Deciphering Interactions from LED ID Tracking Data

Tetsuya Matsuguchi[†], Yasuyuki Sumi[†], Kenji Mase[†‡]

[†]ATR Media Information Science Laboratories, [‡]Nagoya University

1. Introduction

With the advent of ubiquitous and wearable computing technology, we are now capable of recording large amounts of data in various forms simultaneously [1]. However, there is a lack of system that is capable of analyzing an enormous amount of data quickly to output on the fly something meaningful and useful. Audio/Video data are especially problematic due to the time-consuming audio and image processing.

Here we report an implementation of a system that incorporates identity tags with an infrared light emitting diode (LED tags) and infrared signal tracking device (IR tracker) in order to record spatial and temporal context along with audio/video data [2]. We have developed a prototype for analyzing human interactions from large amount of data collected by ubiquitous and wearable computers. This system is designed to analyze intricate schemes of human interactions [1]. Its fast processing without any image processing makes it ideal for real-time applications as well. We discuss our approaches to recognizing human interactions from IR tracker data.

2. Setup and Demonstration

For the demonstration of our system, five booths were set up in the sensor room. Each booth had two sets of ubiquitous sensors that include video cameras with IR trackers and microphones. LED tags were attached to each of the posters and displays at the booths. One presenter at each booth carried a set of wearable sensors, which included a video camera with an IR tracker, a microphone, and an LED tag. A visitor could choose to carry the same wearable system as the presenters or just an LED tag, or nothing at all.

3. Data Analysis

During the two-day demonstration, with the participation of 80 users, we were able to collect ~300 hours of video data and over 380,000 tracker data. Our task is to develop an analysis tool to extract interactions from the tracker data and bring together important video data to create a meaningful summary of their interactions.

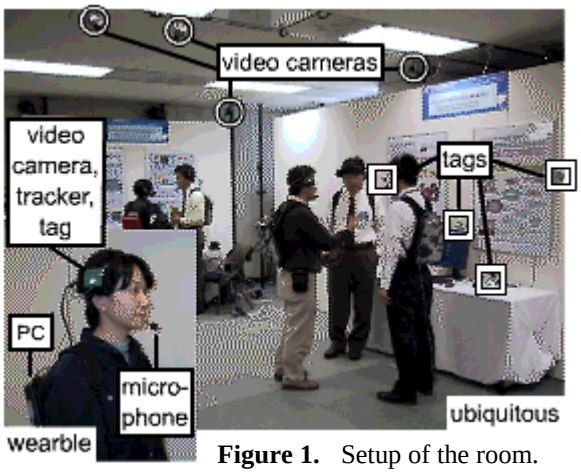

Figure 1. Setup of the room.

4. Tracker Data Analysis

Each tracking data consists of the time, the LED ID, and the coordinate of the object within the view of the tracker (x, y). Unfortunately, due to some hardware constraints, the detection rate was lower and the error rate of the tracker was higher than what we expected. Thus, any single tracking data by itself was not dependable. It was necessary then to distinguish the actual tracking data from the erroneously reported data. To this end, we employed two parameters, *minInterval* and *maxInterval*, to define a CAPTURED event. A CAPTURED event is at least *minInterval* in length, and times between tracking data that make up an event is less than *maxInterval*. The idea is that it is less likely to have erroneous data of the same value repeatedly. The *minInterval* also 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.

In order to adjust the parameters, we picked video clips in which a user is in the view of the video camera and annotated each frame with the values of the parameters that produce a CAPTURED event. This process allowed us to easily visualize the appropriate values for the parameters. Interestingly, we found that ubiquitous sensors (stationary) and wearable sensors (in motion) should have different

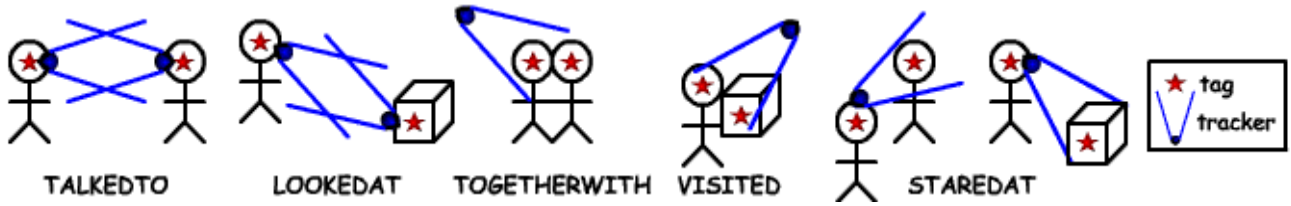

Figure 2. Basic Interaction Events

values for the parameters. As the result of the analysis, we decided to use 5 sec for *minInterval*, 10 sec for *maxInterval* of ubiquitous sensors, and 20 sec for *maxInterval* of wearable sensors. In the future when the detection and error rates are improved, the parameters can easily be changed using the same analysis.

5. Interaction Events

To simplify the analysis of human interactions, we defined interaction events to be the building blocks of an interaction. Five basic interaction events were used: TALKEDTO, TOGETHERWITH, LOOKEDAT, VISITED, and STAREDAT (Figure 2).

- TALKEDTO/LOOKEDAT events occur when UserA captures UserB/ObjectB at the same time as UserB/ObjectB captures UserA. Another words, two users (or a user and an object) are facing each other.
- TOGETHERWITH or VISITED events occur when two users, or a user and an object, were captured by the same IR tracker in the same time interval.
- STAREDAT events are "passive" interaction events, in which a user is capturing another user or an object. STAREDAT events are CAPTURED events that are at least twice the *minInterval*.

6. Scenes: Clustering Events

A scene is made up of several interactive events and is defined based on time. It also has some temporal dependence due to the clustering of sensors at each booth. Precisely, all the events that overlap at least *minInterval*/2 were clustered together to form a scene. In this prototype, we used simple rules to annotate the interactions in each scene. For example, if TALKEDWITH event with UserB and VISITED event at BoothC are clustered together, we inferred that the user was talking UserB at BoothC. Using our interaction analysis, we were able to find ~1800 scenes with an average length of 150 seconds per scene. Although the annotation is very primitive at this stage, this system has been useful and necessary in forming more complex definitions for the analysis.

7. Video Production

Scene videos were created in a linear time fashion using only one source of video at a time. In order to decide which video source to use to make up the scene video, we established a priority list. The priority list used was based on the following basic rules. When someone is speaking (the volume of the audio is greater than 0.1 / 1.0), a video source that shows the close-up view of the speaker is used. If no one that is involved in the event is speaking, use ubiquitous video camera source. In the time intervals where more than one interaction event have occurred, the following priority was used: TALKEDWITH > TOGETHERWITH > LOOKEDAT > VISITED > STAREDAT.

The audio for the scene videos were composed of all audio sources of users and objects that are part of each scene in order to reconstitute conversations and the atmosphere of the exhibition room.

8. Conclusions

At the two-day demonstration of our system, we were able to provide users with their summary at the end of their experience on the fly. In the future, we will develop a system that researchers can query for specific interactions quickly with simple commands and provides enough flexibility to suite various needs. We plan to work together with such researchers to improve our interaction pattern recognition. In addition to the on-the-fly service, we can also use audio and image processing to augment the data when more computing time is available and when detailed reports are necessary. We foresee that this system is useful not only for the study of human interactions, but automatic cataloging of personal video collection.

Acknowledgement

We would like to thank Sidney Fels for helpful discussions and assistance. We also thank the members of ATR MIS for the making the demonstration possible. This research was supported by the Telecommunications Advancement Organization of Japan.

References

[1] Sumi Y., Ito S., Matsuguchi T., Fels S., Utsumi A., Suzuki N., Nakahara A., Iwasawa S., Kogure K., Mase K., Hagita N. Collaborative capturing of interactions by multiple sensors, Interaction 2003, 2003.

[2] Ito S., Sumi Y., Mase K. Interaction Capturing Device with an Infrared ID Sensor, Interaction 2003, 2003.