A Layered Structure of Human Interaction Interpretations

Masashi Takahashi *1*2 Sadanori Ito *2 Yasuyuki Sumi *1*2 Megumu Tsuchikawa *2 Kiyoshi Kogure *3
Kenji Mase *2*4 Toyoaki Nishida *1*2
***1 Graduate School of Informatics, Kyoto University *2 ATR Media Information Science Laboratories**
***3 ATR Intelligent Robotics and Communication Laboratories *4 Information Technology Center, Nagoya University**
takahashi@lab1.kuis.kyoto-u.ac.jp

ABSTRACT

We have started to develop an innovative system to capture and interpret human interactions automatically by using ubiquitous/wearable sensors in order to realize a new interface that exploits human contexts. This paper proposes a systematic framework for the interpretations of human interactions with a bottom-up approach that bridges the gaps among the context levels of data required by various applications.

INTRODUCTION

We are developing the technology for an interaction corpus, a huge collection of human interaction data captured by various sensors with their machine-readable indices, in order to realize a new interface that exploits human contexts in our daily life. The purpose of this study is to develop a systematic framework in which various applications can deal with human contexts represented as machine-readable indices in a uniform manner. Interaction indices also enable us to improve the availability of video/audio sources and to create a new medium for capturing our daily experiences and reusing captured sources efficiently. This paper proposes a layered model for the interpretations of human interactions with a bottom-up approach, systematically assigning layers from the acquired raw data of individual sensors to application semantics. Consequently, it becomes possible to bridge the gaps among the context levels of data required by various applications. This framework enables us to deal with human contexts flexibly in real time in building realistic applications according to the needs concerning real-time performances and abstraction levels of human contexts, without learning the detailed protocols of particular sensors. Moreover, we assume that it is possible to extend our system easily to various situations of our daily life in the near future because each layer has been composed independently of one another according to its semantic level.

We have prototyped the wearable/ubiquitous sensors shown in Figure 1 to capture human interactions from multiple points of view. In addition to cameras and microphones, we adopted an infrared ID system, which identifies persons or objects, in order to estimate the user's state of gazing at a particular person/object or that of staying at a particular place. We also adopted a throat microphone that is used to detect whether or not the user make an utterance.

To demonstrate our system, we set up the sensor room shown in Figure 2 at a poster exhibition site for the ATR Research Exposition 2003. In the room, the so-called Experience Capture Room, five booths were set up, and each booth had ten infrared ID tag units for posters and two sets of ubiquitous sensors to capture exhibitors, visitors, and posters at the front of the booth from the behind.

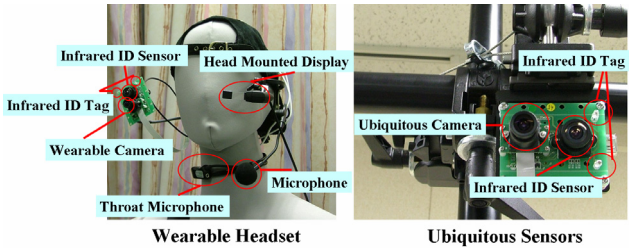

Figure 1. Sensors used to develop an interaction corpus.

Figure 2. Setup of experience capture room.

We are developing an interaction corpus using the room. In the following sections, we explain the layered model for the interpretations of human interactions and illustrate the use of our architecture through three example applications.

LAYERED MODEL OF INTERPRETATIONS

In the previous year, we also made an attempt to capture human interactions at an exhibition in an improvised fashion for the purpose of creating a video summary system [1]. Building on the earlier work, this latest attempt has a significant advance, that is, we systematically developed a framework in which various applications can utilize human contexts in a uniform manner efficiently. We introduce a layered model like Figure 3 to interpret human interactions by using a bottom-up approach. In this model, interpretations of human interactions are gradually abstracted so that each layer has unique semantic/syntactic information represented by machine-readable indices.

First, raw data acquired by individual sensors are stored in the first layer, the RawData Layer. In this layer, data are recorded in sensor-dependent formats. The problems arising from the differences in the characteristics of the various sensors are solved in the next layer.

In the second layer, the Segmentation Layer, the raw data is divided into meaningful clusters to provide information that is necessary for interpreting interactions to the upper layers.

In the third layer, the Primitive Layer, such basic elements of human interactions as 'LOOK_AT', 'TALK_TO', and 'CAPTURE' are extracted from the segments provided by the Segmentation Layer. For example, the situation where the infrared ID sensor worn by UserA captures UserB or ObjectX is interpreted as 'LOOK_AT'.

In the fourth Layer, the Composite Layer, these basic elements are spatiotemporally connected to each other to interpret complicated human interactions. For example, 'JOINT_ATTENTION' occurs when UserA and UserB look at ObjectX at the same time. In other words, a scene where a socially important event attracts many people's attention. And 'TOGETHER_WITH' occurs when UserA and UserB are captured by ObjectX at the same time, a scene where two or more people coexist in the same place.

This model provides applications with a common interface that suits their expected needs. Moreover, this model allows the use of heterogeneous sensors that sense redundant input regardless of the format of the sensor outputs. The abstraction level of the index rises as the layer goes up, but it takes more time to interpret human interactions in the higher layer. Therefore the real-time performance increases as the layer goes down.

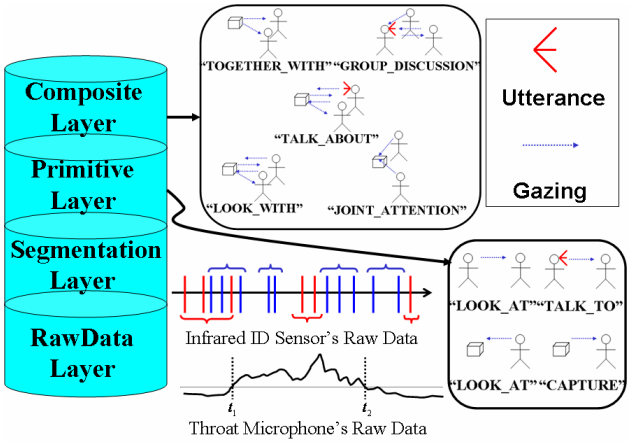

Figure 3. A layered model of interaction interpretations.

VARIOUS APPLICATIONS

In order to demonstrate the effectiveness of these indices, we have developed various applications such as those shown in Figure 4, each of which provides persons with the rich opportunities of sharing their experiences with others.

The first application is a wearable-based personal guidance system that provides the user with beneficial information in a head mounted display (HMD) in real time. This system provides him/her with information on the persons or objects in front of him/her while he/she is looking at them. The persons/objects in the user's sight can be detected in real time by referring to the Segmentation Layer. Moreover, the system provides the user with recommendation concerning the posters suitable to the fields of his/her interest and the persons whose interests lie the closest to his/hers. In this case, the degree of the user's interest or the similarity of action patterns between visitors can be calculated by using the indices of the Primitive Layer.

The second application is a communication robot that provides various experiences to the visitor in front of the robot. The robot refers to the history of the visitor's interactions in the Primitive Layer and communicates with him/her by using this information. They discuss the exhibitions that the user visited and share recommendations.

The third application is a video summary system. The system extracts highlight scenes from interaction indices and summarizes the user's experiences in a short video by using images from multiple points of view. The highlight scenes can be derived from the Composite Layer and the summary video is created in a linear time fashion by using a single source video at a time in each scene.

These applications could work properly with no serious trouble during the two-day demonstration. We found that the layered model in Figure 3 enabled us to develop a wide variety of applications such as these, from those that need information on a high-abstraction level to those that must work strictly in real time.

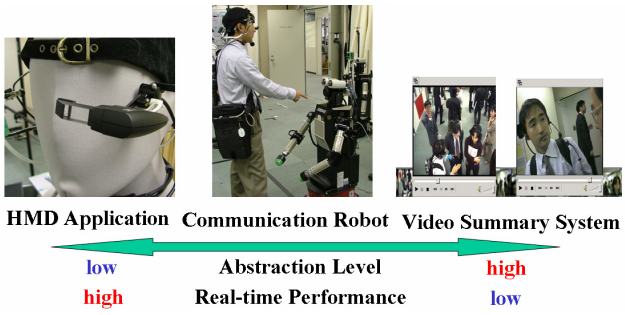

Figure 4. Various applications that refer to indices.

CONCLUSIONS

At the two-day demonstration, we were able to develop a valuable interaction corpus. Now, we are trying to capture human interactions and develop an interaction corpus in various domains, not only at the exhibition sites but also in meetings and lectures.

ACKNOWLEDGEMENT

The research presented here is supported in part by the National Institute of Information and Communications Technology.

REFERENCES

1. Yasuyuki Sumi, Sadanori Ito, Tetsuya Matsuguchi, Sidney Fels, and Kenji Mase. Collaborative Capturing and Interpretation of Interactions, Pervasive 2004 Workshop on Memory and Sharing of Experiences, pp.1-7, 2004.