Deciphering Interactions from Spatio-Temporal Data

Tetsuya Matsuguchi[1], Yasuyuki Sumi[1], Kenji Mase[1,2]

[1]ATR Media Information Science Laboratories
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan
[2]Nagoya University
{tet, sumi, mase}@atr.co.jp
Tel: +81-774-95-1401, Fax: +81-774-95-1408

We have built an interaction corpus in which audio, video, and ID tracking data were recorded and stored in databases by users and ubiquitous sensors. The interaction corpus can be used to analyze human interactions and to provide services such as finding certain events. As a demonstration of the power of the corpus, we provided the users access to the automatic and on-demand creation of a movie, which summarizes user's experiences. To this end, we have explored several approaches and parameters for using the ID tracking data to extract events, interactions, and scenes.

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][2][3]. However, there is a lack of system capable of analyzing these multimedia data quickly to output something meaningful and useful on the fly. 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 LED (LED tags) and infrared signal-tracking device (IR tracker) in order to record context along with audio/video data. This system, powered by the combination of ubiquitous and wearable computing, is designed to analyze the intricate schemes of human interactions such as gestures. We also discuss our approaches to recognizing human interactions from IR tracker data. 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.

2. Setup and Demonstration

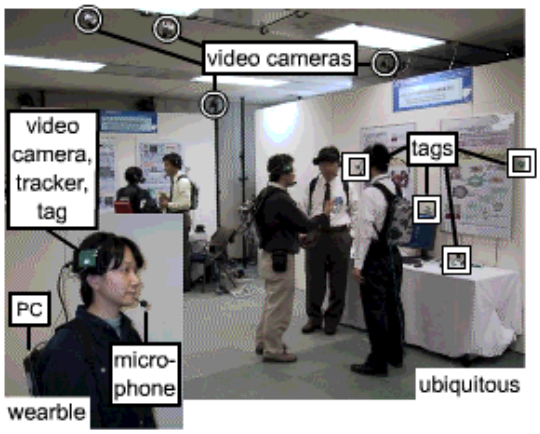

Figure 1. Setup of the Room

For the recording and demonstration of the system, five booths were set up in an exhibition 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.

Figure 2. Top two are taken by the two ubiquitous cameras. The bottom two are recorded by the wearable cameras.

The snapshots of Figure 2 show the four viewpoints of a particular interaction between a visitor and a presenter. The ubiquitous camera views allow us to observe what kind of interactions are going on, and the wearable camera views allow us to study the user's direction of sight, etc. The data recorded during the two-day demonstration include ~300 hours of video data and over 380,000 tracker data. The major advantage of the system is the relatively short time required in analyzing tracker data compared to processing audios and images of all the video data.

3. Tracker Data Analysis

Each tracker data consists of spatial data, the 2-dimentional coordinate of the tag detected by the IR tracker, and temporal data, the time of detection, in addition to the ID of the tag detected. Due to some hardware constraints on the LED tags and IR trackers, 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 then necessary to distinguish the actual tracking data from the erroneously reported data. To this end, we assigned two parameters, *minInterval* and *maxInterval*, to define a CAPTURED event. A CAPTURED event is at least *minInterval* in length, and times between tracker 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.

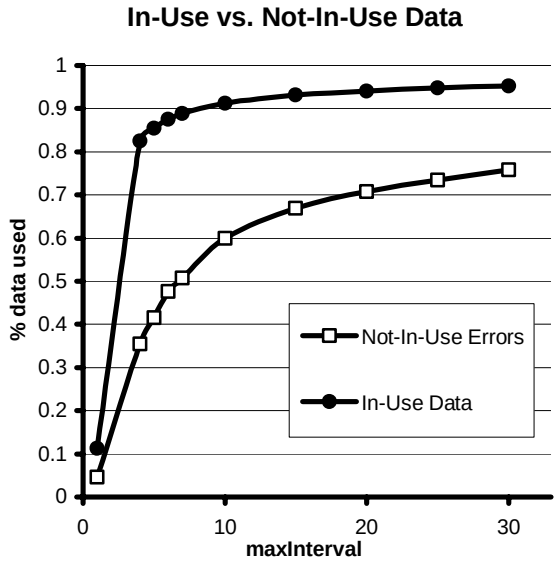

Figure 3A. Percentage of tracker data used with different values of *maxInterval*. Not-In-Use Errors are tracker data of IDs not in use, therefore could not have been real. In-Use Data are those that could or could not be real data.

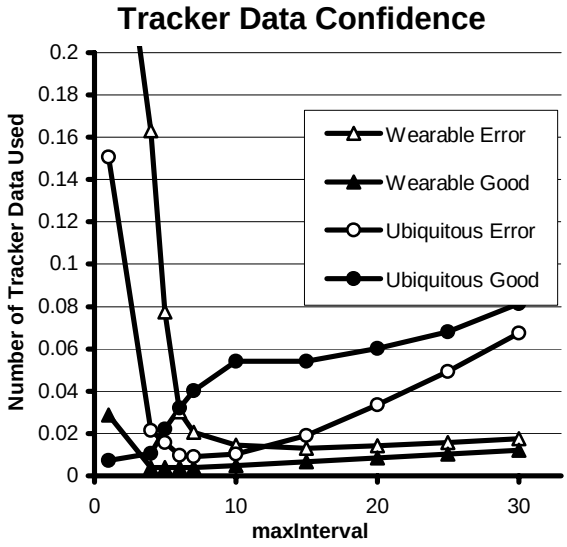

Figure 3B. Average number of tracker data used per event per second with respect to different values of *maxInterval.*

Each LED tag has a 6-bit ID, allowing 64 different ID's. About 24 tags were used for the ubiquitous sensors, leaving 40 tags to be used by the users. When an IR tracker reports an ID erroneously, it can either be an ID in use or not in use. It is simple to ignore IDs appearing when they were not in use, however, if an ID-in-use is reported erroneously, we must distinguish whether or not it is real. Fortunately, we were able to use the ID-not-in-use errors to provide error patterns, which allowed us to distinguish some of the bad data from good data. Figure 3A shows the percentage of tracker data used to create CAPTURED events depending on the value of *maxInterval* with *minInterval* = 5 sec. Although we would like to use as much of in-use data as possible, increasing *maxInterval* increases the use of erroneous data much faster than the use of good data. This provided with us a limit of less than 30 sec for the value of *maxInterval*.

Each CAPTURED event consists of at least two tracker data by definition. The more tracker data are used per CAPTURED event, the more confidence we can have in that event. Figure 3B shows the average number of tracker data used per CAPTURED event per second with varying *maxInterval*. Here we separated the analysis of data captured by ubiquitous sensors and wearable sensors. The reason for this was the apparent difference in the tracking capacity of the IR trackers. We suspect that the difference comes from the fact that the ubiquitous sensors are fixed while the wearable sensors are constantly in motion along with the head of the users. IR trackers seem to produce more errors when the targeting sensors are in motion. Another potential difference comes from different lens angles used for ubiquitous and wearable sensors. Whatever the physical differences there are between ubiquitous and wearable sensors, it is obvious from the Figure 3B that they have different capacities. In ubiquitous sensors, the number of tracker data used per event per second is much different between erroneous data and good (ID-in-

use) data. At *maxInterval* of 10 sec, there is the largest difference in the number of tracker data used. This property can be used to further distinguish erroneous data from real data. In wearable sensors, however, the erroneous data seem to have better confidence when *maxInterval* is below 30 sec. Therefore, this property is not useful for distinguishing erroneous data acquired by wearable sensors.

Figure 4. A snapshot of a video clip with some parameters inserted.

Finally, we used a visual analysis to determine appropriate *maxInterval* values for ubiquitous and wearable sensors. We did this by directly inserting different *maxInterval* values and tracker data into the video clips (Figure 4). This allowed us to easily visualize the effect of *maxInterval* in determining whether certain video clip should be part of CAPTURED event. As the result of the video analysis in addition to the analyses described above, 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.

4. Interaction Events

We defined 5 basic interaction events: TALKEDTO, TOGETHERWITH, LOOKEDAT, VISITED, and STAREDAT (Figure 5). Each basic interaction event is defined with a user or an object as a target. Distinction between interactions with a user and interactions with an object was made in order to provide hierarchy among the events in the future.

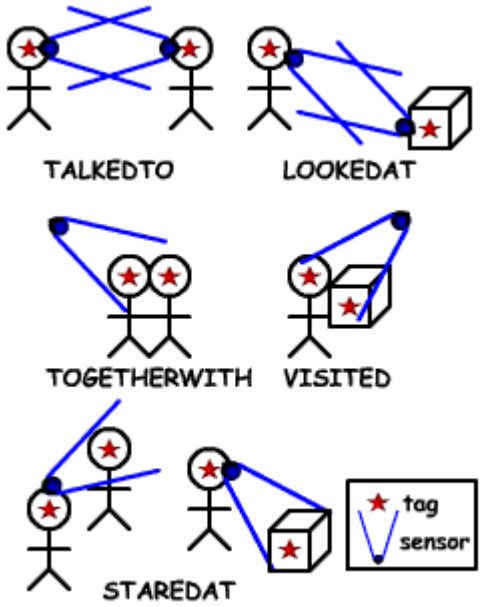

Figure 5. Basic Interactions.

- 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*.

An interaction among three or more users/objects can be inferred by one or more events with overlapping time intervals. We plan to define more complex interactions based on the combination of basic interaction events. This approach should allow more flexibility than

providing extra parameters to define interactions in more detail. For example, in a group discussion among three people, UserA, B, and C, the combination of TALKEDTO and TOGETHERWITH events among UserA, B, and C will signify that they were doing something together. In addition, if one of the users also has LOOKEDAT or VISITED event within the same time interval, it is possible to infer the location in which the group discussion took place.

5. Scene Extraction

A scene is made up of several basic interaction events and is defined based on time. Because of the setup of the exihibition room in which five separate booths had 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.

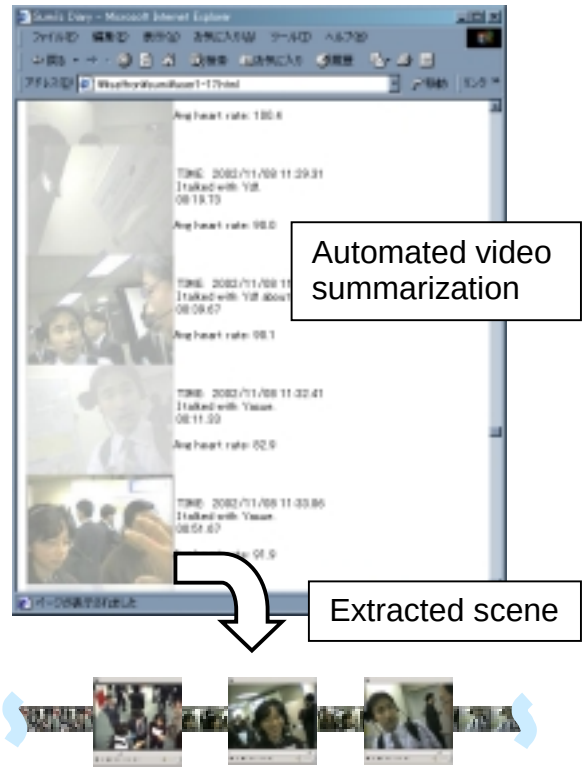

Figure 6. HTML output of scene videos.

6. Scene Video Production

Scene videos were created in a linear time fashion using only one source of video at a time. Other potential forms considered were a linear time fashion with multiple video sources (in panes) and a non-linear fashion in which important events are shown more than once using different viewpoints. In order to decide which video source to use to make up the scene video, we established a priority list. In creating the priority list, we made a few assumptions. One of these assumptions was that the video source of a user associated with the CAPTURED event of UserA shows the close-up view of UserA. Another assumption was that all the components of the interactions occurring in BoothA are captured by the ubiquitous video cameras set up for BoothA.

The actual 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 simply composed of all audio sources of users and objects that are part of each scene. Although all the computers were synchronized with the same time server, the lag between the execution of the record command and the actual start of the recording caused on average 0.5 sec misalignment. This misalignment cannot be fixed in absolute means. We have implemented a function in which audio sources from different computers are compared in attempt to aligning the sound. However, this process is time-consuming and was not used. Therefore, the resulting audio for the scene video often has an echo

if the same sound was recorded by more than one source. As a solution to the time synchronization problem, we plan to implement time stamps within the video feed, which can be extracted later.

7. Summary Video Production

The purpose of the summary video was to provide a quick overview of all the events the users experienced. We used a simple format in which at most 15 seconds of each scene was put together chronologically with fading effects between the scenes.

8. Spatial Data

The discussion thus far was mainly based on the temporal data. In addition to the temporal data, the IR tracker sends x-y coordinates of tags detected in its view. Although this feature is not fully implemented into the system, the spatial data can be powerful indicator of details of the interaction. The initial step to using the spatial data will be the alignment of video camera and IR tracker. Because IR trackers and video cameras have separate optical input, though they are next to each other, the two do not have an absolute relationship. The spatial data will allow us to determine the positions of each person during an interaction. This will be useful in the analysis of interactions in which the focus of attention plays an important role.

9. Conclusions

At the two-day demonstration of the system, we were able to provide users with video clips of their interactions on the fly. We have provided a powerful use for infrared LED signals and infrared tracking cameras as a tracking and tagging system. The system is useful for researchers studying human interactions as well as human-computer interactions. In near 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 will collaborate with interaction researchers to improve our interaction pattern recognition. Once the interactions have been indexed, audio and image processing can augment the data when more computing time is available and when detailed analyses are necessary.

Acknowledgement

We would like to thank Sadanori Ito, Tetsushi Yamamoto, and Sidney Fels for their valuable contributions to this project. 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. Rainer Stiefelhagen, Jie Yang, Alex Waibel. Towards Tracking Interaction Between People. AAAI Spring Symposium on Intelligent Environments, Stanford University, California, March 23-25, 1998. AAAI Technical Report SS-98-02, pp. 123-127.
2. Cory D. Kidd and Robert Orr and Gregory D. Abowd and Christopher G. Atkeson and Irfan A. Essa and Blair MacIntyre and Elizabeth Mynatt and Thad E. Startner and Wendy Newstetter. The aware home: A living laboratory for ubiquitous computing research. *Proceedings of CoBuild'99*, 1999, 190-7.
3. Barry Brumitt and Brian Meyers and John Krumm and Amanda Kern and Steven Shafer. EasyLiving: Technologies for intelligent environments. *Proceedings of HUC 2000*, 2000, 12-29.