Indexing Shared Experience Videos by Detecting Conversation Groups

Kai Toyama
Future University Hakodate
Email: k-toyama@sumilab.org

Yasuyuki Sumi
Future University Hakodate
Email: sumi@acm.org

Abstract—We propose a system to aid the browsing of shared experience data that includes multiple first-person view videos, to avoid boredom induced by having to watch through lengthy videos. Our system aids browsing by showing indices on the video seek-bar based on conversational field information. Conversational field contains information of participants and approximate location of group conversations. We use auditory similarity to detect conversational fields, because this method can detect the dynamics of groups in crowded areas. We conduct an experiment to evaluate the effect of provided indices to decrease the time for finding the specified scene from lifelog videos. Our experimental results suggest that our system can aid the browsing of videos that include one’s own experience, but cannot be proven to aid the browsing of unknown data.

I. INTRODUCTION

We propose a system that helps quickly browsing videos capturing social events by indices of participants’ names and approximate location. The indices are given by our previously proposed system, Neary [1], to detect conversational fields based on similarity of auditory situation among the participants.

The similarity of auditory situation between each pair of the users is measured by the similarity of frequency property of sound captured by microphones of the individual users’ smartphones. Our preliminary experiments show this method can successfully distinguish groups of conversations and track dynamic changes of them. In this paper, we apply Neary to detect user contexts during experience sharing in daily social activities such as attending conferences.

We define conversational field, a topological area where multiple persons join the same conversation. Recognizing group activities like group conversation is one of the important issues to enable context-aware applications for enriching our social activities. We aim to realize a context-aware application based on recognized group situation, i.e., a quick browser of our users’ lifelog videos.

Recently, advancing camera technologies enable us to record lifelog videos for a long period of time. However, it is difficult to locate significant events quickly in such lengthy videos. It is necessary to provide cues that indicate events within the videos. We consider the conversation partners and their approximate location to be important information for recalling their memories. This paper shows a quick browser of such lifelog videos by deploying Neary to detect the partners of group conversations in daily activities.

II. RELATED WORK

Edward Hall [2] introduced a concept called proxemics, measurable distances between people as they interact. Many researchers in the domain of ubiquitous computing try to detect social contexts by estimating the users’ position and mutual orientation based on proximity detection by infrared tags [3], [4], location detection by signal intensity of WiFi access points [5], visual tracking of groups of people [6], [7], etc.

Physical clusters of people could be candidates of conversational field. However, it would be difficult to determine conversational field according to size of the clusters because the physical size of conversational fields would easily vary depending on size and shape of space, crowdedness of people, and situation of social activities. Our system, Neary, detects conversational fields based on similarity of auditory situation among users. The similarity of auditory situation between each pair of the users is measured by the similarity of frequency property of sound captured by body-worn microphones of the individual users.

Intuitively explaining, users whose microphones receive similar sound (voice of a certain person, ambient sound, etc.) are regarded as the members of a conversation. In this method, people situated in the same sound environment are naturally grouped in the same conversational field, not depending on its physical size. The conversational fields detected by this method match to the granularity of our social activities such as meeting, lectures, group touring, etc. The method is also adjustable to various size of conversational fields from ad-hoc chatting to a lecture in a big hall.

There have been some works aiming to estimate ad-hoc groups based on ambient sound similarity. Aoki et al. [8] proposed a method to detect conversation groups from colocated multiple simultaneous conversations. Their method needs prior training with users’ speech data. Wirz et al. [9] shares the aim and approach with us and reported detailed performance evaluation of their method of proximity estimation. We are more interested in application development with simple and light implementation and this paper aims to provide practical findings from our trials in various fields.

Neary is implemented with a simple algorithm and runs on portable PCs. Experimental result shows Neary can success-
fully distinguish groups of conversations and track dynamic changes of them. This paper aims to deploy Neary to track users’ participation in conversational fields in daily activities and provide them with a quick browser of their lifelog videos.

III. DETECTION OF CONVERSATIONAL FIELDS AND ITS APPLICATION

A. Detecting Conversational Fields

1) System Overview: Figure 1 shows our vision of context-aware applications using conversational fields detection in this section. Neary, our previously proposed system, used a small sized computer and detected conversation groups by using Peer to Peer (P2P), but, it is difficult that other systems access conversational fields information. Besides, this method cannot detect conversational fields. Therefore, we used the client-server model, because systems can access conversational fields information in real-time or not.

Additionally, we try to detect approximate locations of conversation groups by installing a device that is the same as the device that the participants wear. The device, which is installed on an object or some part of the environment, is called an environmental-node and it enables us to recognize the position and physical size of conversational fields. Meanwhile, the device-wearing participants are called human-nodes. In this paper, we worked on video indexing system as one of the system, depending on conversational fields based on auditory similarity.

2) Sound Sensing by Mobile Devices: We outline our method of sensing by mobile device for running our client software. We used the Nexus5 as a mobile device, having considered the frequency response of the microphone, because different microphone could record different sound. The microphone on the Nexus5 was used to collect sound data.

Figure 2 shows the state of people wearing the device as human-nodes and the state of the device installed as an environmental-node on a table. Installing the device on the environment enables recognition of positions of conversation groups and size of conversational field, because if human-nodes and the environmental-node on the table are judged to be the same group by our method based on auditory similarity, we can deduce that people are near the table.

3) Detection of Conversation Groups: We describe the detection algorithm based on auditory similarity. Our algorithms are based on those of Neary, because Neary has a sufficiently high precision ratio. However, as the recall ratio of Neary is low, we modified some of the parameters. In addition, the algorithms used in Neary do not perform smoothing, so we also modified some of the algorithm.

In Neary, first, the algorithm obtain the similarity between each pair of devices and judge whether or not there are conversation groups, based on the threshold. The threshold is defined as 0.775, but, we changed it to 0.75, because we use different microphone-equipped smartphone and thus, we should change the adopted values.

Next, we try to perform smoothing. Conversational field information is obtained once every second. If the number of the degree of similarity exceeding the threshold is majority until 80 seconds before, these devices are considered co-located. In addition, perform smoothing with conversational field as two values. However, these parameters were decided with the data, and this algorithm generates time-lag between judgement result and videos, so we synchronized the video to the result.

B. Indexing Shared Experience Videos by Detected Conversation Groups

Figure 3 shows our proposed system that aids the browsing of shared experience videos. This system shows an index on the seek-bar and visualizes conversational fields as a node graphs. Our users can browse multiple videos by an index based on the member, or conversation groups, or their position.

The index is the list of cues for the video and it is expressed in n!-colors (n is the number of participants) depending on conversational fields that contain the information of the conversation group members and their positions. Accordingly, if there are no groups the index is colorless. We explore scenes in multiple first-person videos using the index as a cue.

In addition, if the user searches for a certain conversation group member, this system shows only the index including the chosen member. Thus, if the member who has been searched for does not belong to any conversation groups, the index is expressed as colorless.
IV. EXPERIMENTAL EVALUATION

A. Experiment

We compare our proposed system to existing video playback software (baseline). We measured the time taken to complete some assigned tasks and evaluated by conducting a significance test between our proposed system and the baseline software.

We recruited 3 participants (subjects 1, 2 and 3), and gave them 8 tasks on 2 datasets (datasets A and B). We observed subjects to see their reaction to our proposed system, to confirm whether conversational field information is useful or not. Finally, we conducted a semi-structured interview.

1) Datasets Used for Experiment: We prepared 2 datasets (datasets A and B). These data were recorded in a poster presentation session, because we think our proposed method can detect conversational fields in a crowded area, and detect the dynamics of conversational fields easily, as participants move to listen to a presentation.

Figure 4 shows the situation of the poster presentation in dataset A. Further details of datasets are given in Table I.

![Fig. 4. Circumstances of poster presentation in dataset A](image)

We assigned 4 tasks for each dataset. The 4 tasks for dataset A were referred to as A-1, A-2, A-3 and A-4, likewise for dataset B.

Dataset A included a video that was recorded by subject 1. However, subjects 2 and 3 did not participate in the presentation in dataset A. In other words, dataset A was not familiar to them. Meanwhile, dataset B included a video that was recorded by subjects 2 and 3. Accordingly, subject 1 was unfamiliar with the videos in dataset B.

2) Finding Task for Specified Scene: We prepare tasks such as determining “When did participant A converse with B in the poster X”. Figure 5 shows an example of solving a task.

When given the task question “When did participant A converse with participant B”, one should watch multiple videos recorded by participants A and B, then make a judgement based on their video image and voices. If it is uncertain whether they conversed, it is necessary to watch other videos recorded by other subjects from a third-person viewpoint.

![Fig. 5. Example of finding task for specified scene](image)

TABLE I

| Details of the datasets: The video was recorded by participants wearing human-nodes. |
|--------------------------------|------------------|
| Participants | dataset A | dataset B |
| Presentation | 9 | 61 |
| Human-Node | 6 | 6 |
| Environmental-Node | 3 | 2 |
| Videos | 3 | 5 |
| Video Length | 01:12:00 | 00:57:31 |

![Fig. 3. Our proposed system: This system shows an index and node-graph based on conversational fields](image)
B. Results of time taken to complete tasks

Task completion time for each task and subjects is presented in Figure 6 and Figure 7. The cells colored green in Figure 6 indicate the measured amounts of time taken to complete the task with the help of our proposed system. The figures in red are the measured times the subjects took to complete the task when watching their own data.

Figure 7 shows a comparison of the proposed system and the baseline software. Conducting a statistical significance test ($p < 0.05$) revealed that, in the case of browsing other people’s video data, there was little difference between the proposed system and the baseline software, because the measured time did not vary widely.

In contrast, we observed a significant difference in the case of subjects browsing videos that included their own data (see the light blue and orange bars in the graph in Figure 7). As a result, it was confirmed that our proposed system can aid the watching of videos that include the user’s own data.

C. Observation and interview

We observed the subjects to confirm their reactions while using our system. We noted that, it seemed to be difficult for them to judge who someone converse with, because it is hard to confirm conversation groups from videos. Besides, our index information is nonfigurative and we defined conversational field as groups hearing the same voice or environmental sound. In other words, we regard hearing and conversation as the same, thus, the subjects were confused by the task.

After the tasks were completed, we conducted a semi-structured interview with the subjects. First, we asked them about the indexes, and received the comment that “The index is almost correct”. Next, we asked them why they had difficulty completing the task, and were told “It is difficult to judge instantly, because these videos occasionally losing conversation partner”, and “I was confused by the expression of the task”.

V. CONCLUSIONS

We proposed a system to aid the browsing of shared experience data that includes multiple first-person view videos, to avoid boredom induced by having to watch through lengthy videos. Our system aids browsing by showing indices on the video seek-bar based on conversational field information. Conversational field contains information of participants and approximate location of group conversations. We use auditory similarity to detect conversational fields, because this method can detect the dynamics of groups in crowded areas.

We conducted an experiment to evaluate the effect of provided indices to decrease the time for finding the specified scene from lifelog videos. Our experimental results suggest that our system can aid the browsing of multiple videos that include one’s own experience. On the other hand, the system has not been proven to aid the browsing of unknown data.

REFERENCES