simple process that matches between an actual question and "question type" with other required values for answering [3]. Suppose that a user asks "How much should I cut it? " concerning cutting bonito. Our system gives an answer with a video fragment that explains the "degree" of cutting as shown in Fig. 3. Such a video fragment can explicitly or implicitly give the similar information as a natural language explanation "slice it up with the thickness of around 1cm".

2.3 Related Work

Many works have been reported on video indexing and retrieval, *e.g.*, Informedia project[3], and they introduced various methods for analyzing and structurizing videos. One of the most common ways for video retrieval is to search fro significant words from transcripts, and another is to find relevant video segments in terms of color features. Such kinds of video retrieval, however, are methods of "retrieving related data portions", and is not of "answering questions". In this sense, our approach, that is video management based on question, is unique. Moreover, our framework uses multi-view videos in order to compose comprehensible answers.

As for tagging, although MPEG-7 standard incorporates XML, we currently use our original tagset, since the MPEG-7 standard for semantic description is not completely ready. We will move to MPEG-7 after the semantic portion of its standard is fully fixed.

In the natural language processing and AI field, many researchers have reported their interactive systems, some of which are used for question-answering systems. Our research is different on the point of concentrating on video specific problems, such as video tagging, editing, and the selection of multimedia data. Hopefully, useful techniques of natural-language-based interaction schema can be incorporated into our conversational module.

3 Answering by Video Data

3.1 Question and Answer

In order to develop the data structure based on "question and answer", we intensively checked broadcasted cooking shows and made a list of possible questions for typical indoor works. Table 1 shows typical questions that we gathered, which should be dealt with our framework.

Table 2 shows an example of categorized questions and the information that requested by them. The first column shows the questions for which we categorized into more than 30 types, and the second column shows the information that each type of question requests. We consider that the system can *answer* the questions, if the system can retrieve video portions from which the user can draw the requested information.

3.2 Answering Scheme

Potential answers can be obtained by searching for data closely related to a question. Tagging to data is the common way for specifying this relationship and delineating the location of potential answers. For this purpose, we use tags for specifying raw data such as a bounding-box[4] that encloses an object's figure,

[3] We do not focus on the natural language processing, since we want to concentrate on the problem on handing videos.

[4] We often use a bounding box on image that encloses the object's figure.

Table 1: Typical questions

How can I make a sashimi?
How should I cut it?
What kind of food do I need to prepare?
Why should I add water?
Is there any suggestions?
Which kind of fish is suitable for this dish?
How much sugar do I need to put?
How is the finish form?
How would a professional cook do?
How long does it take?
To which shape do I need to cut?
Salt is running out. What should I do?

Table 2: Typical questions and requested information

Question type	Requested information
Tell me how to (verb)	task, dependency, duration
What should I (verb)?	task, substitution, instrument, patient, dependency
Why do I need to (verb)?	reason, dependency, output
What happens when I (verb) it?	output, method
What should I use?	material, substitution, input/output, reason
How many/much do I need to (verb)?	degree, duration, input-quantity, method, task
Is there anything to pay attention?	note, method, degree, quantity
How will be the result?	input/output, task
Who is (verb)+ing?	agent, location, dependency
What is he/she (verb)+ing?	patient, instrument, state, reason, method
Where is he/she (verb)+ing?	location, task, agent, destination

and also use a tag for an object, tags for a task or a task structure. Every important portion in data is marked-up by those tags.

However, we still have missing links between a question and those data portions that are the candidates for an answer, and we need a formalism that delineates the paths from a question to its potential answers. Moreover, video-based multimedia are compound media, and we have certain degree of freedom in choosing answers: which portion of the data and by which modality of the data we answer. For example, when answer a question "How much do I need to cut?", we can show a moving image of actual cutting, an image of the result, or give just a phrase in a speech "around 1cm thickness". This is one of the most essential characteristics of multimedia.

To deal with this essential aspect of multimedia, we consider a model as shown in Fig. 4. The model has three-stage linking considering the following three types of relations: a relation between each question-type and each requested information type; a relation between each explanation form and each requested information; a relation between each explanation form to each data types. Each element in the model, *e.g.*, i-th requested information, has many to many links to other elements. By traversing the relations among these elements, we obtain virtual paths from questions to data portions.

Here we use Q for representing a set of question types, A for a set of requested information, F for a set of explanation forms, and D for a set of data types. Direct product $Q \otimes A$ repre-

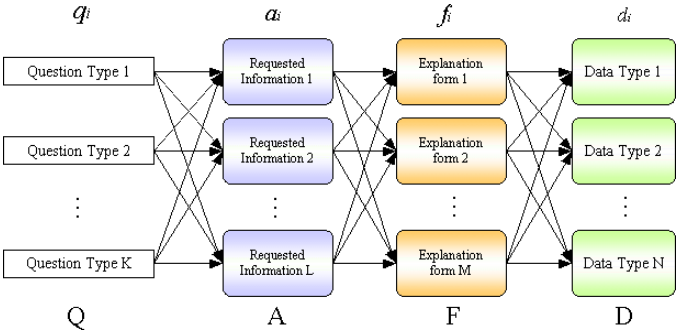

Figure 4: A multimedia QA model

Table 3: Example of explanation forms

name	the target's name that can be person's name, object name, task name, etc.
appearance	image of an object, image of a person, image for explaining location, etc.
movement	target movement, locus, etc.
adjacent object	an object that is always accompanying the target
input/output	input/output of an operation (task)
composition	part(s) that compose a target

sents "which information $a_i \in A$ is requested by each question $q_j \in Q$", which is partially shown in Table 2. We can consider that the value of each matrix element represents the relevance. Similarly, direct product $A \otimes F$ represents "which explanation form $f_i \in F$ is suitable for giving information $a_j \in A$", and direct product $F \otimes D$ holds the relation between an explanation form and a type of data portion. Examples of explanation forms are shown Table 3 and examples of data types are shown Table 4.

By using the above model, we can denote the answering scheme as the following.

$$\text{answering scheme} = Q \otimes A, A \otimes F, F \otimes D$$

After the most relevant data portions are chosen based on this scheme, the data are edited and given to the user.

4 Tagging in QUEVICO

4.1 Tagset

Based on the above idea, we devised the tagset for *marking-up the potential answers to a question*. Most of tags for defining data portions are simple. Physical portions of a video, *e.g.*, areas (regions) in a image, video segments, are marked-up, and they can have attributes for describing them. For example, a bounding box that encloses a person's face is marked-up as a

Table 4: Example of data portion types

image region	an image area that has the target's figure. a bounding box is often used.
video segment	video segment that is a sequence of images
audio segment	audio data in a video
word in a speech	a word in a speech, a word in a transcript
task in a scenario	a task description in a tagged form

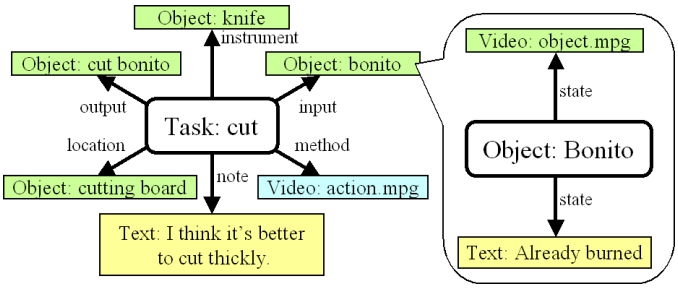

Figure 5: The relationships of task "cutting the bonito"

Table 5: Attributes of the tag for a task

attribute name	description
id	identifier
name	the name of a task
agent	the agent of the action in a task
patient	the objects of the action in a task
input	the input of a task
input-quantity	the quantify of the input
output	the output of a task
output-quantity	the output-quantity of a task
instrument	tools or materials required for performing a task
location	the location where the task is performed
source	the starting point (location) of the action in a task
destination	the end point (location) of the action in a task
time	the time when the task is performed
degree	the degree or the extent which a task is performed
reason	the reason for performing a task is necessary
substitution	alternative tasks that can substitute a task
note	something to pay attention for performing a task
duration	time length necessary for performing a task
dependency	dependence on other tasks

image region whose name is "face" and which has a pointer to the person's name.

For more abstract portions of a video, we have tags "a task" and "an object". They have important roles, since our short-term target is realizing interactive video manual. Those tags can be directly attached to the video data, or they can be attached to a scenario or meta-data if they exist.

Representation of a task and an object: A task is represented by its name and possible attributes as shown in Table 5. A set of tasks is structurally organized based on the orders of the tasks, and we denote the structure as "task tree". The tag is designed based on the questions and the requested information shown in Table 2. An object is represented by the tag as shown in Table 6. An simple example of these representations is shown in Fig. 5.

Note that any of the attributes except "id" and "name" can be omitted. If an attribute value corresponding to required information is directly given by a tag, it will be used as an answer. Otherwise, candidates for an answer are searched by using the scheme described in Section 3.

Table 6: Attributes of the tag for an object

attribute name	description
id	the identifier of an object
name	the name of an object
description	the description for an object
state	the current state of an object
color	the color of an object
shape	the shape of an object
quantity	the quantity of an object
smell	the smell of an object
reason	the reason for requiring an object
substitution	the substitution of an object

Tagging to video data: Figure 6 shows an example of directly adding a tagged description to a video. Here, a tag pair for a task (`<task>` and `</task>`) specifies tasks performed in a video. Two objects are denoted by `<object>`. Video segments are are described by `<video-segment>` whose "stime" expresses start time of the segment, "etime" expresses the end time. Those tags are referred by one another by their "id"s, such as "t1", "v1", and so on.

4.2 Tagging Process

As mentioned in Section 2.3, automatic video indexing is a hot research topic for the effective reuse of vast amount of video archives. Our group is also intensively investigating automatic tagging, such as object tracking, human movement recognition, speech recognition and so on[6][8][7]. One promising approach is automatic alignment between video and its scenario[9]. Some of the indexing technique will be used for our video archive in the near future.

In this paper, however, we skip those techniques and manually add tags to the data, since it is important to clearly separate video composition problems and automatic video indexing problems.

5 Answering Questions

At the current stage, answering processes is not fully fixed as mentioned in Section3. In the followings, we briefly describe the processes in our current prototype system.

5.1 Searching for Answer

The process for selecting relevant data portions is composed as follows:

1. The system receives a question form the user. By simple pattern matching, the system determines the type of the question. By using the the words in the question and current status of the system, the system also delineates for which task or for which object the user is requesting information.
2. According to the requested information, the system searches for the direct answer that is sometimes given as attributes of a tag.
3. If no direct answer is given, potential answers are searched for based on the scheme in Section 3. Retrieved data are scored by the relevance of linking. Examples of $Q \otimes F$ and $F \otimes A$ are shown in Table 7 and Table 8, respectively. If an element is given scores through two or more different paths, the summation of the scores is considered as the element's score. Eventually, the data with the highest score is chosen for the answer.
4. If the selected video is a multi-viewed video, the most appropriate view is selected according to the requested information.

```
<iimd>
<video-set>
  <video-segment id="v1" src="cake1.mpg" stime="10s" etime="62s"/>
  <video-segment id="v2" src="cake2.mpg" stime="67s" etime="90s"/>
</video-set>
<speech-set>
  <sentense>Please bake the cake <span id="p1">until it starts to brawn</span>.</sentense>
</speech-set>
<object-set>
  <object id="o1" name="cake"/>
  <object id="o2" name="fresh cream"/>
</object-set>
<task-set>
  <task id="t1" name="cook" output="#o1">
    <task id="t2" name="bake" patient="#o1" method="#v1" degree="#p1"/>
    <task id="t3" name="make up" patient="#o1" input="#o2" method="#v2"/>
  </task>
<task-set>
</iimd>
```

Figure 6: Tagging example

Table 7: Example of scoring relations between required information and explanation form

	name	appearance	movement	input	···
agent	1.0	0.7	0.3	0.0	···
patient	1.0	1.0	0.3	0.3	···
location	0.7	1.0	0.0	0.0	···
⋮	⋮	⋮	⋮	⋮	⋮
shape	0.3	1.0	0.3	0.0	···
⋮	⋮	⋮	⋮	⋮	⋮

Table 8: Example of scoring relations between explanation form and physical data

	image region	video segment	audio segment	word in a speech	···
name	0.3	0.3	0.7	1.0	···
appearance	1.0	1.0	0.3	0.3	···
movement	0.3	1.0	0.0	0.0	···
⋮	⋮	⋮	⋮	⋮	
input	0.3	1.0	0.3	0.0	···
⋮	⋮	⋮	⋮	⋮	

This process effectively uses the rich information of videos. Even if enough tags are not added or an exact answer is not contained in the video data, we can obtain an answer not far off the truth. Suppose that a user asks the question about an object, *e.g.*, object's color or shape. Although one of the best answers is the textual description such as "blue" or "square", a video clip that captured the object with close-up view can also be a good answer. In this case, we only need to know which view is the object's close-up. In another example, if a user asks "How long do I need to bake ... ?", a video fragment implicitly gives an answer by its length, even if no exact answer is given in the video.

Thus, our schema greatly improves the effectiveness of question-answering mechanism, since we cannot usually add a tag to every detail of video data. Currently, the scores shown in Table 7 and Table 8 are manually and empirically determined. For a future work, we are planning to apply a semi-automatic method with neural network.

5.2 Presentation of Answers

Since a video is a continuous medium and redundant, the users may have difficulty in understanding the presented videos. A simple user interface that only supports video playback is not enough to present the answer, since it may be still time-consuming to find necessary information. We need to use flexible forms for answering various questions.

One possible form is a diagrammatic representation. We proposed the *Video Icon Diagram (VID)* for representing the inner structures of a video[9]. The VID is a graphical representation composed of *video icons* each of which illustrates a video segment such as a shot or scene. The icons are arranged in a diagram according to semantic relationships, such as order, hierarchy, equivalence and so on. By simply viewing the diagram, a user can easily grasp the structure of a video. Other possible solutions are to add captions that emphasize the essence of a video, to make a visible CG narrator who summarizes the contents, and so on.

Currently, we are trying to incorporate the VID. Some examples are shown in the next section (in Fig. 8). Video icons comprehensively represent the contents of a video. Other solution will be reported in the near future.

5.3 Some Examples

Here we shows some examples obtained by our prototype system. The video contents are about cooking, one of which is "How to cook lightly roasted bonito". The videos are taken by four views as shown in Fig. 7: scene view (wide-angled establishing shot), speaker's view (middle shot of a speaker), table view (close-up shot at objects), and workspace view (close-up shot of the hands and manipulation). The speech text, that is a transcript, and the scenario along which the video is taken are attached to the video data. Tags are manually added to this combination of data.

An example of questions and the system outputs are presented in Fig. 8. As we can see here, the answers by the system are satisfactory for a simple question. The system is still under development, and more intelligent functions will be added in the near future.

6 Conclusion

In this paper, we proposed a novel framework QUEVICO for video-based interactive image media that realizes question-

Figure 7: Multi-view videos (how to cook lightly roasted bonito)

answering as a teacher does. We are currently developing a prototype system based on QUEVICO. Although the implemented functions on this system are still simple, the system showed good potential for answering relatively simple questions.

For future works, we still need intensive work to develop the prototype system, and we will need systematic evaluation in order to prove the effectiveness. We also need to add some important mechanism, for example, a function to recognize the user's status or situation.

References

[1] J. Marti'nez, "Overview of the MPEG-7 Standard" ISO/IEC JTC1/SC29/WG11 N4509 Pattaya, 2001

[2] M. Murata, M. Utiyama, and H. Ishihara, "Question Answering System Using Similarity-Guided Reasoning" (in Japanese), Natural Language Processing, pp.135-24, 2000

[3] H. Wactlar, T. Kanade, M. Smith, and S. Stevens, "Intelligent Access to Digital Video: The Informedia Project", IEEE Computer, Vol.29, No.5, 1996

[4] M. Smith and T. Kanade., "Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques" Proc. IEEE CVPR, 1997

[5] H. Jiang and A. Elmagarmid, "WVTDB - A Semantic Content-Based Video Database System on the World Wide Web", IEEE Trans. on KDE, vol.10, NO.6, 1998

[6] Y. Nakamura, "Multimodal Approach toward Intelligent Video Production", Proc. Int'l Workshop on Multimedia Intelligent Storage and Retrieval Management, 1999

[7] M. Ozeki, Y. Nakamura, and Y. Ohta, "Camerawork for Intelligent Video Production —Capturing Desktop Manipulations", Proc. Int'l Conf. on Multimedia and Expo, 2001

[8] M. Ozeki, M. Itoh, Y. Nakamura, and Y. Ohta, "Tracking Hands and Objects for an Intelligent Video Production System", Proc. Int'l Conf. on Pattern Recognition, 2002 (to appear)

[9] M. Murayama, H. Izuno, Y. Nakamura, and Y. Ohta, "Video Icon Diagram: Representation of Video Contents Structure"(in Japanese), IEICE, SIG-PRMU-2001-45, 2001

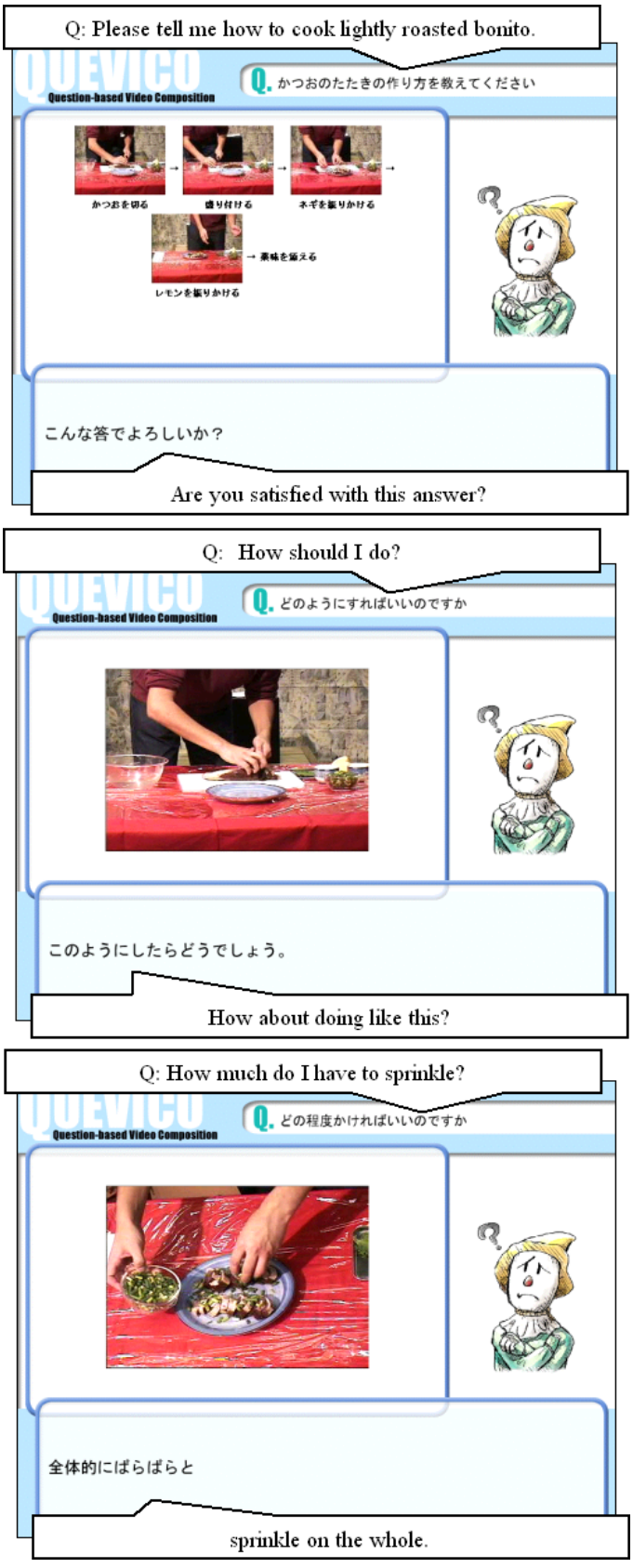

Figure 8: Output of our prototype system