2C1-04

A Framework for Interactive Video Based Explanatory System

Burin Anuchitkittikul Toyoaki Nishida Sadao Kurohashi

Graduate School of Information Science and Technology

The University of Tokyo

A large amount of worthy video data becomes meaningless for the user if he doesn't satisfy the information provided to him especially for the explanatory system. Thus, to satisfy the user, we have to serve him the information that can fulfill not only his explicit requests but also his implicit wants. This paper introduces a framework for an active system that creates video-based contents according to the knowledge about the user that the system can detect via the action that the user performs towards each video. We describe the basis framework of the system by evaluating our framework through the domain of cooking.

1. Introduction

"One picture is much worthier than thousands of words". This classical saying has been heard and well known for a long time. But since the advent of motion pictures, one static picture became less important in the sense of providing information to the user especially for the explanatory system. Consequently, the amount of multimedia information has been instantly increased. Tons of this multimedia data will be worthless without the presence of the appropriate and efficient management. One necessary way to manage this massive amount of multimedia information to be explainable for the users is to provide only the information needed to be consumed by them.

In this paper, we present the framework of the system which interactively creates video-based contents that serve the user what he really needs by considering the knowledge about the user obtained by the user's action or behavior. The focus of our framework is to improve the flexibility of displaying video-based contents that serve the user the essential information in the domain of cooking.

The rest of this paper will be organized as follows: Section 2 describes about researches that are related to ours. In section 3, characteristics of the cooking domain are presented. The framework of our system will be introduced in section 4. Section 5 presents the discussion and the problems occurred in our framework. Finally future works will be mentioned at the end of this paper.

2. Related works

There are two fields of researches that are related to our work namely content-based retrieval system and auto video data summarization system. Both of them share one common goal with our work which is to provide the users only the necessary information.

In the content-based retrieval system, most researches focus on improving the efficiency of the system by trying to increase the precision and recall of the retrieved results but overlook the importance of how to display the retrieved results in order to fulfill the user's needs. Therefore, user may not be satisfied with the retrieved results even though the precision and recall of the retrieved result is high. One well known project that also concerns about how to present the retrieved results to the users is The Informedia Digital Video Library Project [1] at Carnegie Mellon University. This project concerns not only retrieving the video from the libraries but also exploits the video summarization techniques to make a synopsis of the retrieved videos which is called "video skimming"[2] which is to compact the a video by choosing the "significant images and words" from the video. This system characterizes the significance of video through the integration of image and language understanding. Segment breaks produced by image processing can be examined along with boundaries of topics identified by the language processing of the transcript. Though this system concerns how to display the retrieved videos to the user, it doesn't consider what the user is expecting to consume at all.

Researches in the automatic video content summarization field attempt to generate a concise and informative video summary in order to enable a user to quickly figure out the overview contents of a video. Most of the researches in this field extract multiple key frames from shots by using a frame content change computed by features, such as motion activity [3]. Zhuang proposed an unsupervised clustering scheme to adaptively extract key frames from shots [4]. These methods require a pre-defined threshold or key frame number to control the density of key frame in a shot. Besides, these methods don't concern whether the summarized video synopsis meets human expectations or not. Yu-Fei [5] proposed a user attention model for video summarization in order to meet what humans are expecting from the video. The user attention model integrates a set of attention models which are extracted from video sequence. Since human attention is always attracted by visual information element, audio information element and linguistic information element, a complete user attention model is a linear combination of visual

Contact: Burin Anuchitkittikul
Graduate School of Information Science and Technology, The University of Tokyo 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8654 Engineering Building 3, room number: 228
Telephone : 03-5841-6689
Fax : 03-5841-8757
Email : burin@kc.t.u-tokyo.ac.jp

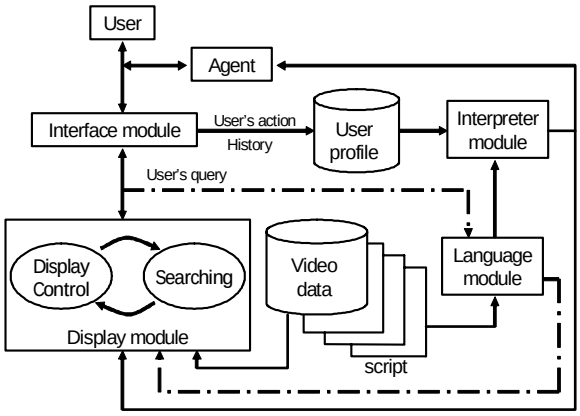

Fig1. Framework of the system.

attention model, audio attention model and linguistic attention model. But the model proposed by this system is modeled for generic users but the attention of each person is subjective and different so individual user's knowledge should be concerned also.

3. Characteristics of cooking domain.

In order to accomplish the mentioned objective, we evaluate our system to the domain of cooking. The data resource that we use in our system comprise the video data from cooking TV program called "Today's cooking" from NHK channel together with the utterance, or script, of the TV program. The reason why we selected cooking as our domain is because of some characteristics that exist in cooking domain. Some of those merits can be described as follows:

- Hierarchical structure

The processes of cooking one dish of food are organized as a hierarchical structure. For example, the processes of cooking cabbage-pork meatball may consist of some main operations such as cutting cabbage, preparing minced pork and so on. But each main operation also consists of sub-operations which show the detail of each main operation such as mixing some vinegar to the mince pork, add an egg into the minced pork and mix them up together. Thus, there are various ways to display the cooking video to the user. For example, showing only the main operations can be considered as the summarization of the video.

- Essential information and Individuality

Cooking domain depends intensely on knowledge and skill of the user. Therefore, individual skill and knowledge affect the required or essential information for each person. Furthermore, each person may pay his attention to different section even if they want to cook the same menu though. Thus, the points needed to be explained vary among each user.

- Suitable for video-based content

Since cooking consists of some processes or operations that are difficult to explain by using textual or static visual information so using a video to explain those operations is more practical. For example, showing the video of cutting cabbage can explain the user how to cut, when to cut and how big should it be.

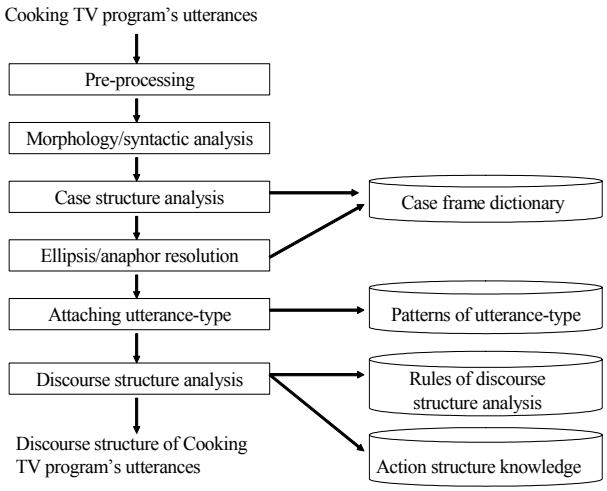

Fig2. The outline of the process for Language module.

- Variety of consequences

One outstanding merit of the cooking domain is that there are varieties of menus but most of them share some common ingredients or cooking actions. The characteristic of cooking domain is that there are alternatives to perform for the same ingredients in order to obtain different menus. In the same manner, the same set of cooking actions can be applied to different ingredients to create new menus also.

4. Framework

It can be seen from the previous sections that the knowledge of the individual user is necessary to flexibly create video-based content providing explainable information that serves what the user really needs especially in the cooking domain. This section describes the basis framework that we proposed to achieve the mentioned purpose. The framework of the overall system can be shown in Fig1.

4.1 Language Module

The main function of the language module is to construct a discourse structure of the utterance of the cooking TV program [6] and then supply this structure information to other modules. The input of the language module is the utterance of the script of the cooking TV program. Then the utterance will be pre-processed by cutting and a group of sentences will be delivered to another process as shown in Fig2. According to these processes, the hierarchical structure of the cooking TV program can be constructed. One example of the result of language module is shown in Fig3.

4.2 User profile

In our framework, the knowledge of individual user to be detected and stored as a user profile consists of 2 types of information which are

- Explicit Information
- Implicit Information

Explicit information is the information that explicitly shows the intention of the users. Some examples of this type of

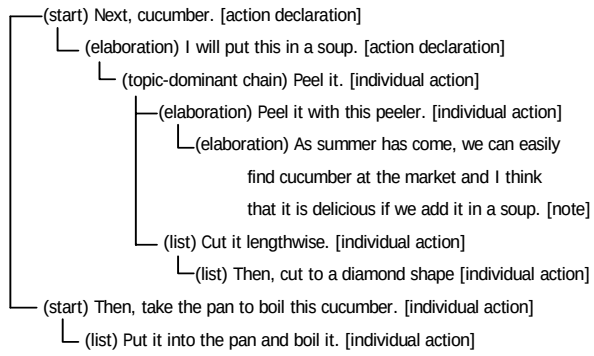

Fig3. Example of the result from Language Module.

information are the query that the user request or the items that the user selected or discarded.

On the other hand, implicit information is the information that doesn't show the intention of the user to the system explicitly. This type of information is as important or maybe more important than the explicit information because the user himself may not realize that this type of information is the one that he really needs.

Even though the actions from the user are explicitly seen but there are some implicit meanings beneath those actions. Therefore, we consider the actions that the user has towards the displaying video as implicit information. Thus, the implicit information of the user defined in our framework consists of user's action and content information of the scene in which the action of the user occurs.

- User's action : skip the video, pause the video, etc.
- Content information : material, process' action and utensil

In order to store this type of information we collect the occurrence's frequency of each user's action according to the content information. Example of the occurrence's frequency of the action "skip" is shown in table1.

Material \ Action	Mix	Fry	etc
Minced pork	10	2	…
Chicken	0	6	…
etc	…	…	…

Frequency (times)

Table1. The "skip" action in the user profile.

As the user uses the system repeatedly, both of this information will be gradually increased in the user profile. Consequently, this user profile can be used as the knowledge about the user that the system could obtain.

4.3 Interpreter module

After we have extracted the explicit and implicit information

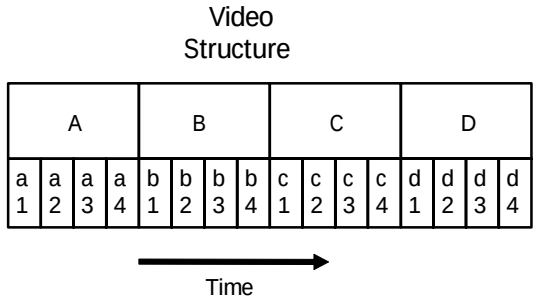

Fig4. Hierarchical Structure of the video

of the user and store them in a user profile, there must be an interpreter to interpret this information. Thus, the main function of the interpreter module is to evaluate the action of the system that should be performed to provide the only information that user needs. Namely, the interpreter module produces a feedback control considering from the user profile and user's current condition.

According to the framework in Fig1, the output of the interpreter module will be sent to the display module which has 2 main sections, Display control and searching. Therefore, the control output from the interpreter module will be different according to these functions. For the searching function, we simply use the scoring system based on the history information stored in the user profile. At the same time, the output control for the display control will be generated by a rule-based method.

As we have mentioned above the action that user does towards the video must contain some meaning, so we set some rules to evaluate those actions. The example of the rule-based method for evaluating the user actions is shown in Table2.

Skip action	Pause action	Action to be performed
Low	Low	Default
Low	High	Show more : Stretch/Detail
High	Low	Skim the Video
High	High	Ask?

Table2. Rules for "Mixing minced pork" scene.

The example of rules shown in Table 2 will be evaluated when the current condition of the user is the state that "Mixing minced pork" appears. Then, this control action will be sent to the display control section of the display module.

4.4 Display module

Display module consists of 2 cooperative sections called Display control section and searching section. These two sections work cooperatively to display the video according to the output control from the interpreter module. Searching section mainly functions as the content-based retrieval system that can retrieve the video according to the query of the user but the one to be

presented to the user will be selected by considering the user profile. In addition, the system can propose alternatives or suggest videos to the user when necessary even the system doesn't receive any requests from the user though. The display control section performs the edition of the video such as shortening or extending the length of the videos to be displayed to the user.

We have described above that we could construct a discourse structure of the utterance from the cooking TV program by using the language module. With this hierarchical structure, there are various ways to present the videos to the user. Suppose that the structure of the video is as shown in Fig4, the output control of the interpreter module according to Table 2 can be executed by the display control section as follows:

- Default :
  a1{10s}→b1{10s}→c1{10s}→d1{10s}
- Skim :
  a1{10}→b1{5s}→c1{10s}→d1{10s}
- Detail :
  a1{10s}→a2{10s}→a3{10s}→c1{10s}
- Stretch :
  a1{10}→b1{15s}→c1{10s}→d1{10s}
- Ask :
  Provide the agent to ask what the user really wants.

5. Discussion

From the framework mentioned above, we can detect the user's knowledge and utilize this knowledge to serve his needs up to some extent but there are still several points needed to be improved. For example, the rules that we used to evaluate the explicit and implicit information in the interpreter module are too simple and limited. The information to be stored in a user's profile is still insufficient to be definitely considered as user's knowledge.

Another issue that we need to focus on is the uncertainty of the implicit information. Since the implicit information is not explicitly expressed from the user, the only way to obtain this information is to deduce from the explicit means performed by the user. Therefore, it can easily be misinterpreted and mixed up with other faulty information. Thus, the discrimination between meaningful explicit information and futile noise becomes a tedious task that we have to concern also.

Furthermore, because so far we have paid attention on display control section and searching section separately and focused more on display control section so there is a gap between these two sections. This gap causes the gap between the user and the system also. One undesirable gap that can obviously be seen is that the feedback to be returned to the user is long-term affected, that is, the user's action doesn't affect the display control section or search section till the user starts a new session. However, the target of the user that we have concerned so far is the type of the user that only watches the cooking video but not simultaneously cook the menu showing in the video. Therefore, it can be acceptable for some level.

6. Future works

As we have mentioned in the previous section that so far we have excluded the real-time typed of user from our target of the user so the future direction of our research is to concern this type of the user also. The main task that we have to pursue is to fill up the gap mentioned above by providing a short-termed affected response to the user. One possible solution to this problem is to exploit the non-verbal means together with the user profile to control the video.

For example, the user who is performing cooking actions might have limited ways to utilize the system. Since he is cooking, both of his hands are too greasy and too dirty to use a mouse or keyboard to control the video explaining how to cook the menu that he is watching. So the non-verbal information of the user such as gaze or gesture should be used together with the verbal information to improve the flexibility of controlling the video.

Therefore, we have to design interface module that allow both verbal and non verbal to be utilized. In addition, we have to clearly define the ontology between the verbal and non-verbal behaviors of the user together with the explicit and implicit information to be inferred from those behaviors. The evaluation of this information should be concerned as one of our future works also.

References

[1] Hauptmann, A., Thornton, S., Houghton, R., Qi, Y., Ng, D., Papernick, N., Jin, R., "Video Retrieval with the Informedia Digital Video Library System"Proc. of the Tenth Text Retrieval Conference (TREC_-2001), Gaithersburgh, Maryland, 2001.

[2] M.A. Smith, T. Kanade, "Video skimming and characterization through the combination of image and language understanding techniques", Proc. of Computer Vision and Pattern Recognition, 1997

[3] W. Wolf, "Key frame selection by motion analysis," Proc. of ICASSP'96, vol.2, pp.1228-1231, 1996

[4] Y. Zhuang, et al, "Adaptive key frame extraction using unsupervised clustering," Proc. of ICIP'98, 1998.

[5] Yu-Fei Ma, Lie-Lu, Hong-Jiang Zhang and Mingjing Li, "A User Attention Model for Video Summarization" Proc of the 2002 ACM Workshops on Multimedia, December 2002

[6] Tomohide Shibata, Daisuke Kawahara, Masashi Okamoto, Sadao Kurohashi, Toyoaki Nishida: Structural Analysis of Instruction Utterances. KES2003: Seventh International Conference on Knowledge Based Intelligent Information & Engineering Systems, to appear, September, 2003