Interaction Corpus for Experience Sharing using Ubiquitous Experience Media

Kenji Mase[†*$], Yasuyuki Sumi[‡*], Megumu Tsuchikawa[*$]
Kiyoshi Kogure[$], Norihiro Hagita[$]

[†] Nagoya University, [*]ATR MIS Labs., [$]ATR IRC Labs., [‡]Kyoto University
mase@nagoya-u.jp

Abstract

A ubiquitous computing environment will become a solid social infrastructure for recording human experiences in the real world and facilitating human activities. The captured activity can be used as a good source of novel communication in the infrastructure. This paper focuses on the non-cumbersome recording, abstracting, and summarizing of experiences as well as sharing these experiences among people. Several devices are developed, such as a ubiquitous and wearable interaction system for recording and a humanoid-robot artificial partner. These are called *ubiquitous experience media (UEM)* that capture, display and facilitate experiences. The recorded experiences are stored in an *interaction corpus* after indexing and segmentation. The functions of UEM and the structure of the interaction corpus are described along with several experience sharing application scenarios.

Keywords: interaction corpus, ubiquitous experience media, experience sharing, artificial partner

1. Introduction

People have developed and used various media to record and present to others their experiences: pen and paper, photography, video recording, and so on. As computers acquire a huge amount of memory capacity in ubiquitous computing environments, computational media will be used as novel and richer media in the near future to support everyday human memorizing and communication activities [1][2].

Capturing large-scale interaction data through ubiquitous computing technology is the very first step for this purpose. In this course, we have proposed the construction of an interaction corpus, which is a semi-structured set of a large amount of interaction data captured by various sensors [3]. The sensors include video cameras, microphones, and other activity sensors to monitor humans as the subjects of the interactions. Thus multimedia information is captured to record the memory of experiences. More importantly, ID tags with an infrared LED and infrared signal-tracking device are incorporated to record the positional and situational contexts very easily along with the audio/video data [4]. With such contextual indices, we can structure the interaction data efficiently and effectively. We aim to use this corpus as a collection of experience elements to describe past and future experiences with other people. Since the captured data is segmented into primitive behaviors and annotated semantically, it is easy to collect highlighted actions, for example, to generate reconstructed diary content [5]. The corpus can, of course, also serve as an infrastructure for researchers to analyze and model social protocols of human interactions.

Research should also cover the media interface design issues of such devices because particular experience activity can be assisted by an association with embodied objects. We are developing artificial partners such as a humanoid robot that can record people's activities with various embedded sensors, including video cameras, microphones, touch sensors, positioning sensors, and so on [6]. A wearable sensor client is also proposed to provide people with “first-person” personal and mobile sensing as well as an environmental client for “third-person” public and stationary sensing. These are integrated and function together to construct experience media for people in a ubiquitous manner. Thus, we call these devices *ubiquitous experience media (UEM).*

A computerized memory aid is a challenging but promising application of the interaction corpus. Various high-density and multi-media digital recording devices have been developed and will continue to be improved. It is nearly possible to record and store the whole life of a person by video and audio onto a small magnetic disk memory [7][8]. The artificial structure of the interaction corpus may lead to a new theory for organizing human memory properly and using memory for exchanging information with other people. Even a system to support simple briefing and reporting of events and experiences to one’s colleagues at places such as business offices, hospitals and schools would be very helpful.

In the following section, the design and development of the interaction corpus and the ubiquitous experience media devices are described. First, the interaction corpus is proposed, followed by the description of the devices.

Next, an interaction interpretation method and its applications for interaction facilitation and experience sharing are presented. Finally, the preliminary experimental results are shown and then analyzed in a discussion.

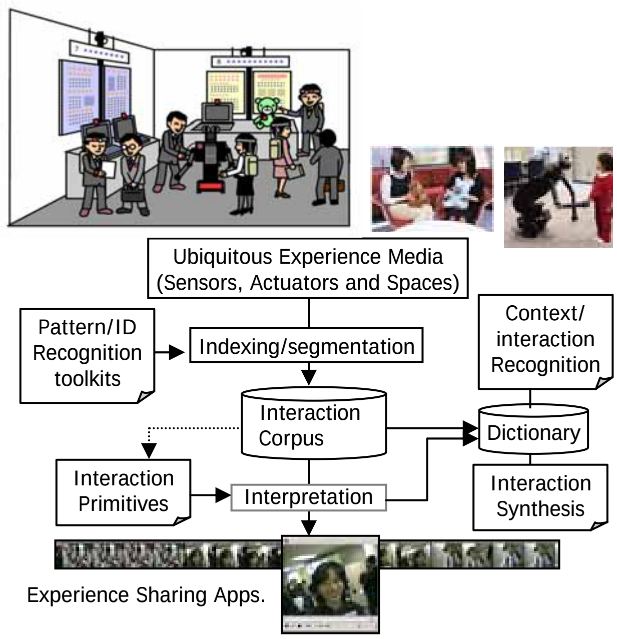


Fig. 1 Interaction Corpus

2. Interaction Corpus

An interaction corpus is 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 effectively anywhere and anytime without the burden of manual manipulation of data on computers. Figure 1 illustrates how the data from sensors is processed to develop the interaction corpus and how these data are used by applications. We call the environment consisting of sensors, actuators and spaces the UEM, where people have experiences, share experiences and recreate experiences. The issues of the interaction corpus include (i) sensing/interaction method, (ii) indexing/annotations, (iii) applications, (iv) dictionary, (v) structure of corpus, and (vi) display.

From sensing to indexing: The experience events are recorded by the UEM, which are described in Section 3. The recorded events and interactions are annotated automatically, semi-automatically or manually depending on the complexity of tasks and objects. For example, natural human movement such as door open/close may be captured automatically by using a door sensor. If we use ID emitting devices such as Infrared Identification (IrID) tags together with IrID receivers, relative positional information can be captured automatically and used to describe human interaction with objects. Intelligent vision system may provide the names of objects and people as annotations in a captured scene in the future. However, because automatic image-based object recognition is not practical at this moment, we use IrIDs instead of vision-based person/object identification. Discussion recording still needs manual highlighting of semantic topics during or after the meeting. We focus on context analysis rather than content analysis as an initial step. When we clearly understand interaction with syntactic structure, we can handle content with semantic analysis in the future. Annotation can be created personally or socially by a group of people. Both types are used to segment and annotate the interactions and to index each interaction. We describe the methods of indexing and annotating interactions and the hierarchical structure of the interaction corpus in Section 4.

Applications: The interaction corpus can also be used as a well-structured experience collection that is shared with other people for communication and creation of further experiences. Here, we demonstrate an application of generating a video-based experience summary that is reconfigured automatically from the interaction corpus. In creating an experience summary, rich annotations are very helpful in interpreting interactions of various granularity: staying in a location, meeting someone, taking speech turns, holding group meetings, maintaining joint-attention, etc. We show in a later section how the annotation/indexing processes are helpful for practical applications. Since the captured data are segmented into primitive behaviors and annotated in the higher layer, it is easy to collect the highlight actions, for example, to generate a reconstructed diary.

We aim to use this corpus, as a context-elements collection, to share past experiences with others. The corpus can, of course, also serve as an infrastructure for researchers to analyze and model the social protocols of human interactions statistically.

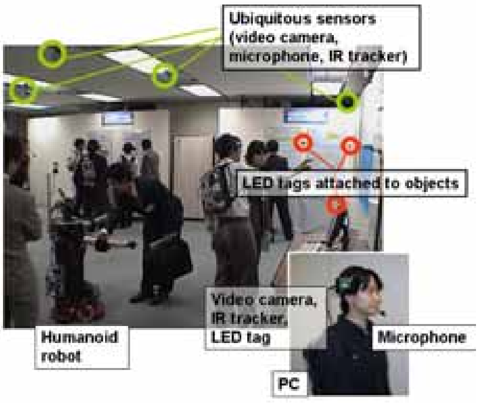

Ubiquitous sensors (video camera, microphone, IR tracker)
LED tags attached to objects
Video camera, IR tracker, LED tag
Microphone
Humanoid robot
PC

Fig. 2 Setup of UEM sensor room

Dictionary: A very large-scale accumulated corpus can provide an important infrastructure in the future digital society for both humans and computers. It will be used as a dictionary for intelligent robots/machines and the environment to understand verbal/non-verbal mechanisms of human interactions.

Data structure: We propose a four-layered data structure for the interaction corpus to manage the interactions systematically. This structure is employed from the analogy of language understanding as explained in a later section. Hierarchical structure is easy to design and utilize for wider applications. Its lower layers are designed to be universal, while the top layer is application-oriented.

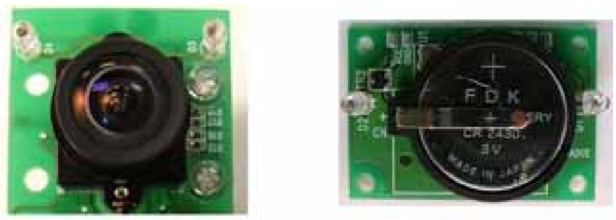


Fig. 3 IR Tracker (left) and IR-LED ID Tag (right)

3. Ubiquitous Experience Media

We prototyped a UEM system to record interactions among multiple presenters and visitors in an exhibition room. The prototype was installed and tested on the occasion of a public open laboratory event involving the presenters and visitors. Figure 2 is a snapshot of the exhibition room set up for the interaction corpus experiment. There were five booths in the room. Each booth had two sets of ubiquitous sensors that consisted of video cameras with IrID trackers and microphones. IrID tags were attached to possible focal points for social interactions, such as on the posters and displays for presentation. Each presenter at his/her booth carried a wearable client, consisting of a video camera with an IrID tracker (Figure 3), a microphone, an IrID tag, a throat microphone, and a head mounted display (HMD) as shown in Figure 4. The headset weights 500 g and the wearable PC backpack weighs 2720 g. A visitor could choose to carry the same wearable client as that of the presenters, just an IrID tag, or nothing at all. During the two-day event in 2003, 105 visitors tried the wearable client, providing about 290 hours of video and audio footage and 6,120,000 points of tracker data. Since 2002, we have captured 590 hours of data altogether.

Our approach to using the interaction corpus is characterized by the integration of many sensors (video cameras and microphones) ubiquitously set up around rooms and outdoors as well as wearable sensors (video cameras, microphones, and physiological sensors) to monitor humans as the subjects of the interactions. The IrID tracker gives the position and identity of any tag attached to an artifact or human in its field of view. By wearing an IrID tracker, the 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, such as humanoid robots, as social actors to proactively collect human interaction patterns by intentionally approaching humans. These devices compose the UEM, collaborative interaction media.

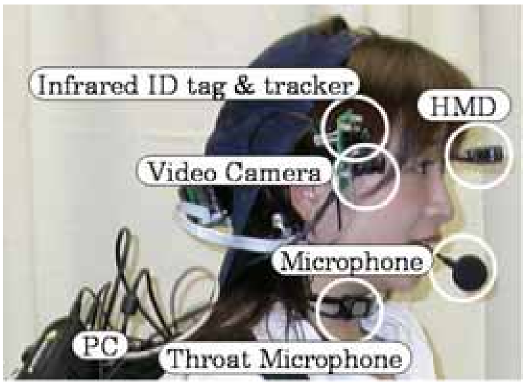


Fig. 4 Wearable sensors and HMD display

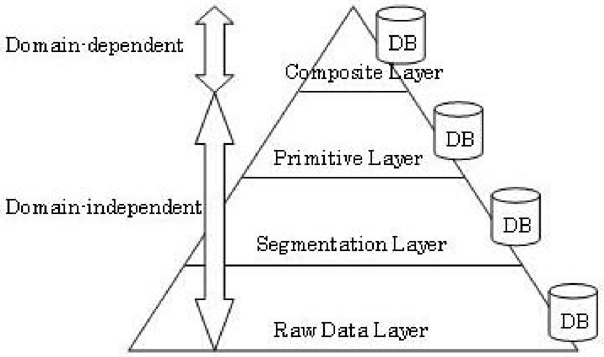


Fig. 5 Structure of Interaction Corpus

4. Interaction Structure

The captured interactions include occurrences such as visiting/leaving posters, presenting posters, listening to presentations, giving attention to things and other people, greeting, meeting and chatting with others, maintaining joint attention to things and others, and looking around.

All of the captured data are stored in the data server. These data are recorded with time information in a hierarchical structure as illustrated in Figure 5. We developed a method to segment interaction scenes from the IR tracker data. We defined interaction primitives, in the third layer, as elemental intervals or moments of activities. For example, a video clip that has a particular object (such as a poster or a user) in it constitutes an event. Since the locations of all objects are known from the IrID tracker and tags, it is easy to determine these events. We then interpret the meaning of events by compositing the primitives in the fourth layer. The hierarchical structure has the following meanings.

First, the lowest (first) layer, called the "Raw Data Layer," stores pairs consisting of raw sensor data and their time stamps. They are instantly obtained and used in applications right away. In the current implementation, IrID tracker, video, microphones (conventional and throat), and motion sensor are used, and their data are stored separately. The data represented in this layer can be thought of as phonemes or letters in human language.

The second layer is called the "Segmentation Layer." The data is segmented out automatically by pre-processing for each modality of data, such as IrID data stream and microphone volume. For IrID data stream, it stores fragments of very short gazing information in the Raw Data Layer. These are combined into a more meaningful cluster by connecting neighborhood gazing fragments if they are close to each other. Some isolated IrID data are considered noise and neglected. The clustering is done sequentially by giving a temporal label to a candidate fragment until it completes the labeling. For the microphone volume, it uses an ordinary threshold technique to define an utterance segment. Segmented data are still stored separately. The layer is analogous to "words" in human language.

The third layer, called the "Primitive Layer," stores interaction primitives, each of which is defined as a unit of interaction. In this layer, an interpretation of interaction is given to the Segmentation Layer data. This process is similar to morphological analysis of human language. We extract the interaction primitives based on gazing information from the segmented ID (tag) data and the utterance information from the segmented utterance. The former is a binomial interaction, while the latter is a monomial interaction. Using the multimodal segments in combination or alone, we define various interaction primitives. We think that both gazing and utterance provide very important data for finding human-human or human-object interactions. For example, when the IrID tracker of a user (A) gazes a tag of another person (B), we call the interaction primitive a "LOOK_AT" situation. When a tracker gazes at an object and an utterance is detected, the interaction is called "TALK_TO." Figure 6 shows some examples of the interaction primitives.

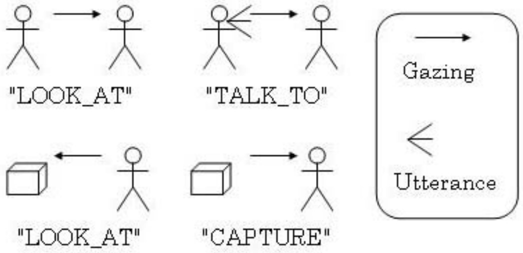


Fig. 6 Interaction primitives

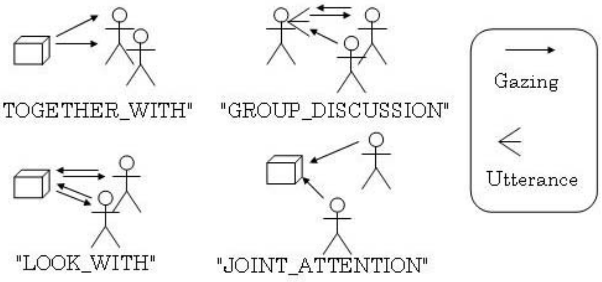


Fig. 7 Interaction Composites

The upper layer, called the "Composite Layer," stores the composite interactions that are more socially oriented and application-dependent. The layer is somewhat equivalent to a sentence in human language. By combining the composites of the interaction corpus, we can represent scenes of interaction. In the exhibition situation, for example, we store correlation interaction data between two or more clients (human's or object's data). If two people talk to each other, the interaction is recorded as "TALK_WITH." When three or more people have the "TALK_TO" or "LOOK_AT" primitives, the composite is interpreted as "GROUP_DISCUSSION." Figure 7 shows examples of composites.

The interaction corpus is created automatically, which makes it very useful for indexing huge quantities of video- and audio-based experience records. We can manipulate multimedia data easily with the interaction corpus.

These primitive notations need to be more fully developed in various contexts in order to utilize the interaction corpus more widely. Worldwide collaboration in this development is necessary based on the intentional capturing and analysis of interactions in a variety of situations.

5. Interaction Corpus Applications

The interaction corpus is used in many applications such as interaction facilitation by humanoid robots and/or head mounted display (HMD) and experience sharing by summarizing experiences. These were demonstrated in an exhibition setting with research poster presentations.

5.1 Interaction Facilitation

Interaction facilitation has two aspects. One is to facilitate interactions for the benefit of a capturing system. In our sensor room setting, wearable, mobile, and stationary sensors are provided. If the position of a user is not appropriate for a stationary sensor (camera, microphone, etc.) due, for example, to its distance and orientation, we may want to bring the user to a suitable location in order to record his/her voice and picture clearly for later use. We introduced facilitating entities (facilitators), such as robots and visual guides, to facilitate capturing. The other aspect is to facilitate people's interaction with other people and/or artifacts in order for people to enjoy talking with others and investigating the events. In our exhibition scenario, visit assistance is helpful to many visitors in visit planning and re-planning phases.

5.1.1 Robot Facilitation

In an exhibition situation, an assistant often helps a visitor pay attention to a specific event. In our prototype system, a humanoid communication robot, Robovie-II, served as such an assistant (facilitator). Robovie wears an IRID tag and tracker and engages in communication with the visitor in front by referring to the already accumulated interaction corpus in order to create new interaction. We have demonstrated several methods of interaction facilitation with Robovie, such as (i) calling the visitor's name as a greeting and to get his/her attention, (ii) guiding the visitor by gesture and voice to other places (booths) of his/her interest based on the room conditions such as being crowded (Figure 8), (iii) talking about other people and (iv) giving informative announcements about the site.

Robovie is very good at attracting the attention of people and guiding them to places because it has a humanoid upper body, eyes, head, and hands. As these body parts move very fast to make gestures very naturally, people are drawn to the robot's facilitation. These services are possible because the robot can discern the situation from the concurrent interaction corpus.

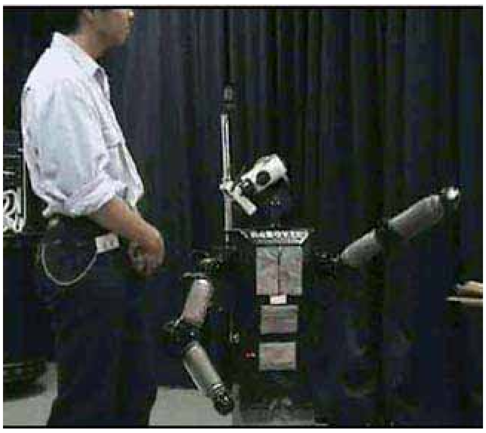


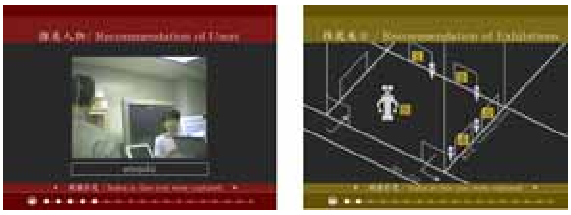
Fig. 8 Robovie's interaction facilitation

5.1.2 Visual and Auditory Facilitation

Robovie's interaction functions provide visitors with public and personal assistance. However, it is only given when the visitor encounters the robot. In order to provide a continuous facilitating service, we gave visitors an HMD with a wearable computer to provide them with visual information (Figure 9). The display recommends persons and booths that the visitor might like to meet and visit. A person of interest is selected from the past interaction corpus data based on the similarity of an interest vector. The vector is formed from the time stayed at each booth as an element. The similarity is computed by taking an inner product of the vectors. The booth recommendation is done similarly.

The HMD also shows augmenting information about the person and the booth in front of the user. The similarity rate of interest between the persons is calculated and displayed on the screen. The popularity of the visited booth is also displayed on the screen. The information is provided based on the following design principles.

1. No active user operation is required.
2. One kind of information is displayed on the screen at one time.
3. Color coding is employed to give assistance to the user's memory about the kind of information.
4. Information is shown in a short time to avoid disturbing the user's real-world focus.



(a) Person recommendation (b) Booth recommendation

Fig. 9 Visual facilitation by HMD

5.2 Experience Summary

We were able to extract appropriate "scenes" from the viewpoints of individual users by clustering events having spatial and temporal relationships. A 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 video's 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 a summary video for a quick overview of the events that the users experienced. To generate the summary video, we used a simple format in which relevant scenes of at most 15 seconds were assembled chronologically with fading effects between the scenes.

The event clips that were used to make up a scene were not restricted to those captured by a single source (video camera and microphone). For example, as shown in

Figure 9 for a summary of a conversational "talked with" scene, the video clips used were recorded simultaneously by the camera (c in the figure) worn by the user herself, the camera (b) of the conversation partner, and a fixed camera (a) on the ceiling that captured both users. Our system selected appropriate video clips to make a summary by consulting the volume levels of the users' individual voices. The LED tag worn by a user is assumed to indicate that his/her face is in the video clip if the associated IR tracker detects it. Therefore, the integration of video and audio from different worn sensors could generate the scene of a speaking face from one user's camera with a clearer voice from another user's microphone.

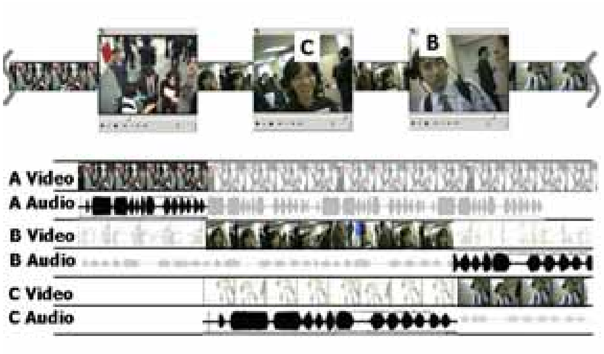


Fig. 9 Experience Summary by three clients

In the current system, the generated summary is a simple sequence of temporarily occurring events only bounded by the vocabulary of primitive scenes. The summary itself effectively expresses the experience. However, we think that a directed summarization would help a person to enthusiastically tell others about the experience. We have already developed the ComicDiary[9], a directed diary generation system using comic presentation techniques. It incorporates pre-defined stories by directors to generate story-telling based on a fraction of experience records stored in a PDA. We plan to integrate the system with the video summary system to generate a story-telling summary based on the interaction corpus.

6. Conclusion

This paper proposed a method for building an interaction corpus by using multiple sensors either worn or placed ubiquitously in the environment. Through a demonstration of our system, we were able to provide users with a video summary at the end of their experience, made on the fly. In the future, we will develop a system that will allow researchers to quickly query for specific interactions by using simple commands and that will provide enough flexibility to suit various needs.

An intelligent human interface has been discussed as an interface for a computer system that performs a certain task for humans. Future intelligent systems will become symbiotic partners of humans as tangible or embedded companions in daily life. A very important requirement of a new paradigm for future intelligent systems is that machine intelligence be built through interactions with humans.

Acknowledgments

Highly valuable contributions to this work were made by Tetsuya Matsuguchi, Sadanori Ito, Tetsushi Yamamoto, Sidney Fels, Shoichiro Iwasawa, Atsushi Nakahara and Yasuhiro Katagiri. This research was partially supported by the National Institute of Information and Communications and the Strategic Information and Communications R&D Promotion Programme.

References

[1] M. Lamming and M. Flynn: "Forget-me-not" Intimate computing in support of human memory", Proceedings of International Symposium on Next Generation Human Interface '94FRIEND21, pp. 150-158 1994.

[2] T. Kawamura, Y. Kono, and M. Kidode, "Wearable interfaces for a video diary: Towards memory retrieval, exchange, and transportation", The 6th International Symposium on Wearable Computers (ISWC2002), pp. 31-38, 2002.

[3] K. Mase and Y. Sumi: "Interaction Corpus and Experience Sharing," ATR Workshop on Ubiquitous Experience Media 2003, pp. 83-87, Sept. 2003.

[4] Y. Sumi, T. Matsuguchi, S. Ito, S. Fels and K. Mase: "Collaborative Capturing of Interactions by Multiple Sensors," Adjunct Proceedings of the Fifth International Conference on Ubiquitous Computing (UbiComp2003) pp. 193-194, Seattle, Oct. 2003.

[5] Brian Clarkson, Kenji Mase and Alex Pentland: "The Familiar: a living diary and companion," CHI2001 extended abstracts, pp. 271-272, Seattle, April 2001.

[6] N. Hagita, K. Kogure, K. Mase and Y. Sumi, "Collaborative capturing of experiences with ubiquitous sensors and communication robots," Proc. 2003 IEEE International Conference on Robotics and Automation, pp. 4166-4171, Taipei, Taiwan, Sept. 2003.

[7] http://www.darpa.mil/ipto/Programs/lifelog/

[8] http://research.microsoft.com/barc/mediapresence/MyLifeBits.aspx

[9] Yasuyuki Sumi, Ryuuki Sakamoto, Keiko Nakao, Kenji Mase: ComicDiary: Representing individual experiences in a comics style, *Ubicomp 2002*, in G.Borriello, L.E.Holmquist (Eds.), Ubicomp 2002, LNCS 2498, pp.16-32, Springer, Sep. 2002.