

Collaborative Capturing of Interactions by Multiple Sensors

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

[†]Kyoto University, [‡]ATR Media Information Science Laboratories, [§]University of British Columbia, [¶]Nagoya University
sumi@acm.org, tet@mit.edu, sito@atr.co.jp, ssfels@ece.ubc.ca, mase@itc.nagoya-u.ac.jp

ABSTRACT

We propose a notion of an interaction corpus, a captured collection of human behaviors and interactions among humans and artifacts. The corpus provides an important infrastructure for a future digital society for both humans and computers to understand verbal/non-verbal mechanisms of human interactions. Our approach employs multiple 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.

KEYWORDS: interaction corpus, ubiquitous/wearable sensors, video summary.

INTRODUCTION

Weiser proposed a vision that computers pervade our environment and hide themselves behind their tasks [1]. To achieve this vision, we need a new HCI paradigm based on embodied interactions beyond existing HCI frameworks such as the desktop metaphor and GUIs. 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, we propose to build an interaction corpus, a semi-structured set of a large amount of interaction data collected by various sensors. This corpus may serve as a venue for researchers to analyze and model social protocols of human interactions.

Our approach is characterized by the integration of many sensors (video cameras, trackers and microphones) ubiquitously set up around the room and wearable sensors (video camera, trackers, microphone, and physiological sensors) to monitor humans as subjects of interactions. Our system incorporates ID tags with an infrared LED (LED tags) and infrared signal tracking device (IR tracker) in order to record position context along with audio/video data. The tracking device is a parallel distributed camera array where any camera can determine the position and identity of any tag in its field of view. By wearing a tracking camera, a user's gaze can be determined. This approach assumes that gazing can be used as a good index for human interactions [2]. We also employ autonomous physical agents like humanoid robots as a social actor to proactively collect human interaction patterns by intentionally approaching humans.

Use of the corpus allows us to infer the captured event to interaction semantics among users by collaboratively processing data of the users who jointly interacted 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.

CAPTURING INTERACTIONS BY MULTIPLE SENSORS

We prototyped a 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.

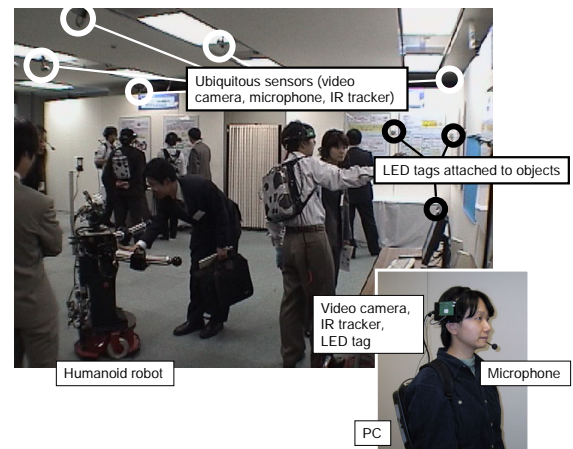


Figure 1: Setup of the ubiquitous sensor room.

Figure 1 is a snapshot of the exhibition room set up for recording an interaction corpus. There 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 or 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 the interactions using the same wearable system as the human presenters.

Eighty users participated during the two-day open house providing ~ 300 hours of video data, 380,000 tracker data along with associated biometric data.

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 at the 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.

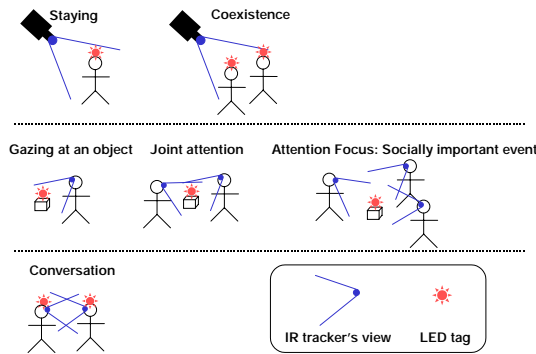


Figure 2: Interaction primitives.

Figure 2 illustrates basic events which we considered.

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

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

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

attention An LED tag attached to an object is simultaneously captured by IR trackers worn by two users: 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 tag: they are facing each other.

VIDEO SUMMARY

We were able to extract appropriate “scenes” from the viewpoints of individual users by clustering events having spatial and temporal relationships. 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. 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, e.g., *I talked with [someone]*; *I was with [someone]*; and *I looked at [something]*.

We also provided 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.

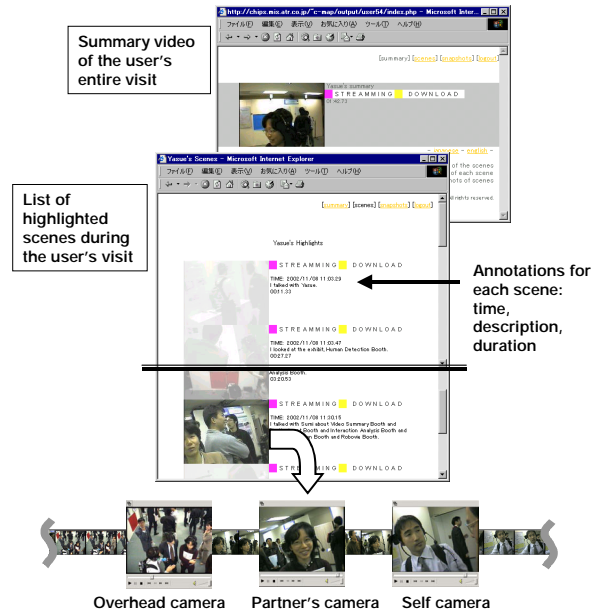


Figure 3: Automated video summarization.

The event clips used to make up a scene were not restricted to only ones captured by a single resource (video camera and microphone). For example, for a summary of a conversation “talked with” scene, video clips used were recorded by: the camera worn by a user him/herself, the camera of the conversation partner, and a fixed camera on the ceiling that captured both users. Our system selected which video clips to use by consulting the volume levels of users individual voices. Remember, 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.

CONCLUSIONS

This paper proposed a method to build an interaction corpus using multiple sensors either worn or placed ubiquitously in the environment. 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. In the future, we will develop a system that researchers (HCI designers, social scientists, etc.) can quickly query for specific interactions with simple commands and provides enough flexibility to suit various needs. We plan to work together with such research groups to improve our interaction pattern recognition and enrich the interaction corpus.

ACKNOWLEDGMENTS

Highly valuable contributions to this work were made by Tetsushi Yamamoto, Shoichiro Iwasawa and Atsushi Nakahara. This research was supported by the Telecommunications Advancement Organization of Japan.

REFERENCES

1. Weiser, M. The computer for the 21st century. *Scientific American*, 265(30):94–104, 1991.
2. Stiefelhagen, R., Yang, J., and Waibel, A. Modeling focus of attention for meeting indexing. In *ACM Multimedia '99*, pp. 3–10, 1999.