Quick Browsing of Shared Experience Videos Based on Conversational Field Detection

Kai Toyama(✉) and Yasuyuki Sumi

Future University Hakodate, Hokkaido, Japan
k-toyama@sumilab.org, sumi@acm.org

Abstract. We propose a system to aid the browsing of shared experience data that includes multiple first-person view videos. Using this system, users can avoid the tedious task of searching through lengthy videos. Our system aids browsing by displaying situational information cues on the video seek-bar, and visualizing node graphs showing members participating in the scenes and their approximate location. Users of our system can search and browse events with the help of cues indicating participant names and their locations. We use auditory similarity to detect conversational fields in order to detect the dynamics of groups in crowded areas. We conduct an experiment to evaluate the ability of our system to decrease the time needed for finding specified scenes in lifelog videos. Our experimental results suggest that our system can aid the browsing of videos that include one's own experiences, but cannot be proven to aid the browsing of unfamiliar data.

Keywords: Smart video viewing · Information cues
First-person view videos · Conversational fields

1 Introduction

We propose a system that helps users to quickly browse videos capturing social events by providing cues showing the chronological history of conversation groups generated during the events. We use a method to detect conversational fields as cues based on the similarity of auditory situations among the participants.

We define conversational fields as a topological area in which multiple persons join the same conversation. As participants in the conversational fields, we consider not only people speaking but also people listening to them. Recognizing group activities such as group conversations is an important technique to enable context-aware applications for enriching social activities, e.g., groupware. For example, at conferences, there are activities of different spatial sizes; these include oral presentations, poster presentations, and social gatherings. We must recognize differences in spatial sizes and perform the appropriate service. The purpose of this paper is to realize an efficient browser for searching a user's lifelog data as a context-aware application based on recognized conversation fields.

K. Murao et al. (Eds.): MobiCASE 2018, LNICST 240, pp. 40–55, 2018.
https://doi.org/10.1007/978-3-319-90740-6_3

Because advancing camera technology enables us to record videos for long periods of time, we can record our experiences and share them with others. However, it is difficult to find a specific event from a lengthy video. We consider conversation partners and their approximate location to be important information for recalling memories. Therefore, we use conversational fields as a cue for browsing video.

The remainder of the paper is organized as follows. Section 2 provides an overview of the problem addressed in our research by presenting related work. Section 3 presents our system for quick browsing of shared experience videos as an application that uses detected conversational fields based on situated sound similarity. Section 4 presents an experimental evaluation showing that our system can provide the ability to quickly find any scene from lengthy lifelog videos. We describe limitations of our study and future work in Sect. 5. We conclude the paper in Sect. 6.

2 Related Work

2.1 Detecting Conversation Groups as a Social Context

Hall [1] introduced a concept called proxemics, i.e., measurable distances between people as they interact. Many studies in the domain of ubiquitous computing have attempted to detect social contexts by estimating users' positions and mutual orientations based on proximity detection via infrared tags [2,3], location detection according to signal intensity of Wi-Fi access points [4], the use of Bluetooth Low Energy (BLE) [5], visual tracking of groups of people [6,7], and various other techniques. In addition, Kendon [8] proposed F-formation as a measure of social interaction. F-formation detection [9–12] involves analyzing physical clusters of people as well as proximity detection.

Physical clusters of people could be candidates for conversational fields. However, it is difficult to determine conversational fields according to cluster size because the physical size of conversational fields can easily vary depending on the size and shape of the space, the crowdedness of people, and the particular social situation.

Previous works have aimed to estimate ad-hoc groups based on ambient sound similarity. Techniques that use sound to detect groups and their locations can be roughly divided into two categories: those that are analogous to fingerprinting [13–15] and those that analyze the similarity of each sound [16–18]. To utilize techniques analogous to fingerprinting, the environmental sound of each place must be recorded in advance. For example, Aoki et al. [19] proposed a method to detect conversation groups from collocated multiple simultaneous conversations. Their method needs prior training with users' speech data. In contrast, Wirz et al. [20], whose aim and approach are similar to ours, reported a detailed performance evaluation of their proximity estimation method. Their method records ambient sound on each microphone synchronously, and calculates similarities. 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.

Nakakura et al. [21] proposed a system called Neary, which detects conversational fields containing people in a conversation by comparing sound similarity among them. The similarity in auditory situations between each pair of users is measured according to the similarity in the frequency properties of sound captured by the users' head-worn microphones.

Intuitively, users whose microphones receive similar sounds (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, independent of its physical size. The conversational fields detected by this method match the granularity of social activities such as meetings, lectures, and group tours. The method is also adjustable to various conversational field sizes, from ad-hoc chatting to a lecture in a large hall (Fig. 1).

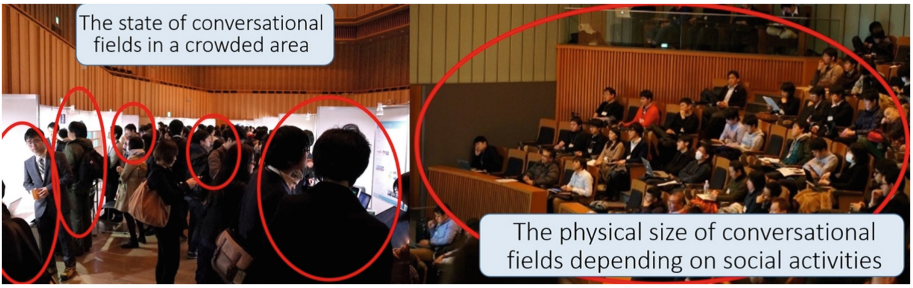

Fig. 1. Distance between participants and the physical size of conversational fields depending on social activity and situation.

Neary is implemented using a simple algorithm and runs on portable PCs. Preliminary experimental results show that Neary can successfully distinguish groups of conversations and track dynamic changes in them. This study aims to deploy Neary to track users' participation in conversational fields during daily activities, and provide a browser that can quickly search the users' lifelog videos.

2.2 Smart Video Viewing

There are two categories of techniques that enable users to browse videos quickly: fast-forwarding techniques [22,23] and video summarization techniques [24–27]. Fast-forwarding techniques reduce camera shaking and fast-forward the video by resampling frames. However, fast-forwarding does not consider the events in the video. On the other hand, video summarization techniques extract scenes in the video based on cue detection achieved by hand activity recognition [28–30], face recognition [31], and activity segmentation [32]. However, users cannot know the

context of an extracted scene. Additionally, if the detection result is wrong, it is possible that an important scene has been missed.

Higuchi et al. [33] proposed EgoScanning, which facilitates rapid browsing of egocentric videos. Hand activity, face recognition, and the movement of the person who recorded the video are used as cues; these are shown on the video timeline to indicate regions of detection. The aims of their study are similar to ours. We facilitate browsing using conversational fields as cues, and show them on the seek bar.

3 Detection of Conversational Fields and Its Application

3.1 Detecting Conversational Fields

System Overview. Figure 2 shows our vision of context-aware applications using conversational field detection, which is described in this section. Neary, our previously proposed system, used a small computer and detected conversation groups by using a peer-to-peer approach. We implemented a Neary server in this study in order to increase the flexibility of access to conversational field information by client applications.

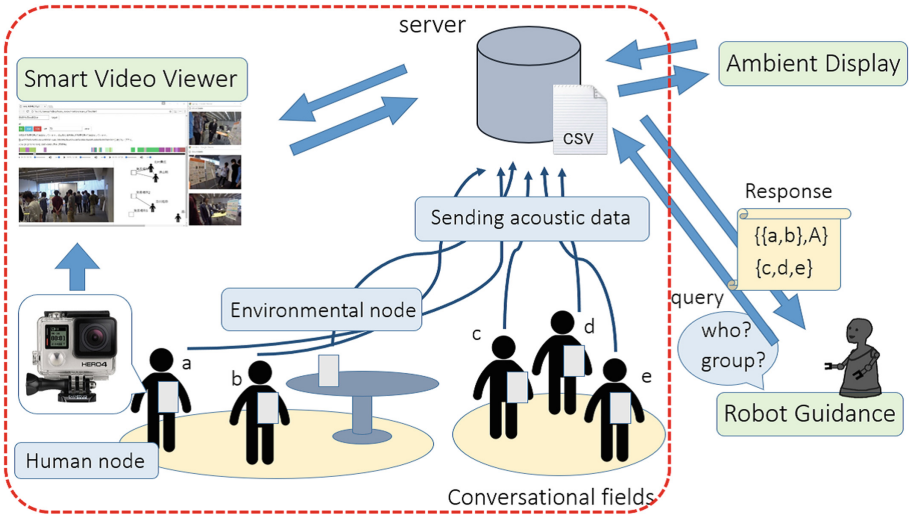

Fig. 2. Overview of our applications based on conversational field detection.

Additionally, we aim to detect approximate locations of conversation groups by installing the same device with ones worn by the participants. 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.

Our goal is to propose applications based on conversational field recognition systems, such as ambient displays or robot guidance. Hence, we use smartphones for detecting conversational fields. For this study, we built a smart video viewer system that depends on conversational fields based on auditory similarity.

Sound Sensing by Mobile Devices. In this subsection, we outline our method of sensing by mobile device, which enables us to run our client software. We used the Nexus 5 as a mobile device, after considering the frequency response of the microphone; in order not to worry about the differences in the sound frequency characteristics between devices, we made all devices Nexus 5.

Figures 3 and 4 respectively show people wearing the device as human nodes and the device installed as an environmental node on a poster. Installing the device in the environment enables recognition of the positions of conversation groups and the sizes of conversational fields, because if human nodes and a group being recorded by the environmental node on the poster are judged by our method to be the same group based on auditory similarity, we can deduce that people are near the poster.

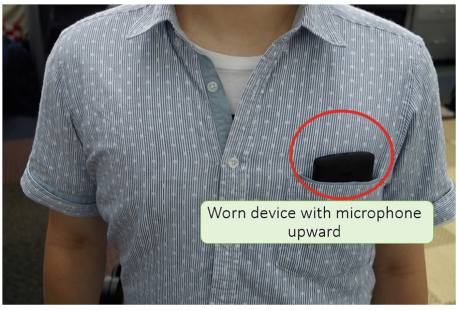

Fig. 3. Worn device (Human node).

Fig. 4. Environmentally installed device (Environmental node).

Detection of Conversational Fields. We employ a detection algorithm that analyzes auditory similarity. Our algorithms are based on those of Neary [21], because Neary has a sufficiently high precision ratio and is lightweight. We optimized the parameters of the algorithm in accordance with the devices used for conversational field detection in this study.

The Neary algorithm obtains the similarity between each pair of devices and judges whether or not there are conversation groups, based on a threshold. The algorithm is as follows:

1. Record audio by microphone-equipped smartphone.
2. Process the sound data using a fast Fourier transform (FFT) every six seconds.

3. Extract frequency characteristics ranging from 50 Hz to 1600 Hz in 1 Hz increments.
4. Compare feature amounts obtained via cosine similarity to that of other devices, using the following formula:

$$Cos(u, v) = \frac{\sum_{i=0}^{1550}(u_i) \times (v_i)}{\sqrt{\sum_{i=0}^{1550}(u_i)^2} \times \sqrt{\sum_{i=0}^{1550}(v_i)^2}} \quad (1)$$

   u and v are device identifiers
5. Every 80 s, count the number of instances in which the threshold of 0.75 is exceeded.
6. If number of instances in which the threshold is exceeded is increasing, judge these devices as belonging to the same conversational fields.

We made a few modifications to the Neary algorithm. In Neary, one second of non-silent content is extracted from a six-second buffer (feature amount extraction step 2); however, its performance suffers owing to the time lag on each device. Therefore, in this study we use the entire six seconds of sound data for calculation. Moreover, the threshold was defined as 0.775 in Neary, but we adjusted it to 0.75 (step 5), because we use different microphone-equipped smartphones.

Next, our system performs the smoothing of sequentially obtained results using the above algorithm. Conversational field information is obtained once every second. If the degree of similarity exceeds the threshold more times than it does not until 36 s before the end of the time period, these devices are considered co-located. In addition, smoothing is performed on conversational fields using two values. However, these parameters were determined based on the data, and this algorithm generates time-lag between judgement results and videos; thus, we synchronized the video to the result.

Scene of Detecting Conversational Fields. We show a scene in which conversational fields are detected and comprehended.

Figure 5 shows movement among three conversation fields. The left side of the figure shows actual video images used as ground truth data. The right side of the figure shows a graphical representation of corresponding scenes based on detection results produced by our system. Human nodes are expressed as human pictograms, and environmental nodes are expressed as rectangles. There were three conversation groups in this scene. In addition, some participants were wearing the devices. Devices were installed on all tables as environmental nodes. The man surrounded by red circles was moving among the conversation groups, and he wore a device identified as 3d6db. His activity is tracked as shown in the node graph.

Figure 6 shows the combining of two conversational fields into one. At the beginning the participants were talking in two different groups, but later they began conversing between the two conversational fields. Therefore, the physical

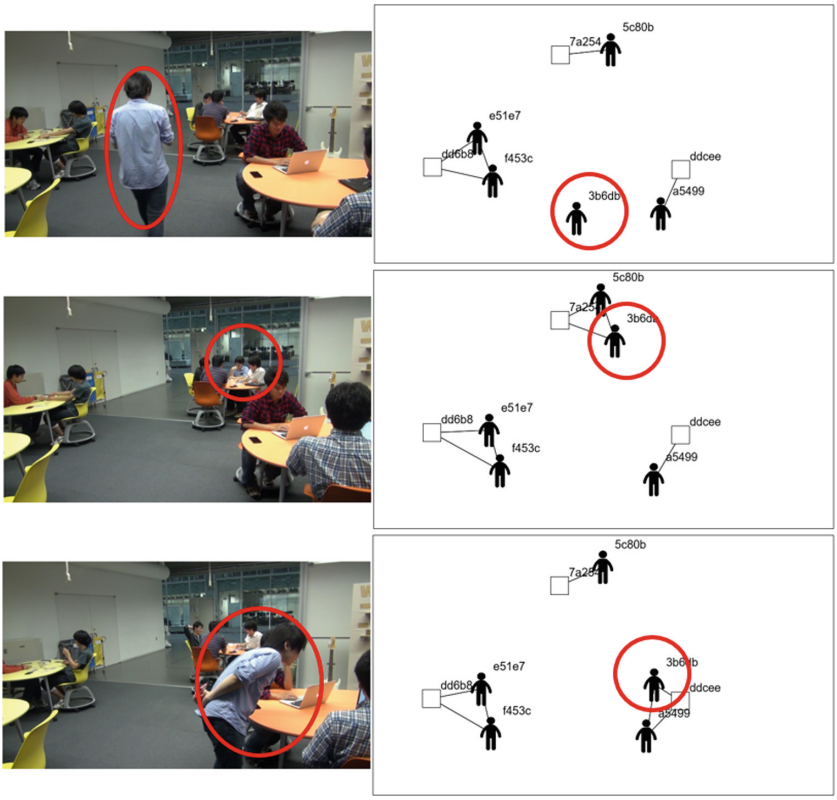

Fig. 5. Tracking a participant who moves across different conversation groups. (Color figure online)

size and dynamics of conversational fields can be detected by our method, as shown in the node graph.

In contrast, Fig. 7 shows a state of separation. The man surrounded by red circles was a moderator, and he was calling participants. However, there were two participants who were talking at presentation booth 1, and ignoring the moderator. In this case, these two participants are close to the moderator, but we believe they should be regarded as another group. However, the classical approach based on physical clusters of people cannot distinguish them. In contrast, our proposed system can distinguish between participants in such situations.

We can recognize approximate physical size and dynamics of conversational fields, and distinguish conversational fields. Using these comprehensions, our proposed system can facilitate video browsing.

3.2 Indexing Shared Experience Videos by Detected Conversation Groups

Figure 8 shows our proposed system that aids the browsing of shared experience videos. The system displays a cue on the seek-bar and visualizes conversational

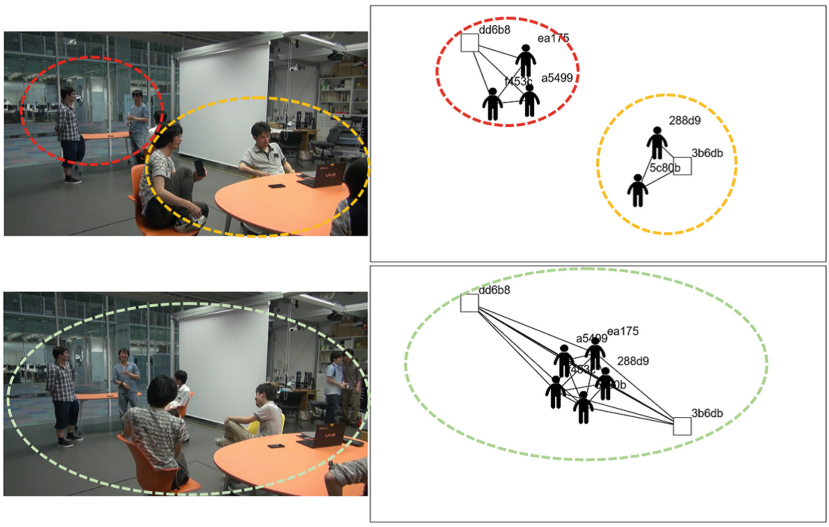

Fig. 6. Two conversational fields merged into one.

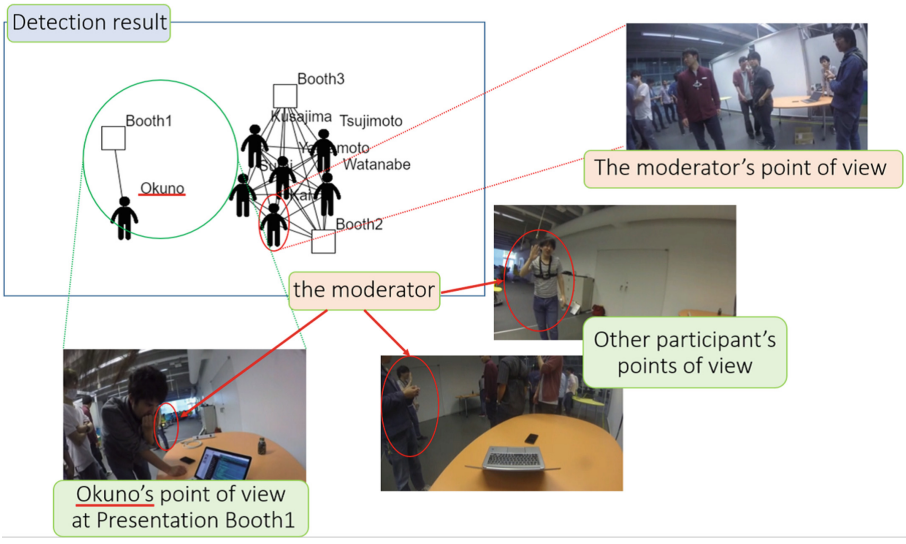

Fig. 7. Two participants are talking while the moderator is speaking. (Color figure online)

fields as node graphs. Users can browse multiple videos using cues based on members, conversation groups, or their positions.

A cue is expressed in n-colors (where n is the number of conversational fields in the data) depending on conversational fields that contain information on conversation group members and their positions. Accordingly, if there are no groups,

the cue is colorless. Moreover, we assign similar colors for similar conversational fields. We explore scenes in multiple first-person videos using cues based on conversational fields.

In addition, if the user searches for a certain conversation group member, the system only shows cues that include the chosen member. Thus, if the member being searched for does not belong to any conversation groups, the cue is expressed as colorless.

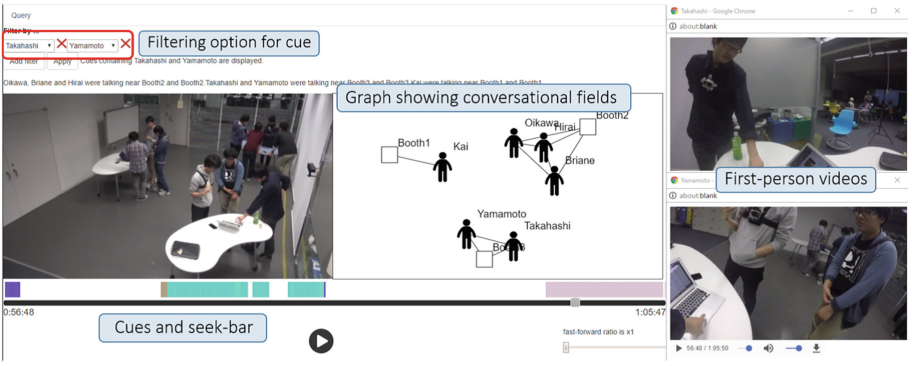

Fig. 8. Proposed system: Showing cues and node graphs based on conversational field detection.

4 Experimental Evaluation

4.1 Experiment

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

We recruited three participants (subjects 1, 2, and 3), and assigned them eight tasks from two datasets (datasets A and B). We observed the subjects to gauge their reaction to our proposed system, in order to confirm whether conversational field information is useful. Finally, we conducted a semi-structured interview.

Datasets Used for Experiment. We prepared two datasets (datasets A and B), and recruited some participants who belong to the same laboratory as the evaluation experiment participants. These data were recorded in a poster presentation session, because we are confident that our proposed method can detect conversational fields in crowded areas and easily detect the dynamics of conversational fields as participants move to listen to a presentation.

Figure 9 depicts poster presentation situations from datasets A and B. These datasets are markedly different in terms of scale. Further details on datasets are listed in Table 1.

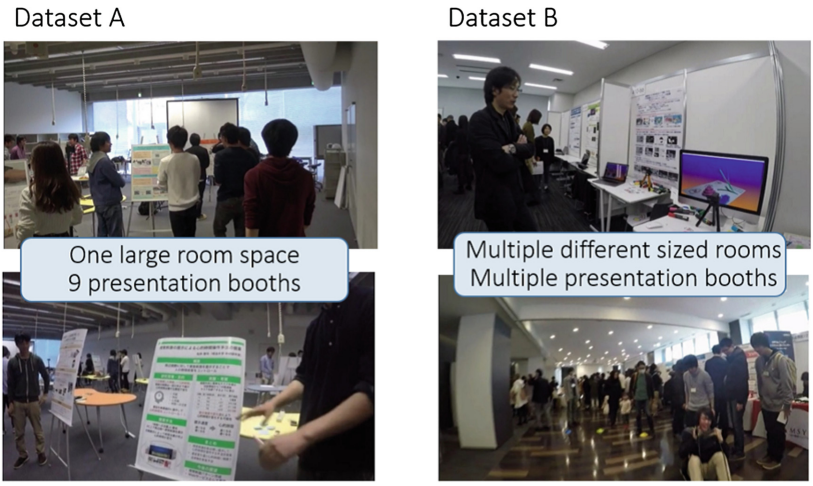

Fig. 9. Circumstances of poster presentation in datasets.

Table 1. Dataset details (video was recorded using cameras worn by participants.

	Dataset A	Dataset B
Participants	about 30	about 600
Presentation booth	9	61
Human node	6	6
Environmental node	3	2
Videos	3	5
Video length	01:12:00	00:57:31
Time elapsed after recording	6 months	2 months

We assigned four tasks for each dataset. The four 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.

Finding Task for Specified Scene. We prepared tasks such as determining "When did participant A converse with B by poster X ?" Figure 10 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 images 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. 10. Example of task for finding specified scenes.

4.2 Results of Time Taken to Complete Tasks

Task completion times for each task and subject is presented in Table 2 and Fig. 11. The cells shaded green in Table 2 indicate the amounts of time taken to complete the task with the help of our proposed system. The figures in red indicate the amount of time the subjects needed to complete tasks that involved watching their own data.

Figure 11 shows a comparison between the proposed system and the baseline software. The blue bar in the graph represents the completion time achieved by

Table 2. Amounts of time needed to complete tasks (in seconds): Red figures represent times measured when subjects watched their own data. Black figures represent times measured when subjects watched the data of others.

	A-1	A-2	A-3	A-4
Subject 1	68	83	18	31
Subject 2	261	128	13	90
Subject 3	58	352	59	71

	B-1	B-2	B-3	B-4
Subject 1	147	42	80	67
Subject 2	43	179	162	216
Subject 3	52	276	30	54

Amounts of the time
needed to complete tasks (in seconds).

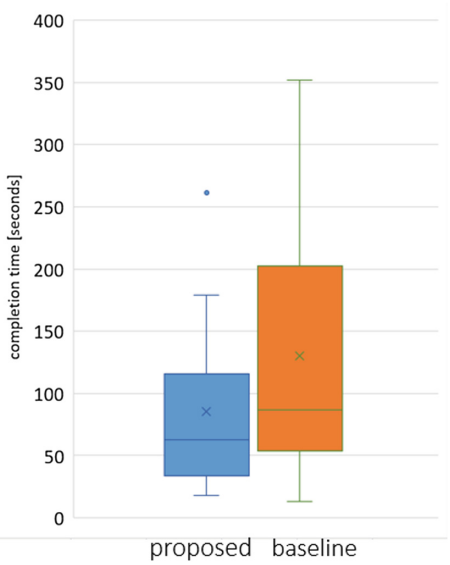

Fig. 11. Task completion times (overall): These graphs show a comparison between our proposed system and the baseline software.

using our proposed system, and the green bar represents the completion time achieved by using the baseline software. Conducting a statistical significance test ($p < 0.05$) revealed that, in the case of browsing video data of other people, 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 left side graphs in Fig. 12). As a result, it was confirmed that our proposed system can aid the watching of videos that include the user's own data.

4.3 Observation and Interview

We observed the subjects to confirm their reactions while using our system. We noted that it appeared difficult for subjects to determine who someone was conversing with, because it is difficult to confirm conversation groups from videos. Moreover, our cue information is nonfigurative, and we defined conversational fields as groups hearing the same voices or environmental sounds. 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 cue, and received the comment that "The cue 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 lose conversation partners," and "I was confused by the expression of the task."

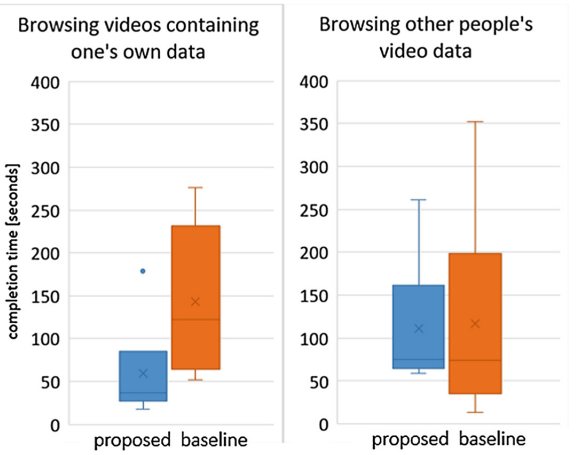

Fig. 12. Task completion times (separated): Graphs on left show a comparison between the proposed system and the baseline software when watching familiar videos; graphs on right represent watching unfamiliar videos.

5 Limitations and Future Work

Our experimental results suggest that our system can help users browse videos that include one's own experience. However, our evaluation and analysis were inadequate. Three participants did not allow adequate assessment of the proposed system. Moreover, the analysis method was insufficient, because we only conducted interviews and measured task completion times in our evaluation. Therefore, our observations and results were not fully supported.

We would like to perform an evaluation with more participants. We will provide data that support observations based on eye-tracking analyses and log data analyses. Moreover, we aim to design a questionnaire that supports the effectiveness of the proposed system.

6 Conclusion

We proposed a system to aid the browsing of shared experience data that includes multiple first-person view videos. With this system, users can avoid the tedious task of searching through lengthy videos. Our system aids browsing by using the video seek-bar to display indices based on conversational field information, including that related to participants and approximate location of group conversations.

We conducted an experiment to evaluate the ability of the indices to decrease the time needed for finding specified scenes in lifelog videos. Our experimental results suggest that our system can aid the browsing of multiple videos that include one's own experiences. On the other hand, the system has not been proven to aid the browsing of unknown data.

References

1. Hall, E.T.: The Hidden Dimension. Doubleday, New York (1966)
2. Borovoy, R., Martin, F., Vemuri, S., Resnick, M., Silverman, B., Hancock, C.: Meme tags and community mirrors: moving from conferences to collaboration. In: Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work (CSCW 1998), pp. 159–168. ACM, New York (1998)
3. Wyatt, D., Bilmes, J., Choudhury, T., Kitts, J.A.: Towards the automated social analysis of situated speech data. In: Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp 2008), pp. 168–171. ACM, New York (2008)
4. Yoshida, H., Ito, S., Kawaguchi, N.: Evaluation of pre-acquisition methods for position estimation system using wireless LAN. In: Proceedings of the Third International Conference on Mobile Computing and Ubiquitous Networking (ICMU 2006), pp. 148–155 (2006)
5. Do, T.-M.-T., Gatica-Perez, D.: Contextual grouping: discovering real-life interaction types from longitudinal bluetooth data. In: IEEE 12th International Conference on Mobile Data Management (MDM 2011), vol. 1, pp. 256–265, June 2011
6. Intille, S.S., Davis, J.W., Bobick, A.F.: Real-time closed-world tracking. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 1997), pp. 697–703, June 1997
7. McKenna, S.J., Jabri, S., Duric, Z., Rosenfeld, A., Wechsler, H.: Tracking groups of people. Comput. Vis. Image Underst. **80**(1), 42–56 (2000)
8. Kendon, A.: Conducting Interaction: Patterns of Behavior in Focused Encounters. Cambridge University Press, Cambridge (1990)
9. Vascon, S., Mequanint, E.Z., Cristani, M., Hung, H., Pelillo, M., Murino, V.: A game-theoretic probabilistic approach for detecting conversational groups. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9007, pp. 658–675. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16814-2_43
10. Vascon, S., Mequanint, E.Z., Cristani, M., Hung, H., Pelillo, M., Murino, V.: Detecting conversational groups in images and sequences: a robust game-theoretic approach. Comput. Vis. Image Underst. **143**, 11–24 (2016). Inference and Learning of Graphical Models: Theory and Applications in Computer Vision and Image Analysis
11. Alameda-Pineda, X., Yan, Y., Ricci, E., Lanz, O., Sebe, N.: Analyzing free-standing conversational groups: a multimodal approach. In: Proceedings of the 23rd ACM International Conference on Multimedia (MM 2015), pp. 5–14. ACM, New York (2015)
12. Vázquez, M., Steinfeld, A., Hudson, S.E.: Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2015), pp. 3010–3017, September 2015
13. Lane, N.D., Georgiev, P., Qendro, L.: DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2015), pp. 283–294. ACM, New York (2015)
14. Kannan, P.G., Venkatagiri, S.P., Chan, M.C., Ananda, A.L., Peh, L.-S.: Low cost crowd counting using audio tones. In: Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems (SenSys 2012), pp. 155–168. ACM, New York (2012)

15. Azizyan, M., Constandache, I., Choudhury, R.R.: SurroundSense: mobile phone localization via ambience fingerprinting. In: Proceedings of the 15th Annual International Conference on Mobile Computing and Networking (MobiCom 2009), pp. 261–272. ACM, New York (2009)
16. Zhang, B., Trott, M.D.: Reference-free audio matching for rendezvous. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010), pp. 3570–3573, March 2010
17. Nirjon, S., Dickerson, R., Stankovic, J., Shen, G., Jiang, X.: sMFCC: exploiting sparseness in speech for fast acoustic feature extraction on mobile devices - a feasibility study. In: Proceedings of the 14th Workshop on Mobile Computing Systems and Applications (HotMobile 2013), pp. 8:1–8:6. ACM, New York (2013)
18. Tan, W.-T., Baker, M., Lee, B., Samadani, R.: The sound of silence. In: Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems (SenSys 2013), pp. 19:1–19:14. ACM, New York (2013)
19. Aoki, P.M., Romaine, M., Szymanski, M.H., Thornton, J.D., Wilson, D., Woodruff, A.: The Mad Hatter's cocktail party: a social mobile audio space supporting multiple simultaneous conversations. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2003), pp. 425–432. ACM, New York (2003)
20. Wirz, M., Roggen, D., Tröster, G.: A wearable, ambient sound-based approach for infrastructureless fuzzy proximity estimation. In: International Symposium on Wearable Computers (ISWC 2010), pp. 1–4, October 2010
21. Nakakura, T., Sumi, Y., Nishida, T.: Neary: conversational field detection based on situated sound similarity. IEICE Trans. Inf. Syst. **E94–D**(6), 1164–1172 (2011)
22. Kopf, J., Cohen, M.F., Szeliski, R.: First-person hyper-lapse videos. ACM Trans. Graph. **33**(4), 78:1–78:10 (2014)
23. Poleg, Y., Halperin, T., Arora, C., Peleg, S.: EgoSampling: fast-forward and stereo for egocentric videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), pp. 4768–4776 (2015)
24. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012), pp. 1346–1353, June 2012
25. Arev, I., Park, H.S., Sheikh, Y., Hodgins, J., Shamir, A.: Automatic editing of footage from multiple social cameras. ACM Trans. Graph. **33**(4), 81:1–81:11 (2014)
26. Lee, Y.J., Grauman, K.: Predicting important objects for egocentric video summarization. Int. J. Comput. Vis. **114**(1), 38–55 (2015)
27. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013), pp. 2714–2721, June 2013
28. Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: Proceedings of the 2011 International Conference on Computer Vision (ICCV 2011), pp. 407–414. IEEE Computer Society, Washington, DC (2011)
29. Li, C., Kitani, K.M.: Pixel-level hand detection in ego-centric videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013), pp. 3570–3577, June 2013
30. Cai, M., Kitani, K.M., Sato, Y.: A scalable approach for understanding the visual structures of hand grasps. In: IEEE International Conference on Robotics and Automation (ICRA 2015), pp. 1360–1366, May 2015
31. Yonetani, R., Kitani, K.M., Sato, Y.: Ego-surfing first person videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), pp. 5445–5454, June 2015

32. Poleg, Y., Ephrat, A., Peleg, S., Arora, C.: Compact CNN for indexing egocentric videos. In: IEEE Winter Conference on Applications of Computer Vision (WACV 2016), pp. 1–9 (2016)
33. Higuchi, K., Yonetani, R., Sato, Y.: EgoScanning: quickly scanning first-person videos with egocentric elastic timelines. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI 2017), pp. 6536–6546. ACM, New York (2017)