Collaborative Capturing and Interpretation of Interactions

Yasuyuki Sumi†‡ Sadanori Ito‡ Tetsuya Matsuguchi‡ß Sidney Fels¶ Kenji Mase§‡

†Graduate School of Infomatics, Kyoto University
‡ATR Media Information Science Laboratories
¶The University of British Columbia
§Information Technology Center, Nagoya University
ßPresently with University of California, San Francisco

sumi@i.kyoto-u.ac.jp, http://www.ii.ist.i.kyoto-u.ac.jp/~sumi

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.

KEYWORDS: interaction corpus, experience capturing, ubiquitous sensors

INTRODUCTION

Weiser proposed a vision where computers pervade our environment and hide themselves behind their tasks[1]. 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 (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[2]. We also employ autonomous physical agents, like humanoid robots[3], 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

[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.

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.

CAPTURING INTERACTIONS BY MULTIPLE SENSORS

We developed a prototype 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 two-day research laboratories' open house.

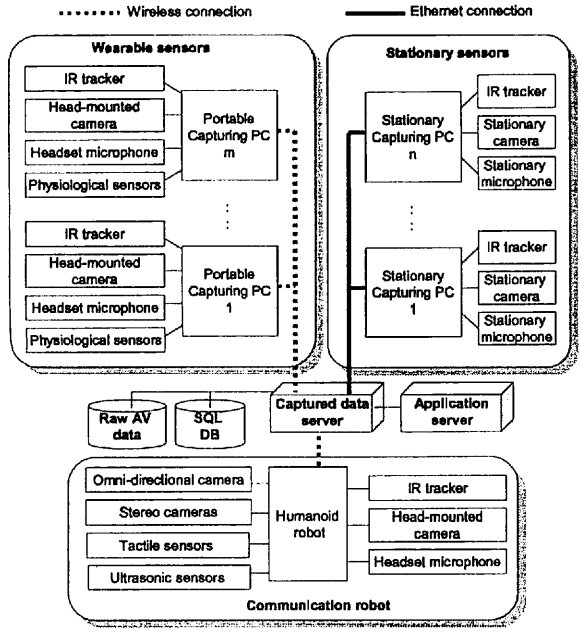


Figure 1: Architecture of the system for capturing interactions.

Figure 1 illustrates the system architecture for collecting interaction data. The system consists of sensor clients ubiquitously set up around the room and wearable clients to monitor humans as subjects of interactions. Each client has a video camera, microphone, and IR tracker, and sends the data to the central data server. Some wearable clients have physiological sensors.

Principal data is video data sensed by camera and microphone. Along the video stream data, IDs of the LED tag captured by the IR trackers and physiological data are recorded in the database as indices of the video data.

The humanoid robots in the room record their own behavior logs and the reactions of the humans with whom the robots interact.

RELATED WORKS

There have been many works on smart environments for supporting humans in a room by using video cameras set around the room, e.g., the Smart rooms[4], Intelligent room[5], AwareHome[6], Kidsroom[7], and EasyLiving[8]. The shared goal of these works was recognition of human behavior using computer vision techniques and understanding of the human's intention. On the other hand, our interest is to capture not only an individual human's behavior but also interactions among multiple humans (networking of their behaviors). We then focus on the understanding and utilization of human interactions by employing an infrared ID system to simply identify the human's existence.

There also have been works on wearable systems for collecting personal daily activities by recording video data, e.g., [9] and [10]. Their aim was to build an intelligent recording system used by single users. We, however, aim to build a system collaboratively used by multiple users to capture their shared experiences and promote their further creative collaborations. By using such a system, our experiences can be recorded by multiple viewpoints and individual viewpoints will become obvious.

This paper shows a system that automatically generates video summaries for individual users as an application of our interaction corpus. In relation to this system, some systems to extract important scenes of a meeting from its video data were proposed, e.g., [11]. These systems extract scenes according to changes in the physical quantity of video data captured by fixed cameras. On the other hand, our interest is not to detect the changes of visual quantity but to segment human interactions (perhaps derived by the humans' intentions and interests), and then extract scene highlights from a meeting naturally.

IMPLEMENTATION

Figure 2 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, 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

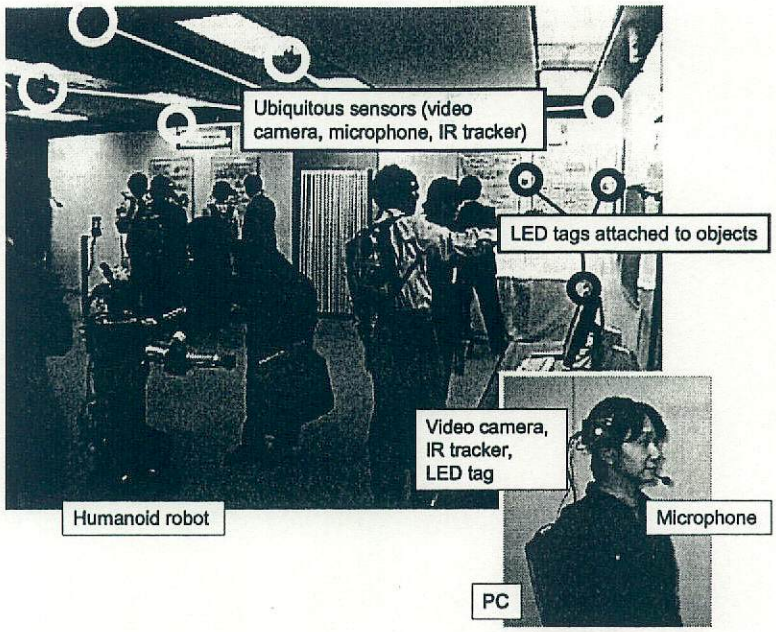


Figure 2: Setup of the ubiquitous sensor room.

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 UNIX file server via samba server. Index data given to the video data were stored in an SQL server (MySQL) running on another Linux machine. In addition, we had another Linux-based server, called an application server, for generating a video-based summary by using MJPEG Tools[2] .

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.

Figure 3 shows the prototyped IR tracker and LED tag. 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 (see Figure 4).

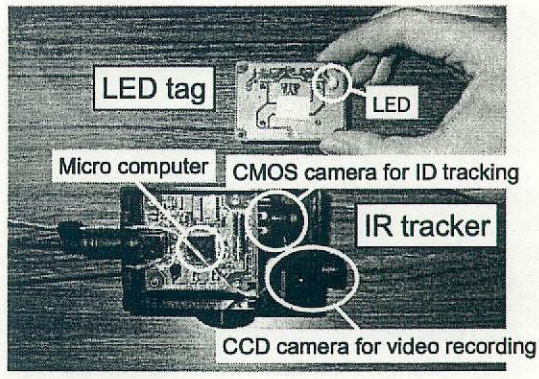
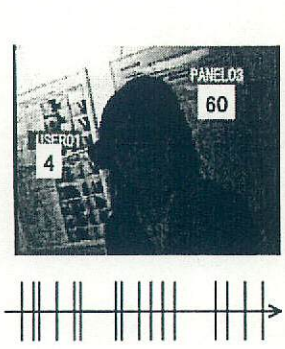


Figure 3: IR tracker and LED tag.

A few persons attached three types of physiological sensors – a pulse physiology sensor, skin conductance sensor, and temperature sensor – to their fingers[3] These

[2] A set of tools that can do cut-and-paste editing and MPEG compression of audio and video under Linux. http://mjpeg.sourceforge.net

[3] We used Procomp+ as an AD converter for transmitting



ID	TIME	X	Y
4	1036571603.137000	61	229
60	1036571603.448000	150	29
4	1036571603.878000	61	228
60	1036571604.319000	149	28
4	1036571604.659000	62	227
60	1036571605.440000	152	31
60	1036571605.791000	150	28
60	1036571606.131000	148	30
4	1036571606.472000	64	230
60	1036571607.163000	150	30
60	1036571608.074000	150	30
60	1036571608.385000	148	29
60	1036571608.725000	146	28
4	1036571609.066000	65	228

Figure 4: Indexing by visual tags.

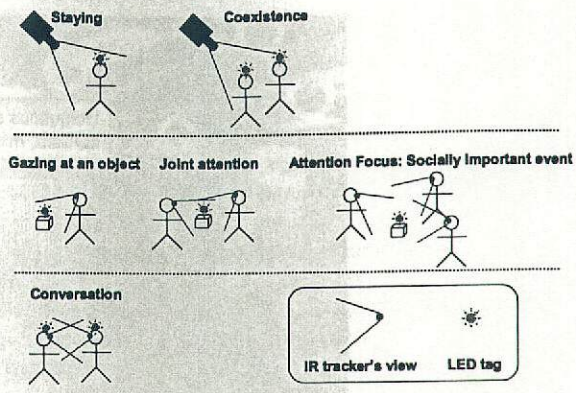


Figure 5: Interaction primitives.

data were also sent to the SQL server via the PC.

Eighty users participated during the two-day open house providing $\sim$ 300 hours of video data, 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.

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 5 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.

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.

sensed signals to the carried PC.

VIDEO SUMMARY

We were able to extract appropriate "scenes" from the viewpoints of individual users by clustering events having spatial and temporal relationships.

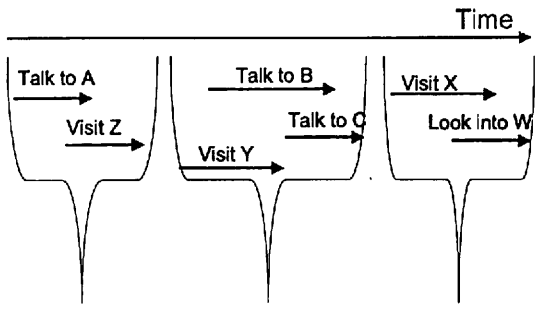


Figure 6: Interpreting events to scenes by grouping spatio-temporal co-occurences.

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 (see Figure 6).

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. In creating the priority list, we made a few assumptions. One of these assumptions was that the video source of a user associated with a 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 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, the ubiquitous video camera source is used.

Figure 7 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.

CORPUS VIEWER: TOOL FOR ANALYZING INTERACTION PATTERNS

The video summarizing system was intended to be used as an end-user application. Our interaction corpus is also valuable for researchers to analyze and model human social interactions. In such a context, we aim to develop a system that researchers (HCI designers, social scientists, etc.) can query for specific interactions quickly with simple commands that provides enough flexibility to suit various needs. To this end, we prototyped a system called the Corpus Viewer, as shown in Figure 8.

This system first visualizes all interactions collected from the viewpoint of a certain user. The vertical axis is time. Vertical bars correspond to IR trackers (red bars) that capture the selected user's LED tag and LED tags (blue bars) that are captured by the user's IR tracker. Many horizontal lines on the bars imply IR tracker data.

By viewing this, we can easily grasp an overview of the user's interactions with other users and exhibits, such as mutual gazing with other users and staying at a certain booth. The viewer's user can then select any part of the bars to extract a video corresponding to the selected time and viewpoint.

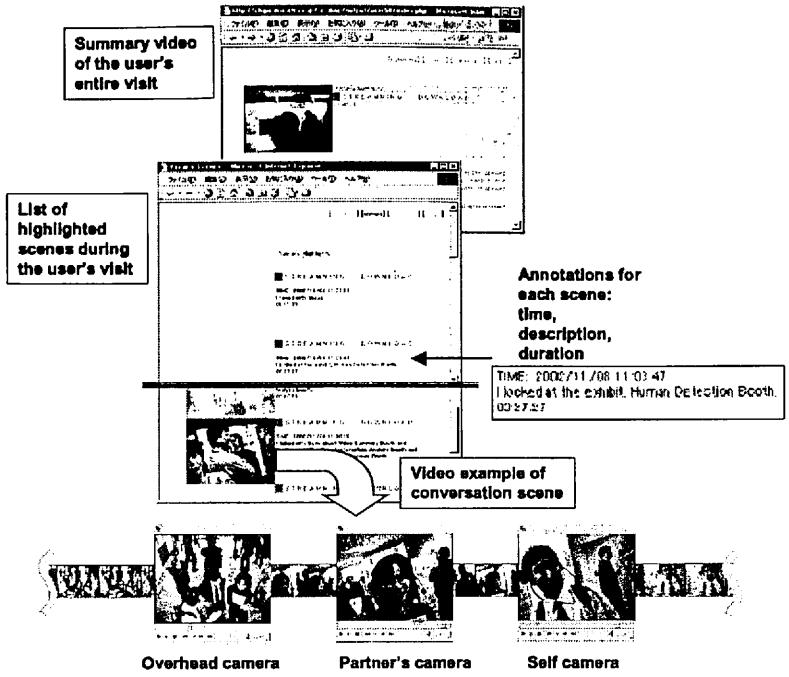


Figure 7: Automated video summarization.

We have just started to work together with social scientists to identify patterns of social interactions in the exhibition room using our interaction corpus augmented by the Corpus Viewer. The social scientists actually used our system to roughly estimate sufficient points from a large amount of data by browsing clusters of IR tracking data.

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. We also developed a prototype system to help social scientists analyze our interaction corpus to learn social protocols from the interaction patterns.

ACKNOWLEDGEMENTS

We thank our colleagues at ATR for their valuable discussion and help on the experiments described in this paper. 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. Mark Weiser. The computer for the 21st century. *Scientific American*, 265(30):94–104, 1991.
2. Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing. In *ACM Multimedia '99*, pages 3–10. ACM, 1999.
3. Takayuki Kanda, Hiroshi Ishiguro, Michita Imai, Tetsuo Ono, and Kenji Mase. A constructive approach for developing interactive humanoid robots. In *2002 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2002)*, pages 1265–1270, 2002.
4. Alex Pentland. Smart rooms. *Scientific American*, 274(4):68–76, 1996.
5. Rodney A. Brooks, Michael Coen, Darren Dang, Jeremy De Bonet, Josha Kramer, Tomás Lozano-Pérez, John Mellor, Polly Pook, Chris Stauffer, Lynn Stein, Mark Torrance, and Michael Wessler. The intelligent room project. In *Proceedings of the Second International Cognitive Technology Conference (CT'97)*, pages 271–278. IEEE, 1997.
6. Cory D. Kidd, Robert Orr, Gregory D. Abowd, Christopher G. Atkeson, Irfan A. Essa, Blair MacIntyre, Elizabeth Mynatt, Thad E. Startner, and Wendy Newstetter. The aware home: A living laboratory for ubiqui-

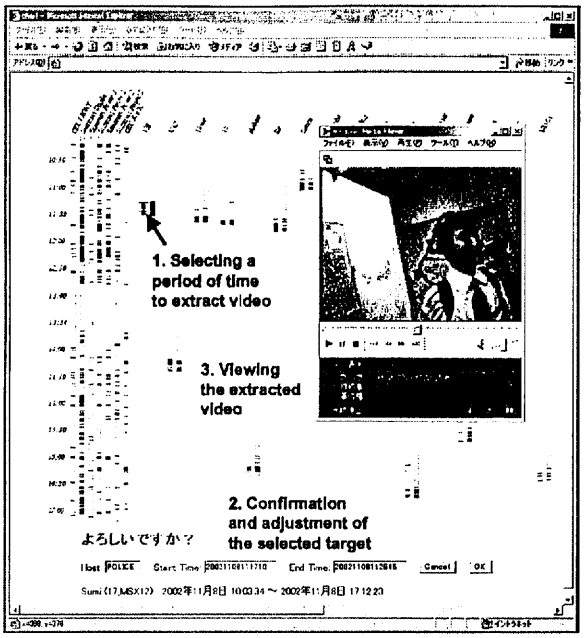


Figure 8: Corpus viewer for facilitating an analysis of interaction patterns.

tous computing research. In *Proceedings of CoBuild'99 (Springer LNCS1670)*, pages 190–197, 1999.

7. Aaron F. Bobick, Stephen S. Intille, James W. Davis, Freedom Baird, Claudio S. Pinhanez, Lee W. Campbell, Yuri A. Ivanov, Arjan Schütte, and Andrew Wilson. The KidsRoom: A perceptually-based interactive and immersive story environment. *Presence*, 8(4):369–393, 1999.
8. Barry Brumitt, Brian Meyers, John Krumm, Amanda Kern, and Steven Shafer. EasyLiving: Technologies for intelligent environments. In *Proceedings of HUC 2000 (Springer LNCS1927)*, pages 12–29, 2000.
9. Steve Mann. Humanistic intelligence: WearComp as a new framework for intelligence signal processing. *Proceedings of the IEEE*, 86(11):2123–2125, 1998.
10. Tatsuyuki Kawamura, Yasuyuki Kono, and Masatsugu Kidode. Wearable interfaces for a video diary: Towards memory retrieval, exchange, and transportation. In *The 6th International Symposium on Wearable Computers (ISWC2002)*, pages 31–38. IEEE, 2002.
11. Patrick Chiu, Ashutosh Kapuskar, Sarah Reitmeier, and Lynn Wilcox. Meeting capture in a media enriched conference room. In *Proceedings of CoBuild'99 (Springer LNCS1670)*, pages 79–88, 1999.