COLLABORATIVE CAPTURING AND INTERPRETATION OF EXPERIENCES

Yasuyuki Sumi[*], Sadanori Ito[†], Tetsuya Matsuguchi[‡], Sidney Fels[§], Kenji Mase[¶]

Abstract

This paper proposes a notion of interaction corpus, *a captured collection of human behaviors and interactions among humans and artifacts. Digital multimedia and ubiquitous sensor technologies create a venue to capture and store interactions that are automatically annotated. A very large-scale accumulated corpus provides an important infrastructure for a future digital society for both humans and computers to understand verbal/non-verbal mechanisms of human interactions. The interaction corpus can also be used as a well-structured stored experience, which is shared with other people for communication and creation of further experiences. Our approach employs* wearable *and* ubiquitous *sensors, such as video cameras, microphones, and tracking tags, to capture all of the events from multiple viewpoints simultaneously. We demonstrate an application of generating a video-based experience summary that is reconfigured automatically from the interaction corpus.*

1. Introduction

Weiser proposed a vision where computers pervade our environment and hide themselves behind their tasks[2]. To achieve this vision, we need a new HCI (Human-Computer Interaction) paradigm based on embodied interactions beyond existing HCI frameworks based on desktop metaphor and GUIs (Graphical User Interfaces). A machine-readable dictionary of interaction protocols among humans, artifacts, and environments is necessary as an infrastructure for the new paradigm.

As a first step, this paper proposes to build an *interaction corpus*, a semi-structured set of a large amount of interaction data collected by various sensors. We aim to use this corpus as a medium to share past experiences with others. Since the captured data is segmented into primitive behaviors and annotated semantically, it is easy to collect the action highlights, for example, to generate a reconstructed diary. The corpus can, of course, also serve as an infrastructure for researchers to analyze and model social protocols of human interactions.

Our approach for the interaction corpus is characterized by the integration of many sensors (video cameras and microphones), ubiquitously set up around rooms and outdoors, and wearable sensors

[*] Graduate School of Infomatics, Kyoto University, http://www.ii.ist.i.kyoto-u.ac.jp/~sumi

[†] ATR Media Information Science Laboratories

[‡] University of California, San Francisco

[§] The University of British Columbia

[¶] Information Technology Center, Nagoya University

(video camera, microphone, and physiological sensors) to monitor humans as the subjects of interactions[1]. More importantly, our system incorporates ID tags with an infrared LED (LED tags) and infrared signal tracking device (IR tracker) in order to record positional context along with audio/video data. The IR tracker gives the position and identity of any tag attached to an artifact or human in its field of view. By wearing an IR tracker, a user's gaze can also be determined. This approach assumes that gazing can be used as a good index for human interactions[1]. We also employ autonomous physical agents, like humanoid robots, as social actors to proactively collect human interaction patterns by intentionally approaching humans.

Use of the corpus allows us to relate the captured event to interaction semantics among users by collaboratively processing the data of users who jointly interact with each other in a particular setting. This can be performed without time-consuming audio and image processing as long as the corpus is well prepared with fine-grained annotations. Using the interpreted semantics, we also provide an automated video summarization of individual users' interactions to show the accessibility of our interaction corpus. The resulting video summary itself is also an interaction medium for experience-sharing communication.

2. Capturing Interactions by Multiple Sensors

We developed a prototype system for recording natural interactions among multiple presenters and visitors in an exhibition room. The prototype was installed and tested in one of the exhibition rooms during our research laboratories' open house.

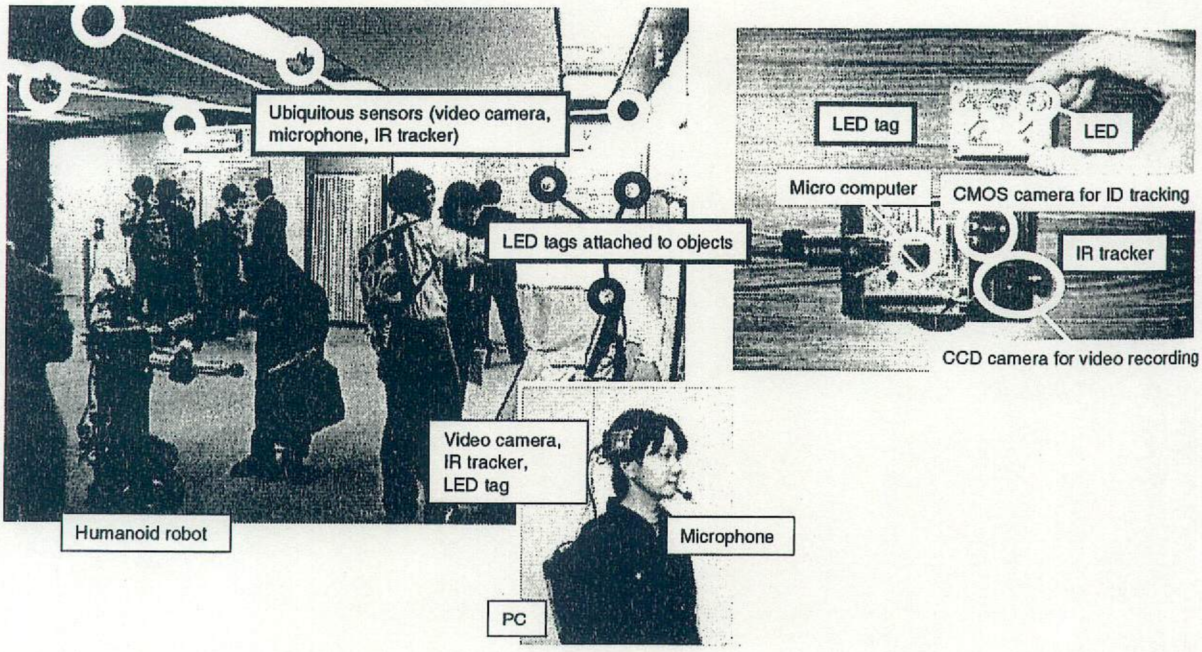

Ubiquitous sensors (video camera, microphone, IR tracker)
LED tags attached to objects
Humanoid robot
Video camera, IR tracker, LED tag
Microphone
PC
LED tag
LED
Micro computer
CMOS camera for ID tracking
IR tracker
CCD camera for video recording

Figure 1. Setup of the ubiquitous sensor room

Figure 1 shows a snapshot of the exhibition room set up for recording an interaction corpus. There

[1] Throughout this paper, we use the term "ubiquitous" to describe sensors set up around the room and "wearable" to specify sensors carried by the users.

were five booths in the exhibition room. Each booth had two sets of ubiquitous sensors that include video cameras with IR trackers and microphones. LED tags were attached to possible focal points for social interactions, such as on posters and displays. Each presenter at their booth carried a set of wearable sensors, including a video camera with an IR tracker, a microphone, an LED tag, and physiological sensors (heart rate, skin conductance, and temperature). A visitor could choose to carry the same wearable system as the presenters, just an LED tag, or nothing at all. One booth had a humanoid robot for its demonstration that was also used as an actor to interact with visitors and record interactions using the same wearable system as the human presenters.

The clients for recording the sensed data were Windows-based PCs. In order to incorporate data from multiple sensor sets, time is an important index. We installed NTP (Network Time Protocol) to all the client PCs to synchronize their internal clocks within 10ms. Recorded video data were gathered to a file server. Index data given to the video data were stored in an SQL server (MySQL) running on another machine. In addition, we had another server, called an application server, for generating a video-based summary by cut-and-paste editing of audio and video. At each client PC, video data was encoded into MJPEG (320 x 240 resolution, 15 frames per second) and audio data was recorded in PCM 22 KHz 16 bit monaural.

The prototyped IR tracker and LED tag are shown in Figure 1. The IR tracker consists of a CMOS camera for detecting blinking signals of LED and a micro computer for controlling the CMOS camera. The IR tracker was embedded in a small box with another CCD camera for recording video contents. Each LED tag emits a 6-bit unique ID, allowing for 64 different IDs, by rapidly flashing. The IR trackers recognize IDs of LED tags within their view in the range of 2.5 meters, and send the detected IDs to the SQL server. Each tracker data consists of spatial data, the two-dimensional coordinate of the tag detected by the IR tracker, and temporal data, the time of detection, in addition to the ID of the detected tag.

A few persons attached three types of biometric sensors – a pulse physiology sensor, skin conductance sensor, and temperature sensor – to their fingers. These data were also sent to the SQL server.

Eighty users participated during the two-day open house providing $\sim$ 300 hours of video data, and 380,000 tracker data, along with associated physiological data. The major advantage of the system is the relatively short time required in analyzing tracker data compared to processing audio and images of all the video data.

3. Interpreting Interactions

To illustrate how our interaction corpus may be used, we constructed a system to provide users with a personal summary video at the end of their touring of an exhibition room on the fly. We developed a method to segment interaction scenes from the IR tracker data. We defined interaction primitives, or "events", as significant intervals or moments of activities. For example, a video clip that has a particular object (such as a poster, user, etc.) in it constitutes an event. Since the location of all objects is known from the IR tracker and LED tags, it is easy to determine these events. We then interpret the meaning of events by considering the combination of objects appearing in the events.

Figure 2 illustrates basic events that we considered.

stay A fixed IR tracker at a booth captures an LED tag attached to a user: the user *stays* at the booth.

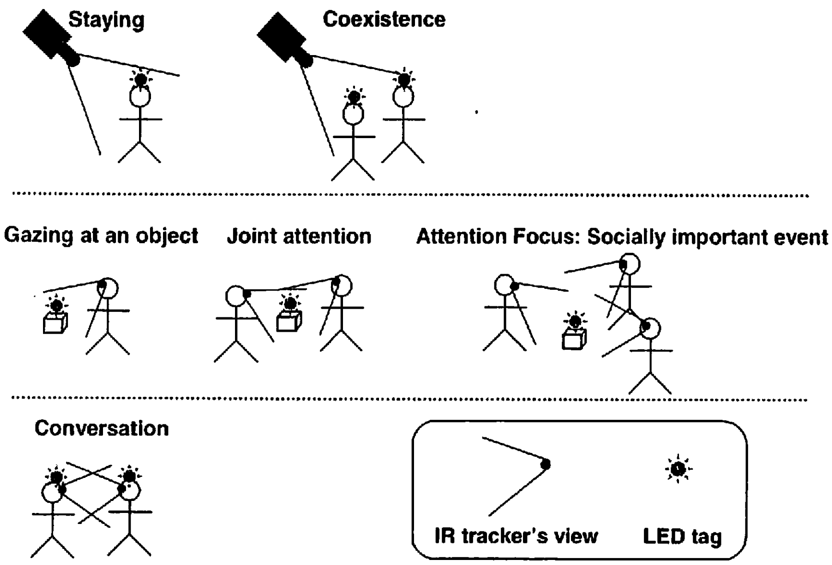


Figure 2. Interaction primitives

coexist A single IR tracker captures LED tags attached to different users at some moment: the users *coexist* in the same area.

gaze An IR tracker worn by a user captures an LED tag attached to someone/something: the user *gazes* at someone/something.

attention An LED tag attached to an object is simultaneously captured by IR trackers worn by two users: the users jointly pay *attention* to the object. When many users pay attention to the object, we infer that the object plays a socially important role at that moment.

facing Two users' IR trackers detect each others' LED tags: they are facing each other.

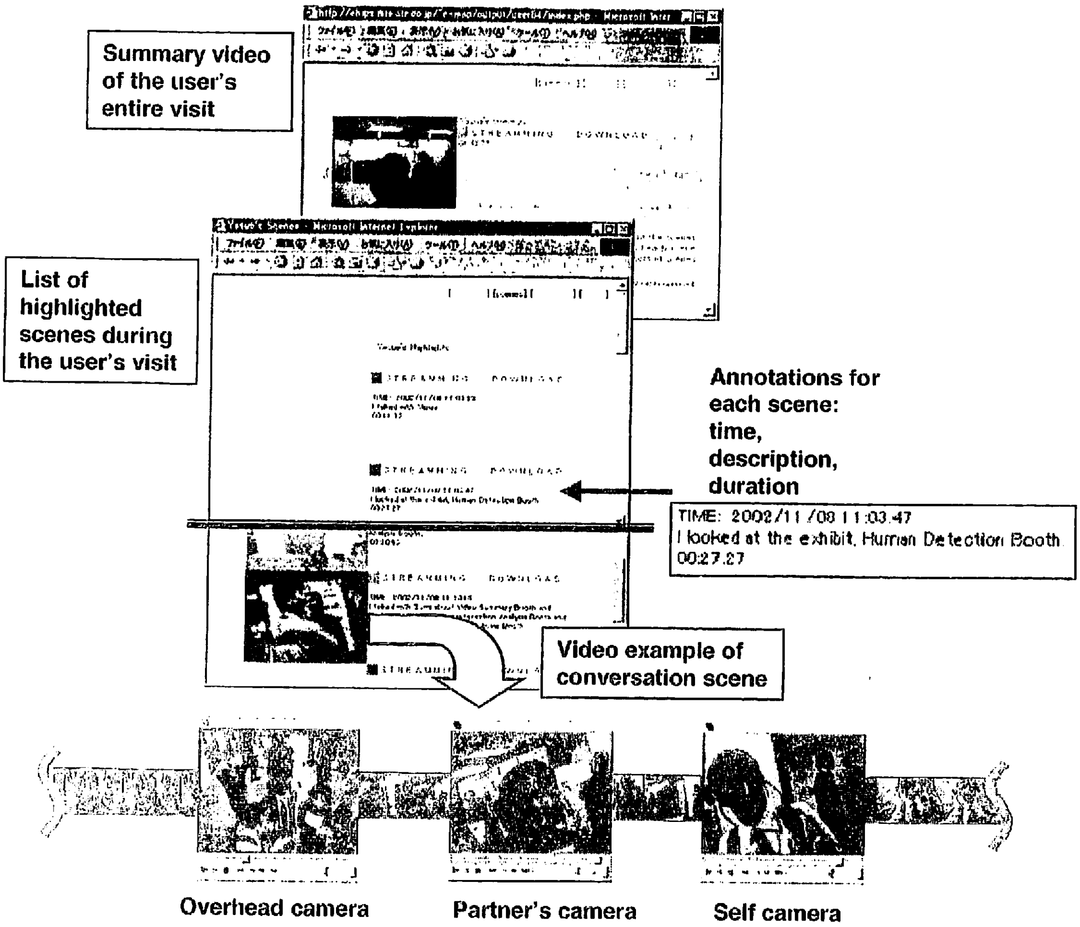
Raw data from IR trackers are just a set of intermittently detected IDs of LED tags. Therefore, we first group the discrete data into interval data implying that a certain LED tag stays in view for a period of time. Then, these interval data are interpreted as one of the above events according to the combination of entities attached by the IR tracker and LED tag.

In order to group the discrete data into interval data, we assigned two parameters, *minInterval* and *maxInterval*. A captured event is at least *minInterval* in length, and times between tracker data that make up the event are less than *maxInterval*. The *minInterval* allows elimination of events too short to be significant. The *maxInterval* value compensates for the low detection rate of the tracker; however, if the *maxInterval* is too large, more erroneous data will be utilized to make captured events. The larger the *minInterval* and the smaller the *maxInterval* are, the fewer the significant events that will be recognized.

For the first prototype, we set both the *minInterval* and *maxInterval* at 5 sec. However, a 5 sec *maxInterval* was too short to extract events having a meaningful length of time. As a result of the video analyses, we found an appropriate value of *maxInterval*: 10 sec for ubiquitous sensors and 20 sec for wearable sensors. The difference of *maxInterval* values is reasonable because ubiquitous sensors are fixed and wearable sensors are moving.

4. Video Summary

We were able to extract appropriate "scenes" from the viewpoints of individual users by clustering events having spatial and temporal relationships. A scene is made up of several basic interaction events and is defined based on time. Because of the setup of the exhibition room, in which five separate booths had a high concentration of sensors, scenes were location-dependent to some extent as well. Precisely, all the events that overlap at least *minInterval* / 2 were considered to be a part of the same scene.


Summary video of the user's entire visit
List of highlighted scenes during the user's visit
Annotations for each scene: time, description, duration
Video example of conversation scene
Overhead camera
Partner's camera
Self camera

TIME: 2002/11/08 11:03:47
I looked at the exhibit. Human Detection Booth
00:27.27

Figure 3. Automated video summarization

Figure 3 shows an example of video summarization for a user. The summary page was created by chronologically listing scene videos, which were automatically extracted based on events (see above). We used thumbnails of the scene videos and coordinated their shading based on the videos' duration for quick visual cues. The system provided each scene with annotations, i.e., time, description, and duration. The descriptions were automatically determined according to the interpretation of extracted interactions by using templates, as follows.

TALKED WITH I talked with [someone].

WAS WITH I was with [someone].

LOOKED AT I looked at [something].

In the time intervals where more than one interaction event has occurred, the following priority was used: TALKED WITH > WAS WITH > LOOKED AT.

We also provided a summary video for a quick overview of the events the users experienced. To generate the summary video, we used a simple format in which at most 15 seconds of each relevant scene was put together chronologically with fading effects between the scenes.

The event clips used to make up a scene were not restricted to those captured by a single resource (video camera and microphone). For example, for a summary of a conversation TALKED WITH scene, the video clips used were recorded by the camera worn by the user him/herself, the camera of the conversation partner, and a fixed camera on the ceiling that captured both users. Our system selects which video clips to use by consulting the volume levels of the users' individual voices. The worn LED tag is assumed to indicate that the user's face is in the video clip if the associated IR tracker detects it. Thus, the interchanging integration of video and audio from different worn sensors could generate a scene of a speaking face by camera with a clearer voice by his/her microphone.

5. Conclusions

This paper proposed a method to build an interaction corpus using multiple sensors either worn or placed ubiquitously in the environment. We built a method to segment and interpret interactions from huge collected data in a bottom-up manner by using IR tracking data. At the two-day demonstration of our system, we were able to provide users with a video summary at the end of their experience on the fly. Currently, we are developing a system that researchers (HCI designers, social scientists, etc.) can query for specific interactions quickly with simple commands and provides enough flexibility to suit various needs. We also work together with such researchers to improve our interaction pattern recognition.

Acknowledgements

Valuable contributions to the systems described in this paper were made by Tetsushi Yamamoto, Shoichiro Iwasawa, and Atsushi Nakahara. We also would like to thank Norihiro Hagita, Yasuyhiro Katagiri, and Kiyoshi Kogure for their continuing support of our research. This research was supported in part by the Telecommunications Advancement Organization of Japan.

References

[1] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing. In *ACM Multimedia '99*, pages 3–10. ACM, 1999.

[2] Mark Weiser. The computer for the 21st century. *Scientific American*, 265(30):94–104, 1991.