Master Thesis

Assisting Collaborative Activity Analysis through Hand Cue in Multiple Egocentric Videos

(複数一人称視点映像中の手を手掛かりとした
協調的行動分析支援)

Author Nathawan Charoenkulvanich
Advisor Prof. Yoichi Sato

Department of Information and Communication Engineering
Graduate School of Information Science and Technology
The University of Tokyo

July 2018

Abstract

Behavioral psychology or Computer Supported Cooperative Work (CSCW) research has aimed to analyze and understand key behavioral elements of collaboration. In virtue of the tedious manually video coding procedure, the researchers have approached for automatically detecting key behavioral element with the only available supporting technology such as speech recognition, eye-gaze, and facial expression detection. Motivated by a recent development in head-mounted cameras and hand detection technology through egocentric video, we decide to investigate the potential of hand cue as a new key element for assisting the collaborative behavioral research.

In this thesis, we design an interface system and an evaluation process for verifying the effectiveness and usefulness of the hand cues in the collaborative behavioral research. We interview the expert researchers in this field in order to figure out the difficulty during conducting their researches. Afterward, we analyze the past works and found that hand's appearance in the egocentric video and its identity have a potential for identifying activity state of each member in a group activity as following: individual working, collaborative working, observer, and passive worker. Then, we conduct the experiment to evaluate the potential of identifying the activity state by the following hand cues: no hand, only owner's hand, both owner and other's hand, and only other's hand detected.

Lastly, we analyze the collected data and discuss the effectiveness and usefulness of each hand cue as a potential indicator of an individual's activity state in collaborative work.

Acknowledgements

I would like to express my sincere appreciation to all those who provided me with the opportunity, support, and encouragement to complete this thesis. This thesis would not be able to complete without all the support and help from many individuals who provided expertise which greatly assisted the research.

First, I would like to express my deepest gratitude to my supervisor, Professor Yoichi Sato. He provided me such a warmed welcome and wonderful environment for conducting the researches. His kind, accurate and valuable advice greatly assisted my research. Under his guidance, I doubtlessly enhance my research skill and make this thesis possible. Next, I would like to give my special thank to my friend, Rie Kamikubo, who gave me lots of helpful insight in HCI aspects, introduced me to many expert researchers in this field, and continuously encourage me. Next, I would like to thank Dr. Ryo Yonetani, Dr. Keita Higuchi, and Dr. Minjie Cai who constantly supervised and assisted me in this research. There is a time when I need to do the data collection and experiments which I received a lot of help from every member of the laboratory. Therefore, I would like to express my gratitude to them.

Furthermore, I would like to thank my friend, Nontawat Charoenphakdee, who gave me lots of advice and brought me to various interesting spots for cheering me up. Moreover, I would like to thank my friend, Mungmuang Promsen who accompanied me for data collection until a very late time. I also would like to say thank to Dr. Wiennart Mongkulmann who gave a vast amount of suggestion which covered all important thing from simple tips

while living in Japan until the advice for my research. I also would like to thank Jamorn Sriwasansak and Assistant Professor Vorapong Suppakitpaisarn for giving me a lot of kind and funny comments.

I express my gratitude to the Ministry of Education, Culture, Sports, Science and Technology for offering me the MEXT scholarship. They gave me this precious opportunity for studying at the University of Tokyo.

Finally, I would like to express my deepest thank to my family members: my parents, my sister and my brother for the strong belief in myself, and constantly supported both my health and my mind.

July 11th, 2018

Nathawan Charoenkulvanich

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

1.1.1 Collaborative work analysis

Collaborative work has various factors that affect its process and outcome of the group task. Engagement levels of individuals in a group, equality in contributions of each member, leadership and follower relationships, or social bonding aspects are all examples that influence the quality of collaborative work. Research in behavioral psychology or Computer Supported Cooperative Work (CSCW) has thus aimed to analyze and understand key behavioral elements of collaboration. For example, Suzuki *et al.* [1, 2, 3] evaluated human factors that influence the outcome of group activities by analyzing joint attention or turn-taking conversations. Moreover, the work of Cukurova *et al.* [4, 5] analyzed head directions and hand positions of students in a group to investigate their collaborative problem-solving skills. Such analysis requires researchers to observe various behavioral elements and activities found in a group.

The fundamental observation factor can be divided into two groups: verbal and nonverbal elements. While analyzing the recorded verbal communication information via audio recording is a popular method to extract the explicit intention of each member in a group, it is difficult to extract the implicit

intention. On the other hand, nonverbal factors are known to be able to infer both explicit and hidden intentions as well as personal habits, and in addition of the current development of the technologies which allow the researcher to capture the nonverbal factors more accurately. As a result, nonverbal factors have received more attention recently from the research community. Currently, the common nonverbal elements used for analysis are gaze direction, face direction, facial expression, hand gesture, and body motion.

To elaborate the analysis components of non-verbal elements, hand's state has been approached in observing collaborative work. For instance, Cukurova *et al.* [4, 5] observed students' behavior through their hands' active status and face position from several side view cameras. Their objective was to investigate the difference of each individual student's collaborative problem-solving skill. They defined three statuses of each student as follows: the active state indicates that a student's hands are working with objects related to the task, the semi-active state indicates that a student is facing towards another student who is in an active state or object that is subjected to the task, and the passive state otherwise.

1.1.2 Method of video analysis in a collaborative behavioral study

In most cases, researchers observe and analyze the behaviors and interactions of multiple people simultaneously using videos captured from collaborative task experiments. The most popular method to analyze video is using a software called ELAN [6] as shown in Figure 1.1. It offers multiple videos syncing and browsing with an annotation area for annotating videos in multiple layers. Another software called Wavesurfer [7] is used by Suzukie*t al.* [3] as shown in Figure 1.2 for additional gaze transcription and with the different layout of data visualization.

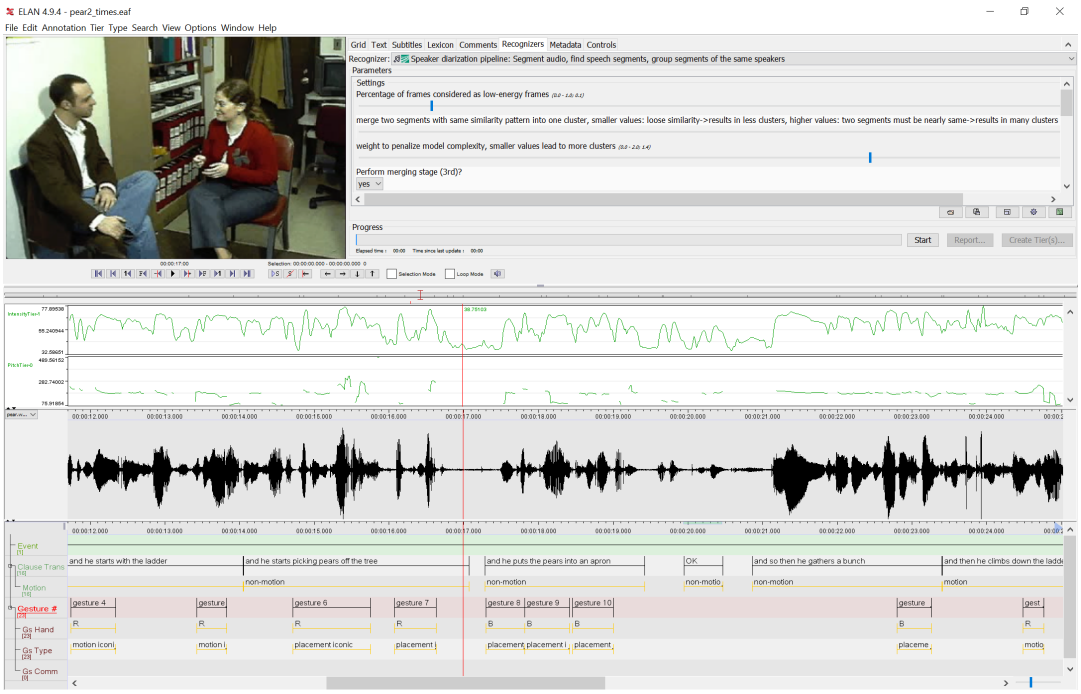

FIGURE 1.1: ELAN interface [8]

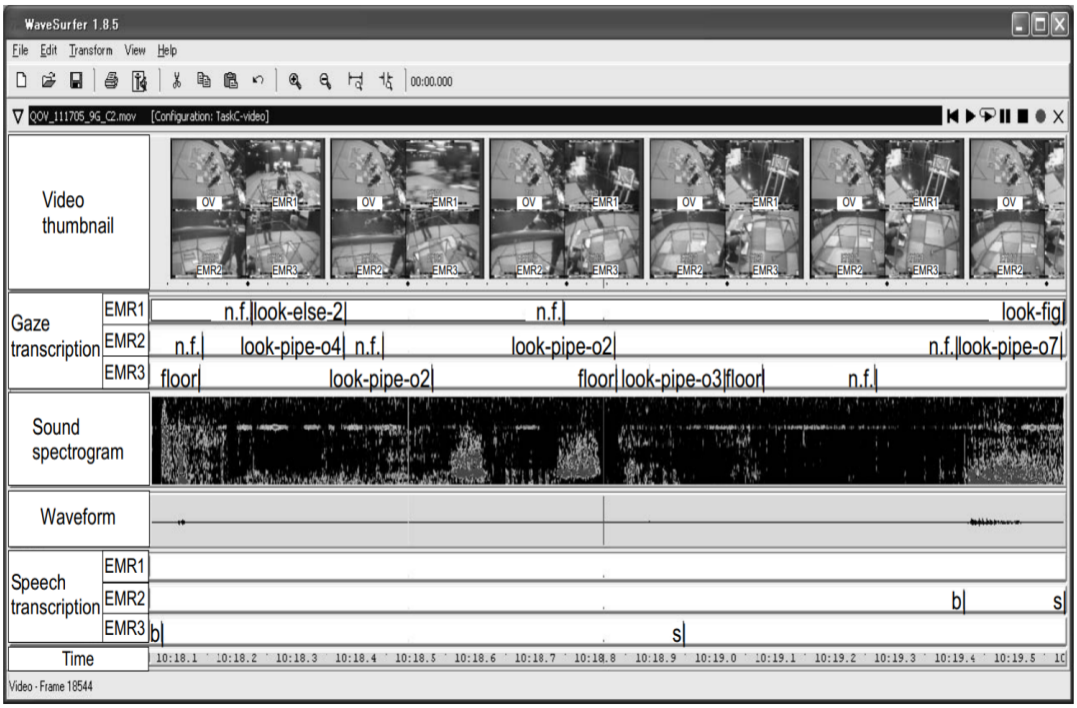

FIGURE 1.2: Wavesurfer interface used within the work of Suzuki *et al.*[3]

Even though there is a notable amount of analysis using video recordings, this demanding task of observing and analyzing the behavior of multiple people simultaneously significantly adds effort and complexity in the research. Researchers often observe certain collaborative scenes such as object exchanges in multiple-people interactions by looking at the pattern of eye gaze, body and hand position of people one by one, and then find the related participant of that event from all participants [9, 3]. The annotation task has to be done manually which consumes long work hours since the annotation results are needed to be precise. Not only that, sometimes it is unavoidable to deal with the human error or disagreement among the researchers. For this reason, the automatic key factors extraction has been approached to mitigate this problem effectively.

1.1.3 Automatic key behavior extraction

To support the analysis of collaborative behaviors, automatic detection of key behaviors has been approached to lessen the effort of analysis. Recent collaborative behavioral researches used different automatically detected behavioral cues such as speech [10, 11, 12], turn-takings, visual attention [13, 14], face detection [15], and body gestures [16].

To our knowledge, there is no direct approach to hand activities as a factor to evaluate collaborative activity due to its complexity. On the other hand, there is an increasing amount of research to classify hand actions, especially through first-person perspective videos. For instance, Chen *et al.*[17] classified hand interaction pattern for each part of the sewing machine in order to automatically create a manual of machine operation for a new user. They divide the hand shape and motion into four types of action as follows: push, put, rotate and slide. This classification is a signal that we can utilize hand data into a more fine-grained action of a human, and considering hand as a new proxy in the analysis of collaborative behaviors is an interesting goal.

1.2 Major contributions of this work

We propose to explore a computer-vision assisted detection to expand the scope of analysis techniques of collaborative behaviors through hand-detection within multiple-videos browsing interface. Importantly, we investigate the effect of hand cues provided in analyzing collaborative behaviors from first-person videos. State-of-the-art computer vision techniques allow us to extract information that the researchers want to observe automatically from videos and shorten the time usage in researches.

This research has to overcome several challenges. The main difficulty comes from the ambiguous way to use the extract hand data as a part of factors helping in collaborative behavior study which we need to figure out since there have not been obviously declared how to utilize this data in the collaborative behavior research field before. After that, we need to design the interface and data visualization to utilize this extracted data. At the same time, the tasks used in the experiment are also considered in the way that we can investigate the benefit of hand information to the study of collaborative behaviors.

After we analyze how the researchers in the collaborative study benefit from hand cue, we then find that hand activity of each worker in a group are used to determine their working status, so we propose an interface which designed for evaluating the effectiveness and usefulness of hand-detection within the first-person view from multiple videos. Then, we divide hand-detection categories based on the idea that we want to inspect the hand working status of each worker such as a private working, a collaborative working, observing and passive status through the first-person perspective. Furthermore, with the advantage of the first-person view that the owner of perspective's hand tends to show up inside their visual field during the interaction involving their hand[18, 19], we can observe participants' attention

and action with less occlusion. As a result, we decide to divide the hand data categories into no hand detected, only the owner's hand detected, owner's hand together with others' hand detected and only other's hand detected within the egocentric view. In short, our contributions are as follows:

- We propose a multiple first-person video browsing interface visualized with four hand-related cues in order to evaluate the potential of our proposed hand information.
- We evaluate the potential of incorporating our proposed hand information into the existing framework using the first-person video. It can be observed that the hand information is potentially useful to represent human activity during collaborative work.

1.3 Thesis outline

The rest of this thesis is organized as follows, Chapter 2 introduces past researches which focus on collaborative work analysis, existing automatic nonverbal cues extraction for assisting collaborative work analysis, and elaborates on the potential of hands in assisting events analysis. Then, Chapter 3 summarizes the preliminary feedback from the experts about procedures and difficulty. Next, Chapter 4 explains our proposed interface. Chapter 5 reports data collection's procedure. Chapter 6 explains the protocol of evaluation and shows the analyzed result from usability testing. Finally, Chapter 7 concludes this thesis and discuss ideas for future research direction.

Chapter 2

Related work

In this chapter, we introduce two topics that are related to our proposed idea. The first topic is the list of past works which worked on automatically extracting the nonverbal behavior for assisting video analysis. Then, we introduce existing work that investigates hand action in their researches about human collaboration and analyze how they use hand action for their researches.

2.1 Automatic extraction of nonverbal information for human behavior analysis

There have been multiple works to facilitate the observation of collaborative scenarios through automatic detection of key human behaviors and activities.

Wang *et al.* [16], they offer an automatic segmentation of body gesture in order to assist the video labeling process by generating predicted annotation. Users of the system are able to edit the annotations which might be needed in the case when the generated result from the system is not desirable.

Oertel *et al.* [20] introduced a model of individual engagement and group involvement from eye-gaze detection with different aspects of gaze patterns.

Hung *et al.* [21] used audio cues for estimating turn-taking, and video cues from close view cameras for upper-half body motion detection then combines the data from both sources into audio-visual cues for analyzing group's cohesion through group behavior.

Furthermore, Zancanaro *et al.* [22] used time interval of speech event to automatically detect group functional role in face to face situation. They proposed the coding scheme for classifying the roles of the group members based on simple features of the visual and acoustical scene.

To the best of our knowledge, no attempt has been made to use automatically hand activities detected as a proxy to assess collaborative behaviors before due to the deficient of accuracy and supported framework. Nowadays, the state-of-the-art made it possible to implement a satisfactory accurate hand detection method using neural networks. Therefore, we propose to incorporate the hand cue to expand the analysis features of collaborative behavioral research.

2.2 Analysis of events through hand actions

While hand activities have been discussed to be used in the applications to support new analysis techniques, we first need to understand how researchers are currently analyzing collaborative behaviors and extract what they are currently looking to measure in the new collaborative analysis platform.

Cukurova *et al.* [5] created a model for analyzing collaborative problem-solving (CPS) in practice-based learning activities by focusing on interaction in a physical environment. They identify the differences in groups collaborative problem-solving behaviors by classifying each student status through

human observation, hand position and heads direction. They divide observed behavior into an active, semi-active and passive state. The participants who use their hand working on the interactive toy is regarded as an active state. In the same time, if there is another participant observe active state person, that person is annotated as a semi-active state. Then, the rest who is not engaged in the activity is marked as a passive state.

The authors found that the high CPS competent groups have more equality in the distribution of time in their problem-solving stages than the low CPS competent groups. Whereas, the other group spent more time in identifying knowledge and skill deficiencies. They found that their coding scheme which based on students physical activity data provides useful data pattern to identify group differences. However, their current work does the annotation process manually by the expert researchers which they mention that their result shows the potential of using a machine for automatically detect behavioral factors after their coding scheme in order to help the study of collaborative research. For that reason, we can confirm that our idea of using the first-person video for finding hand existence has a potential in the collaboration activities analysis. Moreover, we design our evaluation process based on their study in the part of the searched event such as individual working, observing and passive status.

Suzuki *et al.* [1] observed several factors divide into two categories: psychological indexes and physical performance evaluation, and used them to analyze the effects of group size in the furniture assembly task. The psychological indexes are divided into three behavioral indexes: degrees of contribution, satisfaction, and familiarity. In the same way, the physical performance evaluation also divided into three factors: degree of completion which is a ratio of successful and unsuccessful groups among three different group sizes, time-to-completion, and duration of interaction with materials. With the collected data, they analyzed and found that social loafing effects have

varied directly after the increasing number of participants, and the member in a small group felt more satisfaction on self-contribution more than a large group which has less time on individual work.

One of the data which interested us the most is the duration of interaction with materials. Within their research, they need to observe the time when each member in the group come to interact with the material through side view video as shown in Figure 2.1. Their result shows that the interval time of hand active state is an interested candidate for analyzing collaboration work. Consequently, we design one of our evaluation tasks based on finding the interval time of hand active state per each worker in a group.

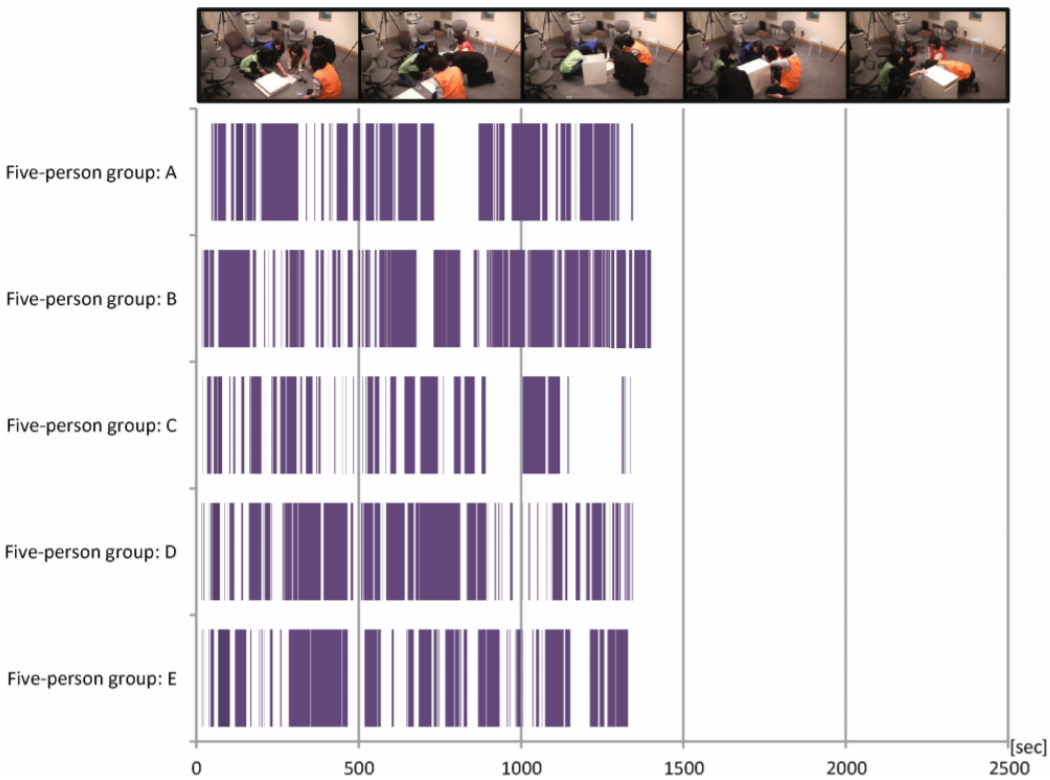

FIGURE 2.1: Distribution of duration of interaction with the material. Figure are taken from Suzuki *et al.*[1]

Chapter 3

Preliminary expert feedback

Looking into more details of the problem when researchers are working on the human behavior in collaborative work research, we interviewed two expert researchers in the collaborative behaviors study. One of them is academic researcher who experienced in collaborative work analysis and the other one is a research and development researcher in group dynamic analysis. Then, we divide the process into two phases: capturing phase and analyzing phase. We utilize all summarized data as a guideline to design the system interface.

3.1 Capturing phase

Capturing data is the phase that induces the complexity level of an analysis phase due to there are multiple types of recorded data. In case of the interviewee who is the expert, the data which is needed to capture altogether were speech information from wireless microphones, overview video from multiple side-view cameras, eye gaze from attached marker appearing in the footage recorded by a wired video recorder and motion coordination from motion capture technology.

3.1.1 Difficulty in capturing phase

Through all the data that the researchers needed to capture in order to observe and analyze all theorized related factors, there are difficulties as they collecting data as the following:

- Synchronizing all videos and data recording for analyzing part are being done by using the idea of film clapperboard. However, the synchronization is still not correct and is needed to be edited after getting all of the data. Furthermore, there are overall five videos and three data of body coordination which all of them are collected by different frequency depending on which the device they use. As a result, this can be considered as a time-consuming step for the research.

- The most ideal situation for collecting data of participants' position for observing their collaboration with others is to set the video recorder on the top-down view. However, there is a difficulty in setting up the device, so they have to use side view video recorders from two angles instead. Unfortunately, this increases the number of videos that are needed to process and analyze.

3.2 Analyzing phase

After capturing all data as explained in the previous section, the next stages are to process all recorded data which we called data coding and then observe the processed data in order to extract the information for the studies.

3.2.1 Difficulty in analyzing phase

There are a number of difficulties in analyzing phase since there are multiple types of data and also multiple instances for each type that are needed

to be observed in a parallel manner. The difficulty while analyzing data can be listed as the following:

- Multiple videos annotating

  Normally, annotating a 10-minute video could take more than one work hour for a researcher, and it takes even more time when they need to annotate multiple videos per one timeline. This is unavoidable in order to study the collaborative behaviors when the multiple videos are given.

- Discovering all related videos for each event

  Long working hours on multiple videos are not the only concern while annotating. The researcher also stated that because the videos are related to each other, when they found one interesting event especially on collaborative events such as mutual gaze or joint attention, they need to as well go through all other videos in the list to find which video is the one that related to that event. This task is one of the subtasks within video annotation that make workers confuse and exhausted.

- Worker's identity identification

  Worker's identity is also an important information that the researcher needs to use to understand each person behavior. However, when each worker is working, they did not look directly at each other all the time. This implies that we cannot always rely on face detection to identify the behaviors. In their case, they let their workers wore clothes with a different color and recognize it manually later when they annotate the videos.

- Occurring occlusion from side-view video

  The conventional way to observe overview video is to observe from multiple side-view videos, however, there are still several times when

they want to specify who is working and what are they working on but there is an occlusion occurred which is caused by the position of workers in a group. They might be able to see a part of the action or in the worst case, cannot see what is happening in the video at all. When this happens, they can do nothing but to guess all the answers. This is also one of the reasons which they think the top-down view video should be able to assist them in this task.

Chapter 4

Designed system

From chapter 2, we found that individual working state is an important factor that researchers use for determining the amount of engagement to their assigned task. The researchers use this quantity of individual work engagement as one part from all aspects that they need for analyzing the performance of group work. Therefore, we designed an interface based on the preliminary expert feedback in Chapter 3 and the hypothesis that the existence of the owner hand in an egocentric video is a potential cue for identifying working state of people.

Our hypotheses are as follows:

- If the owner's hand appears in their view, the attention of owner should be at his/her own hands and have a high possibility of an active state.
- We can divide their working state into individual or collaborative working by the co-existence between owner view's hand and other members' hands inside the video.
- The disappearance of the owner's hand in their perspective can use as a signal that their attention is at other place exclude themselves. Therefore, we can divide it into simple two categories. First, the owner's perspective is observing another member working on their own which

mean we should probably see other's hand inside the view of the observer. The following is the case of no hand detected which should signal as a passive state such as look around, walking, or resting.

From our hypotheses, we propose an interface that will be used to evaluate the effectiveness and usefulness of hand detection cues in assisting collaboration activity research.

4.1 Design requirement

According to the summarization of preliminary expert feedback, we choose the design requirement by focusing on the difficulty in analyzing the phase for the proposed system. The design requirement can be summarized according to this list:

- Multiple videos labeling for data coding

  We design our interface to satisfy this part by visualizing the detected information on a timeline with different color and have a more clear-cut on the enlarge timeline for guiding the annotation place.

- Discovering all related videos for each event

  With the visualization on the parallel timeline, we can see the relationship between each member. For example, member 1 is working with worker 2 while worker 3 is working on his own, in this case, we should be able to differentiate that the color data visualized on timeline 1 and timeline 2 is similar while timeline 3 have a different visualization. This visualization on the parallel timeline can decrease the workload from searching through all of the videos.

- Worker's identity identification

According to this requirement, we focus on giving the identity of owner's perspective first by highlighting the owner's hand area with light blue color. After that, combining this visualization and data visualization on the parallel timeline, then we should be able to identify who is related to each event.

- Occurring occlusion from side-view video

  The occlusion of personal work action from side-view can be decreased by observing all events through egocentric view as an additional information. Therefore, we designed this interface by focusing on extracting data from multiple first-person perspective videos.

4.2 User interface overview

Our proposed interface is designed according to the design requirement from preliminary expert feedback in Chapter 3. The interface has 4 main parts which are videos, timeline for data visualization, button for activating hand cues, and video player controller button as shown in Figure 4.1

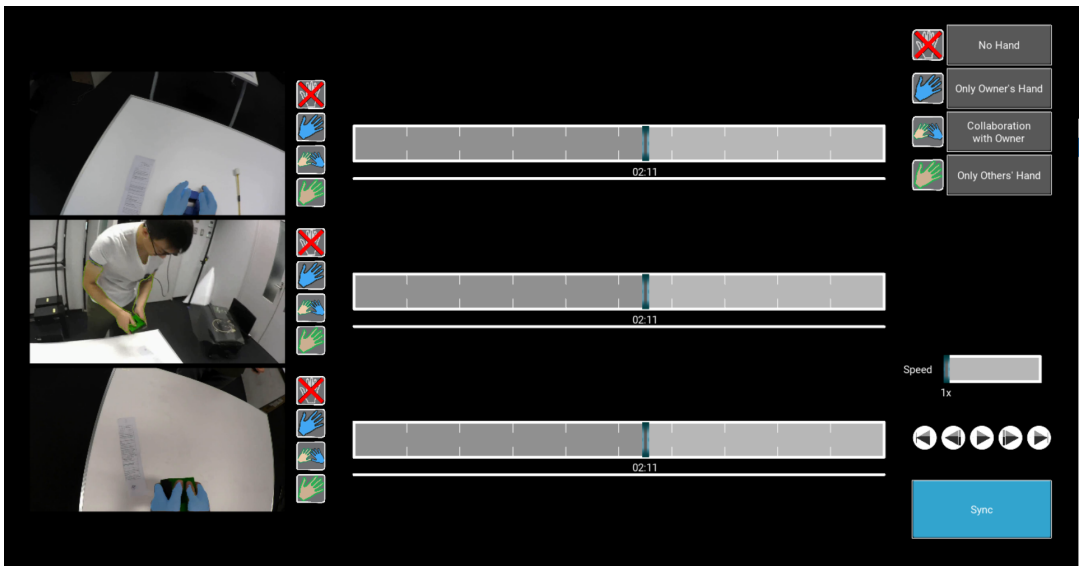

FIGURE 4.1: Interface overview for testing

4.3 Hand detection and identity visualization

4.3.1 Hand detection procedure

In order to verify the hypothesis, the first thing we need is the segmentation of the hand region and then following with owner's identity distinction. The hand region detection is based on the procedure in Cai *et al.* [23] to deal with the rapid change in an environment within egocentric videos. The model hand detector composed of hand pixel classifiers is trained and fine-tune with the 20 frames of hand mask for each change of environment within the dataset. After we get a video of probability map generated from the hand model detector, the next step is to select the hand region from the generated probability map by filtering out face region using OpenFace library[24] and region which is less than 1 percent of video area. Then, we define the owner's hand contour by checking if the contour comes from the bottom line of video, and in case of more than 2 contours come from the bottom line, we select the centermost since the owner of perspective's hand tends to be in the center of the frame. The example of successful detection are shown in Figure 4.2, 4.3, 4.4, and 4.5.

4.3.2 Hand visualization

Hand identity visualization is divided into two types: the owner's hand and other's hand. We decided to highlight owner's hand area with light blue color in order to not become a distraction to the viewer by its color, but still easy to recognize with the whole area highlighted as shown in Figure 4.3. Alternatively, the other's hand is not highlighted within the whole area and the visualization is only the contour of the detected area because we want to prioritized owner's hand first. However, to be able to detect apart from owner hand and still able to easily percept the location of other's hand, we

choose light green contour as a represent for the detected area of other's hand as shown in Figure 4.5, and when both owner's hand and other's hand are together in the scene as shown in Figure 4.4, the owner's hand will be more noticeable than the other's hand which showing up at the same time.

4.4 Data visualization buttons

From the hypothesis, we provide four data categories with four different colors on the timeline. When the user activates each button or icon, these buttons and icons will change its background color to their represent color in order to remind each color visualization meaning. With the current setting, every same function buttons and icons are set to have a synchronized state. When one button from the group is activated, the others in the group will be activated at the same time.

Furthermore, these features are allowed to activate in parallel for using multiple data for finding a more complex scene and observe the flow of events in the videos.

4.4.1 No hand detected

No hand detected is represented with black color due to the meaning of nothing in the scene is detected. when the user enables this button, the black lines will show up on a timeline at the frame that the system cannot detect hand as shown in Figure 4.7. While the button and icons will change as shown in Figure 4.6. This button has the potential to use when the user wants to observe the passive status of each worker. Moreover, in the case of intense collaborative work, this button can use to observe when all group member is resting, so we can also understand the flow of their work.

FIGURE 4.2: Example of no hand detected scenes

4.4.2 Only owner's hand detected

Light blue color is chosen for representing that the system can detect only the owner's perspective hand inside the frame in sync with the color visualization of the owner's hand on the video. The hand color represents in icon also set the whole area to blue tone to map with a presentation of the owner hand in the video. When this button or icons are enabled, the light blue line will appear on the timeline to represent the frame which only the owner's hand is existing as shown in Figure 4.8. At the same time, the button and icons will change as shown in Figure 4.6. This data can pose as a hint for the time when each worker is individually working in their private space.

4.4.3 Owner's hand and other's hand detected

When the owner of perspective's hand and other's hand are detected together in the frame, the pinkish red will be visualized on the timeline when the user activates this function as shown in Figure 4.9. While the button and icon will change as shown in Figure 4.6. This feature is created for testing the potential of collaboration action detection with only hand detected.

4.4.4 Only other's hand detected

This feature when activated, the yellow line will show up on all timeline to depict the frame which can detect other's hand but cannot detect the owner view's hand. The represented icon is designed to have a light green contour to represent that other's hand in a video is visualized with light green contour over the detected area. The interface when this function activated is shown as in Figure 4.10, and the related button and icons is depicts as in Figure 4.6

4.5 Multiple timeline

Our interface show one timeline per one video to emphasized the visualization data according to each video. Moreover, we have restricted the position of all timelines to have the same start position on the x-axis to be able to use the advantage of data visualization as shown in Figure 4.11 to figures out who is working with whom and able to decrease confusion while searching for the related video.

4.6 Video control functions

The user is allowed to control video as follows:

- Play and pause

  Users can play and pause video through two ways: click at play/pause button or toggle by spacebar key on the keyboard. This play/pause function control all the video in the list. When the users enable play or pause function, the videos will play or pause all at once.

- Video playback speed control

  While playing the video, users can increase or decrease the speed of video playback by changing the speed at speed bar in the right panel control. The maximum speed is set to 20 times of original playback speed.

- Move to a specific time on the timeline

  Ability to go to the desired location fast and precise are necessary for easing user when they work on numerous amount of data. Therefore, we design to let the user go moving through timeline in 3 different size steps as following: large movement can be done by dragging cursor through timeline, one second step forward or backward

by scrolling middle wheel of mouse, and the smallest step moving 1 frame per step by using left/right arrow on keyboard or click on the icon next/previous frame as shown in Figure 4.12.

- Synchronized and Unsynchronized video The synchronized button is located at the bottom right corner of the interface. The default of this button is set to the synchronized state because we prioritized the synchronization between the videos. Nevertheless, we still allow the user to deactivate the synchronized feature through this button when they want to focus on only one video or in case they want to compare scene with a different time in a different video.

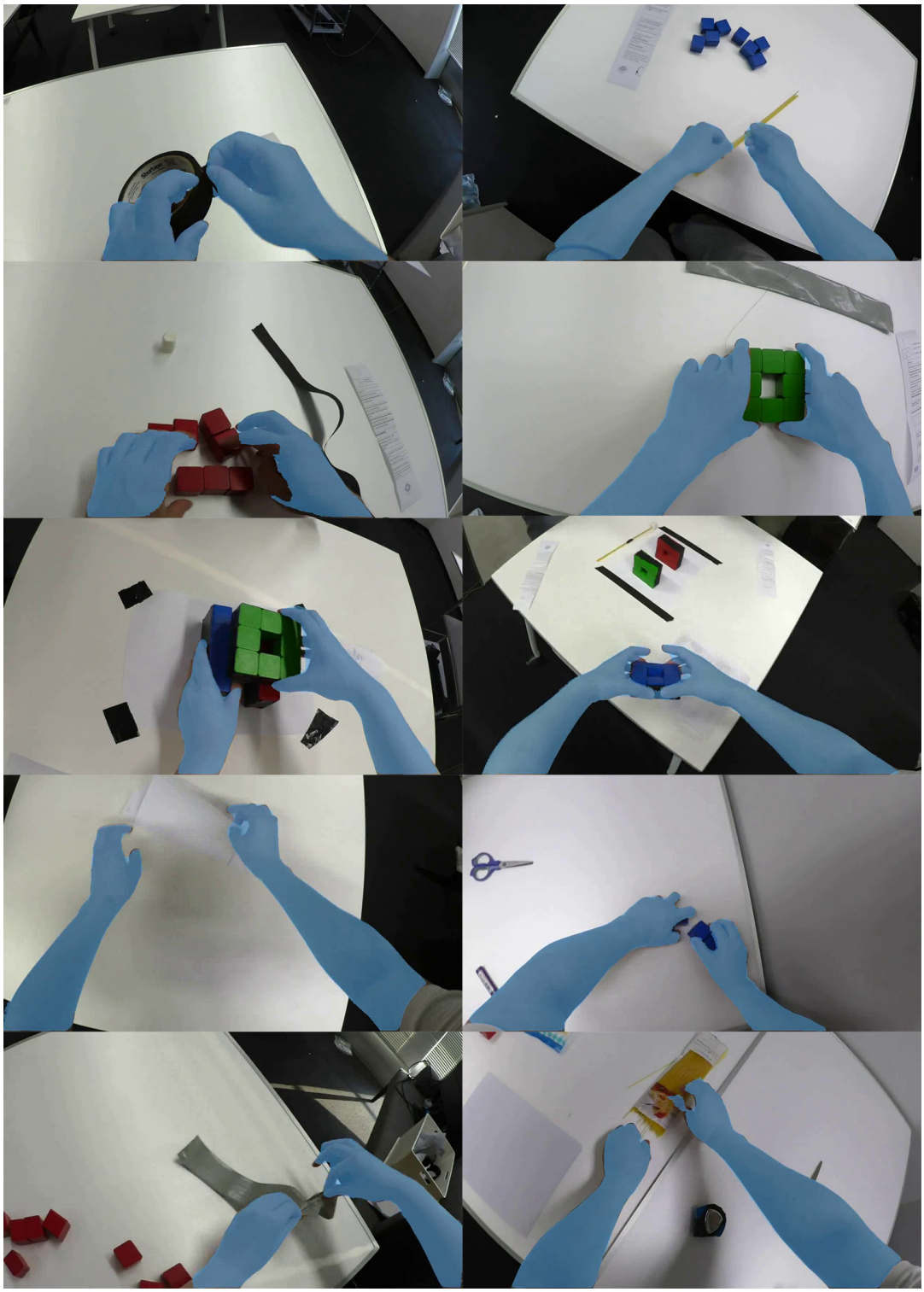

FIGURE 4.3: Example of the only owner of the point of view's hand detected scenes. The detected owner's hand is highlighted in light blue color

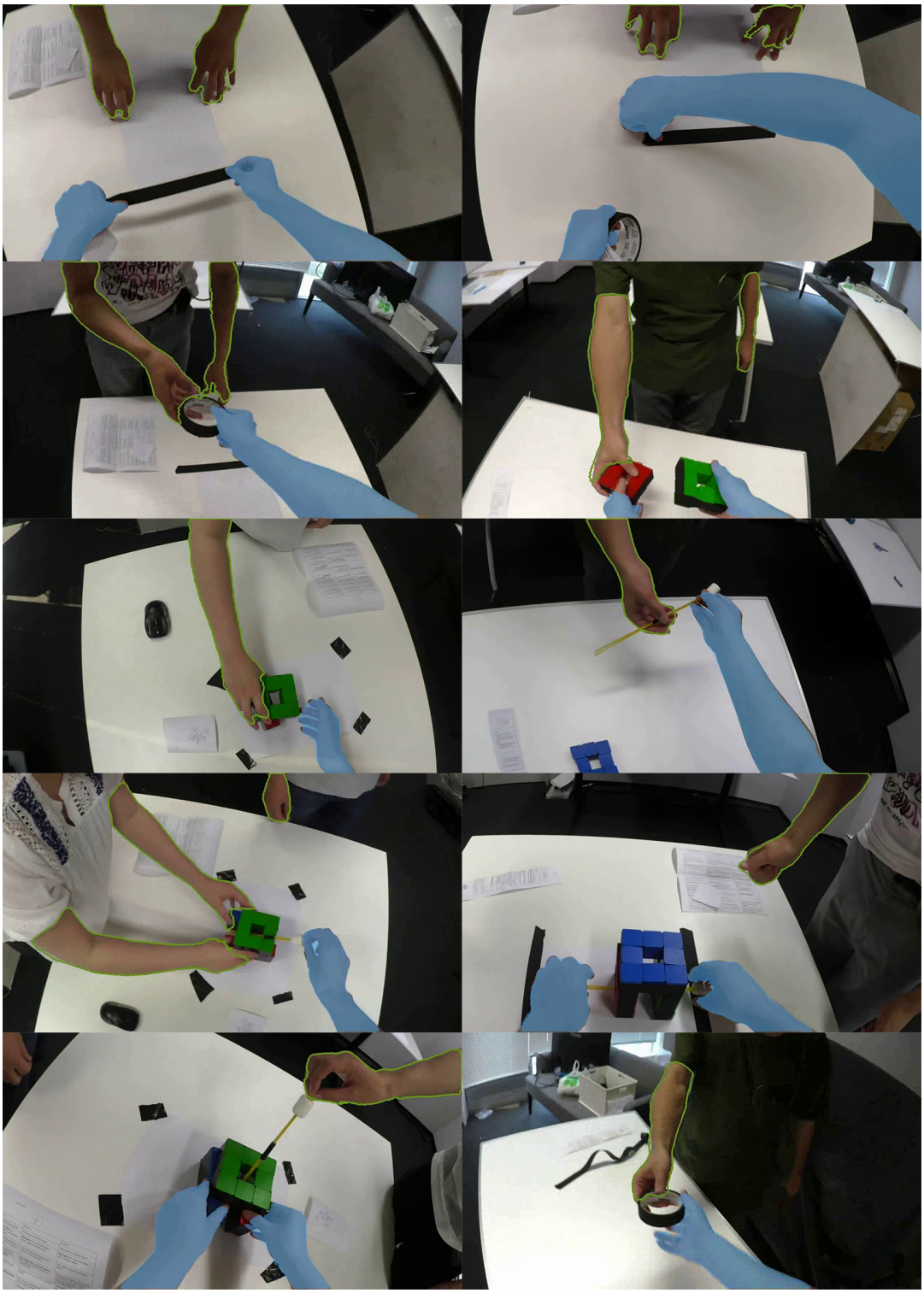

FIGURE 4.4: Example of both owner and other's hands detected scenes. The detected owner's hand is highlighted in light blue color, while the detected other's hand is highlighted with light green contour

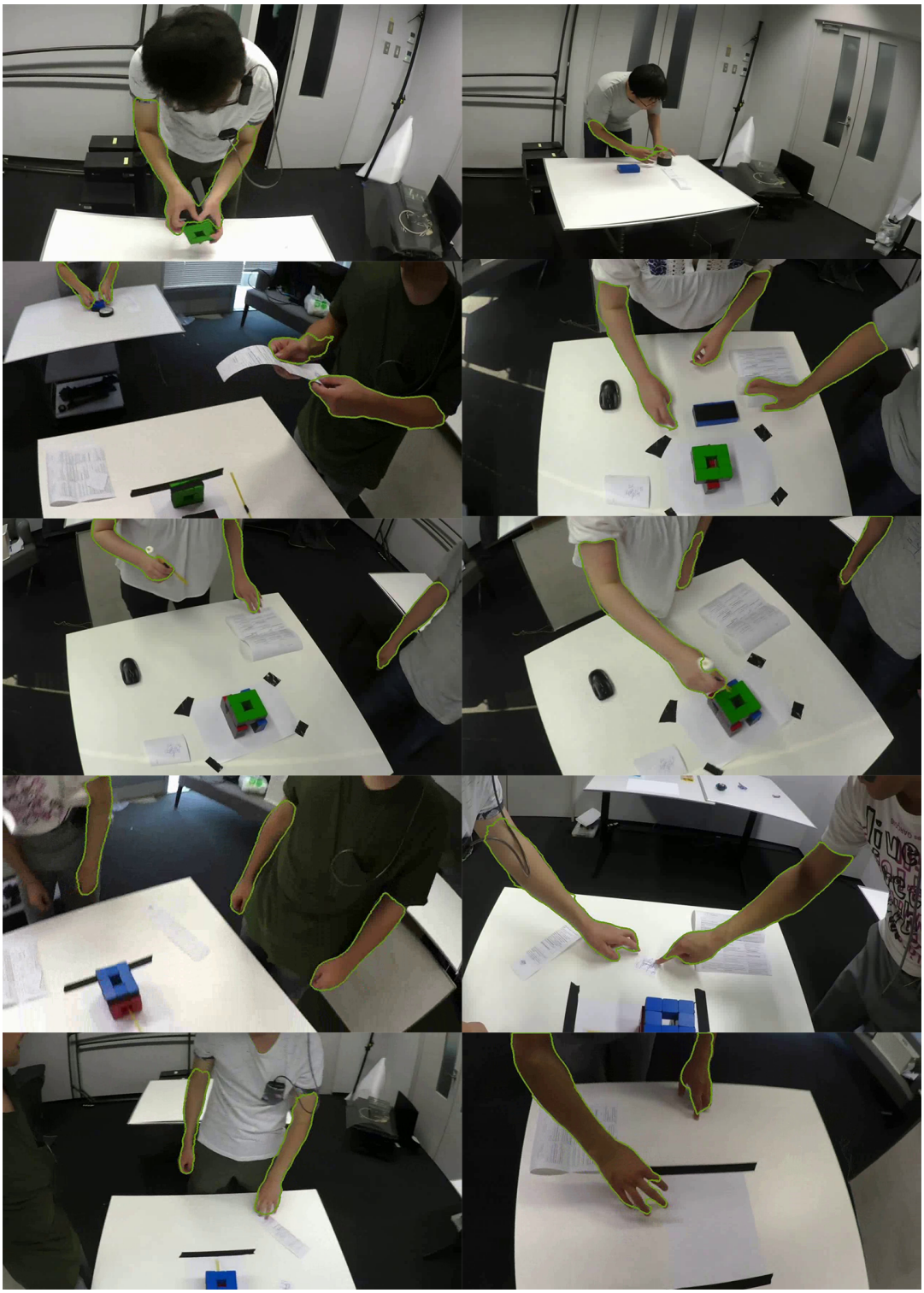

FIGURE 4.5: Example of only other's hand detected scenes. The detected hand classified as other's hand is highlighted with light green contour

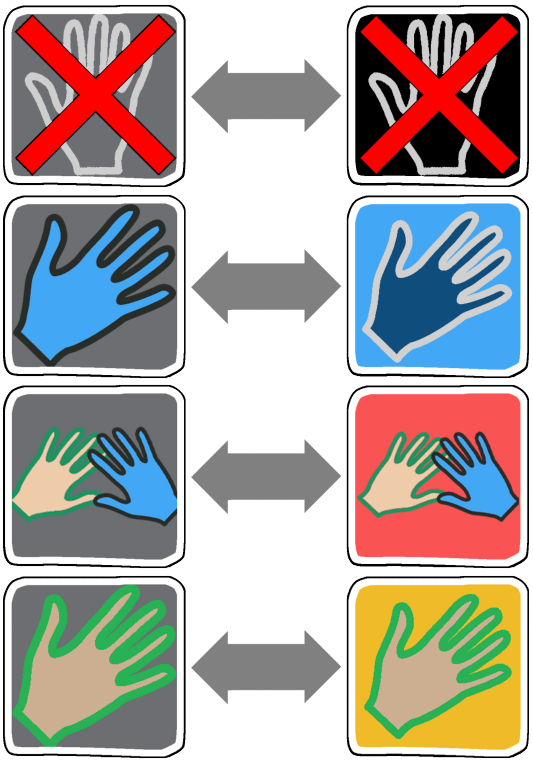

FIGURE 4.6: Each cue icon's image for the icon's state. (left) Gray is used as a background color for representing deactivate state. (right) Specific color (black, light blue, pinkish red, and yellow) is used as a background color for representing activate state. The icons from top to bottom are represented hand cue as follows: no hand detected, only owner view's hand detected, both owner view's hand and other's hand detected and only other's hand detected

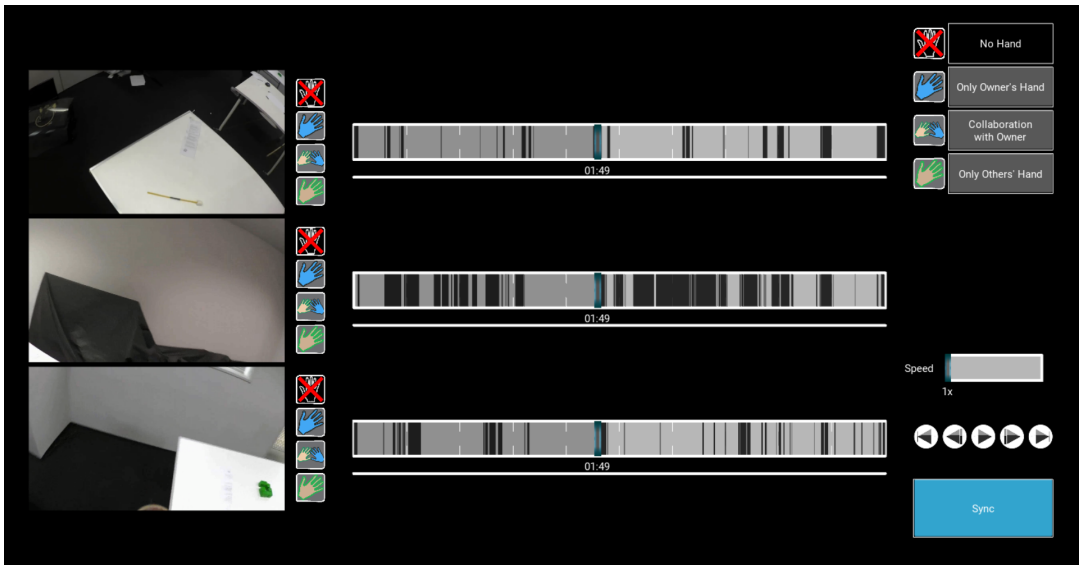

FIGURE 4.7: Interface when no hand cue is activated

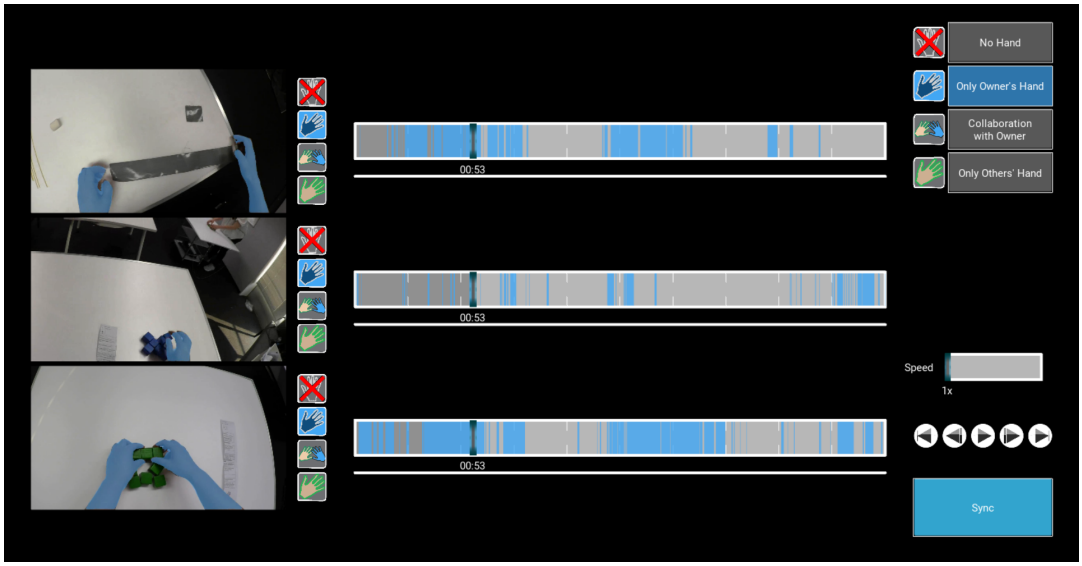

FIGURE 4.8: Interface when only owner's hand cue is activated

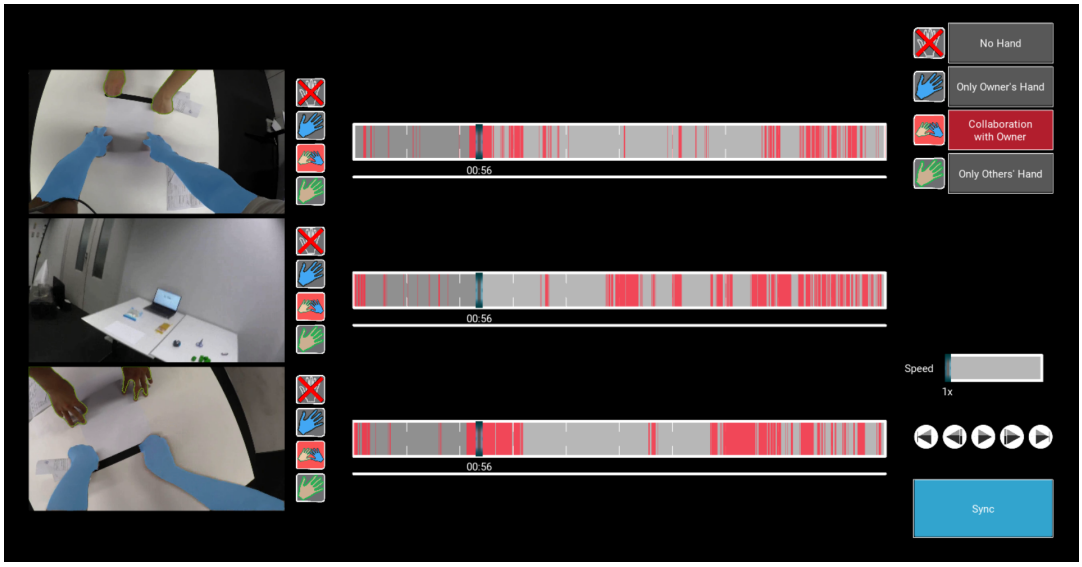

FIGURE 4.9: Interface when collaboration with owner cue is activated

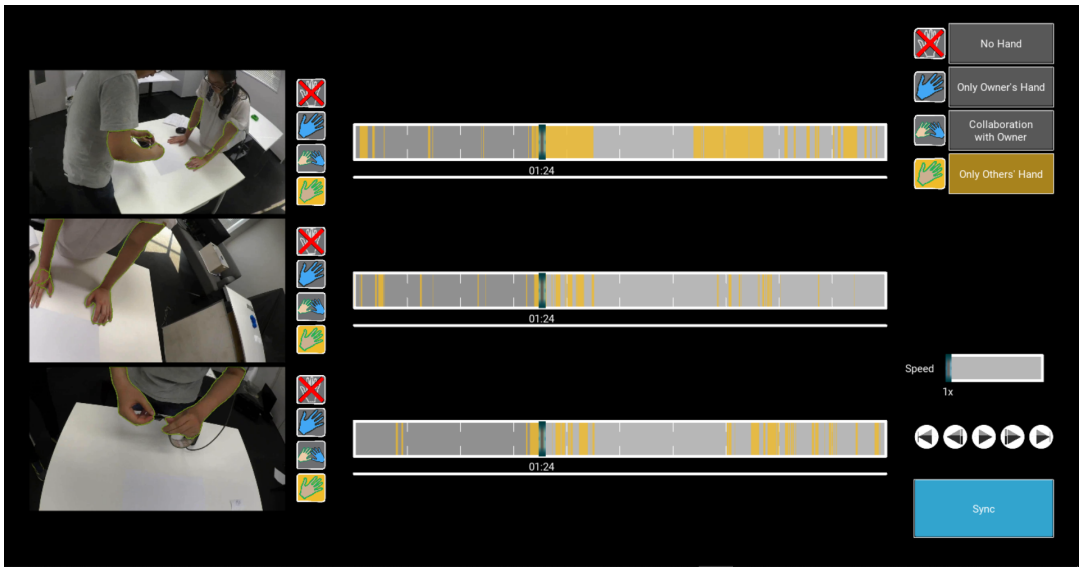

FIGURE 4.10: Interface when only others' hand cue is activated

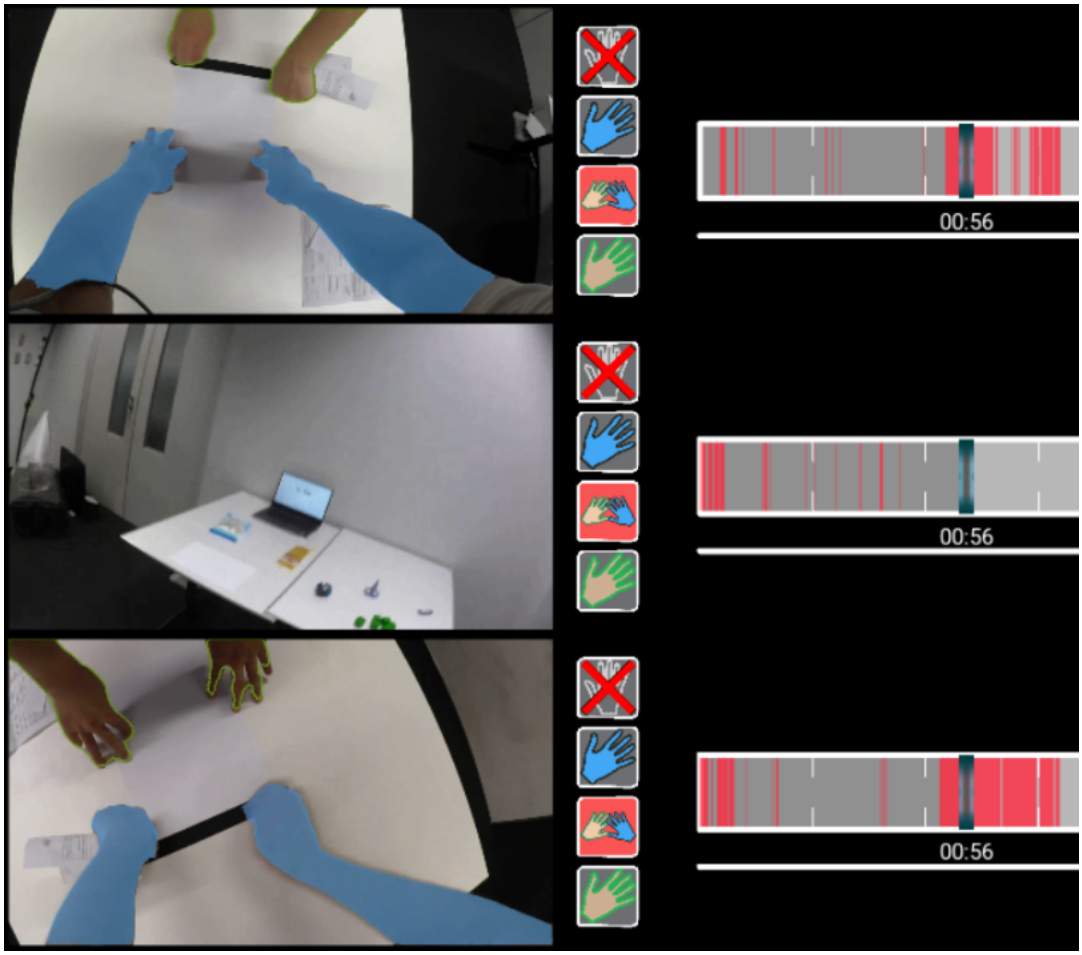

FIGURE 4.11: Visualization on multiple timelines can help as a hint for who is working with whom

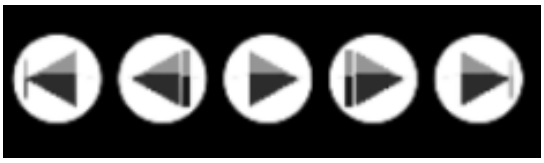

FIGURE 4.12: Video controller button. a) go to start, b) go to the previous frame, c) play and pause, d) go to the next frame, e) go to end

Chapter 5

Data collection

We have collected five datasets for using as a representative video records. The content in the dataset is a semi-control scripted for the block assembling task in a group created for the evaluation tasks.

5.1 Procedure

Before starting the recording, The room structure which has resource space, shared working space and personal working space as shown in Figure 5.1 was explained to each actor who acts as a worker in a group. Each actor participated in the record was given a scripted personal order of actions and a part of assembling blocks picture at the beginning of the record. After they finish working on their personal work, they need to come to the shared space in the middle of the room looking at the final goal and assembling all the parts together. There was a timer to control the whole group activities to finish within the length of four minutes.

Some actions are scripted in order to control the number of happened events for the evaluation task which are as follows:

- The number of times handing objects to other workers
- The number of times getting new materials from the resource table.
- The number of times each worker need to work in his/her private space

But still maintaining the randomness within the four different datasets by changing some related factors as follows:

- The order of action
- The worker identity in the joint action
- The list of interacted objects for each worker
- Personal working space position
- Actor identity

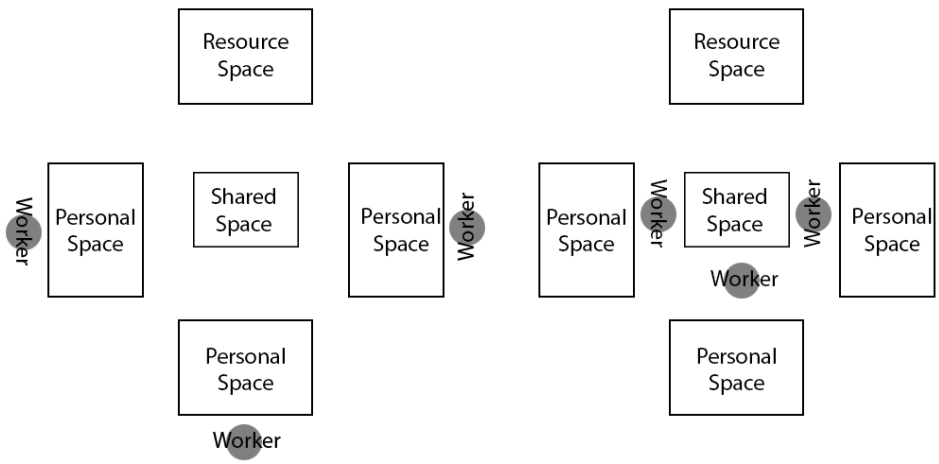

FIGURE 5.1: Recorded room structure. (left) Position for each worker working on their personal work, (right) Position when the workers are working together at shared space

5.2 Participants and devices

We used six different actors to form two different groups of actor and recorded four different sequences of action video for the equality of randomness in the evaluation process. Each actor wore a wearable camera (Panasonic wearable camera model HX-A500E) for recording video in first person view.

5.3 Material

The interacted objects use in the task are composed of following:

- Eight pieces of the wood cube in three sets of blue, green and red color
- One black tape
- One piece of A4 size paper
- Five sticks of spaghetti
- One piece of marshmallow

Chapter 6

Usability testing

6.1 Method and participants

The designed interface prototype as shown in Figure 6.2 was evaluated for its usability through four tasks with 12 peoples (Female: 4). Half of them are graduate students in computer science and engineering fields and half of them are from diverse industries which frequently use a computer in their work. The first three tasks are designed for measuring its effectiveness which we compare it with the base interface as shown in Figure 6.1 and the last task is designed for exploring their usefulness.

The first task is designed for testing the effectiveness of both owner of perspective's hand and other's hand detected cue can lead to specific collaborative action between two members in a group. Moreover, we also test if the effect of data visualization on the parallel timeline can lead to assisting collaborators' identity distinction within the group.

Second, the task is designed for testing the effectiveness of owner's hand detection cue if it can lead to the range of individual working time of each worker within a group.

Third, the task is designed for measuring the effectiveness of the other's hand detection cue if this visualized data can lead to finding observing phase from a worker. However, we understand that it will be too broad using only

other's hand detection feature for finding specific events because depending on this data alone will be too noisy data. Therefore, we add the combination of an owner of perspective's hand cue for observing the effectiveness of using multiple cues at the same time too.

Finally, in task 4 we designed this task for exploring the way to use these hand cues for finding the appointed scenes. These appoint scenes are also designed for finding the personal work, collaborative work and observing state with a different context. While we allow the tester to freely select the provided features as they see fit for each situation.

FIGURE 6.1: Interface without assisted feature

6.2 Procedure and setting

The protocol for the usability testing is asking the participant to find the assigned scene while using the provided interface. We divide the participants into 4 groups in order to generate a counter-balanced order of interface and video contents. The procedure order of each group is shown in Table 6.1. While the video alpha have totally different environment content and

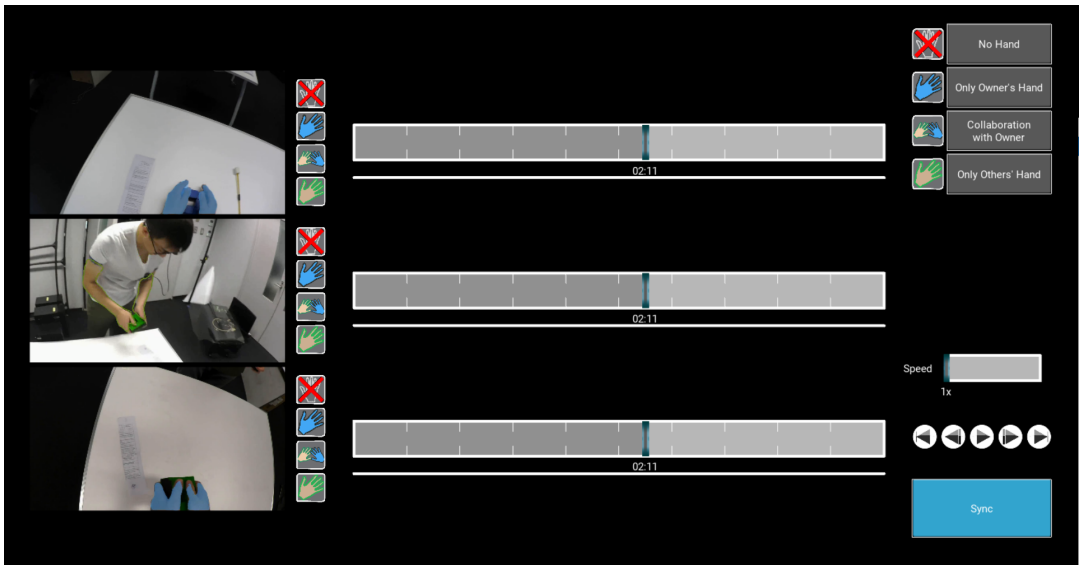

FIGURE 6.2: Interface with assisted feature

only two workers compare to the video use in testing phase which has three workers, it still has similar content of action for the sake of mutual understanding of the searched events definition such as handing object directly to another worker, pick up new resources, observing other member and assembling blocks in the personal space. The first set of the actor is in video 1 and 2, and the second set of the actor is in video 3, 4 and beta.

We start with an explanation of each component in the interface. Then, we explain the meaning of four cues to the participant by enabling each feature and show the example scene of each cue. After that, we explain how to mark the scene on multiple timelines.

When the tester found the specified scene, there are two types of annotation: one frame within events and time interval that the events happened. The type of annotation that the tester needed to use depend on each task. If the user wanted to annotate one frame, they need to move the cursor to the desired time and then press **1**, **2** or **3** to specify which timeline they want to annotate, then the white line will appear on the annotation area as shown in Figure 6.3. Together with the long-range annotation, they need to specify

Brief explanation of component in assisted UI				
Tutorial with assisted UI using Video alpha (2 workers)				
Task	**Group 1**	**Group 2**	**Group 3**	**Group 4**
Task 1	no assisted UI Video 1	no assisted UI Video 2	assisted UI Video 1	assisted UI Video 2
	assisted UI Video 2	assisted UI Video 1	no assisted UI Video 2	no assisted UI Video 1
	Questionnaire Task 1			
Task 2	no assisted UI Video 1	no assisted UI Video 3	assisted UI Video 1	assisted UI Video 3
	assisted UI Video 3	assisted UI Video 1	no assisted UI Video 3	no assisted UI Video 1
	Questionnaire Task 2			
Task 3	no assisted UI Video 2	no assisted UI Video 4	assisted UI Video 2	assisted UI Video 4
	assisted UI Video 4	assisted UI Video 2	no assisted UI Video 4	no assisted UI Video 2
	Questionnaire task 3			
Task 4	assisted UI Video beta			
	Questionnaire Task 4			
Feedback				

TABLE 6.1: Procedure and order of video datasets and interface used for each group

starting point by pressing **1**, **2** or **3** and ending point by pressing **4**, **5** or **6** according to each timeline. For more clear example, when the user wanted to mark timeline of the first worker(top), then they need to move the cursor to starting time then press 1 and then move the cursor to ending time then press 4 and then the mark will appear on the mark timeline as shown in Figure 6.4.

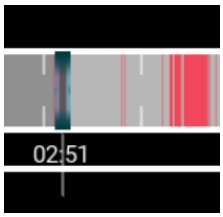

FIGURE 6.3: The white line appears on the annotation line when the user press according number to mark 1 frame

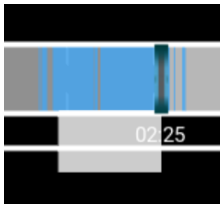

FIGURE 6.4: The white band appears on the annotation line when the user press ending number (4, 5 or 6) to mark for range annotation

After finish the explanation of the controller within the interface, we went on letting them practice using this interface with sample tasks and asking them to press **A** before starting each task, and press **S** when they finish the task in order to record the operation time. The video content used in the tutorial task consists of two workers as shown in Figure 6.5

After the participants got accustomed to the controller inside the user interface, we move on to the testing task with the order according to the group which that participant belongs to without their awareness. The common protocol is asking the participant to find the scenes in videos which match with the provided description. For the first three task, the participant needs to do the same task on two different interfaces with different content of videos, and

only one interface on the fourth task. After finish each task, we give them the questionnaire asking about the task. Then continue on until finish every task. Then we ask them for additional feedback.

FIGURE 6.5: Interface of tutorial phase

6.2.1 Task 1: Handing object (Collaboration action)

The participant is told to find 4 handing object events between 2 workers, and mark only 1 frame per event, but they have to identify who are the ones related to the event by marking on the timeline of related workers. The example of handing object event is shown in Figure 6.6 Therefore, they have to mark 4 pairs of a white line on the timeline.

6.2.2 Task 2: Assembling blocks on private space (Individual work)

In this task, the participants need to annotate all the interval time that each worker is assembling blocks on their own space. The participants are told that the maximum interval per worker is two and the minimum number

FIGURE 6.6: Example of directly handing object between two workers

of the interval is one. The interval is count as separate when the worker in the video change their action from assembling to other actions such as handing an object to other member or observing other members.

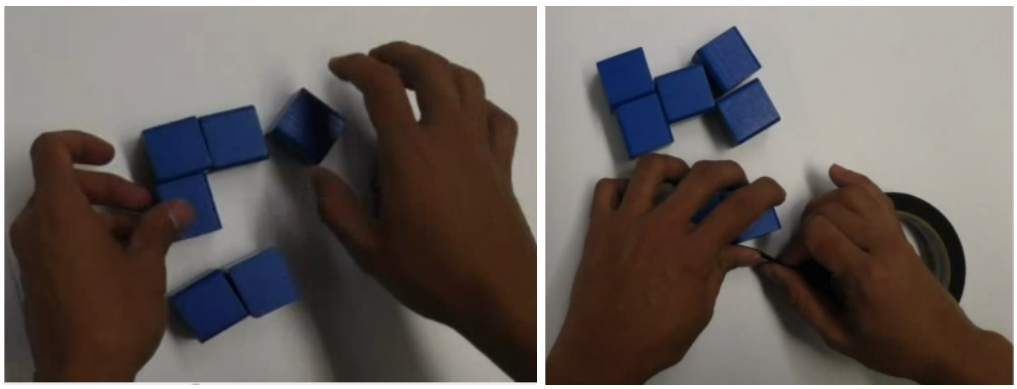

FIGURE 6.7: Example of individual assembling blocks on private space of each worker

6.2.3 Task 3: Observing another worker assembling blocks (Observer)

In order to evaluate if we can get observer state with provided cues, the participant is given with the only other's hand detected cue and the only owner's hand detected cue, and need to identify the interval time that each worker is observing when another member is assembling blocks in their own

personal space. The participant had been told that each worker will have only one interval for controlling the number of answers.

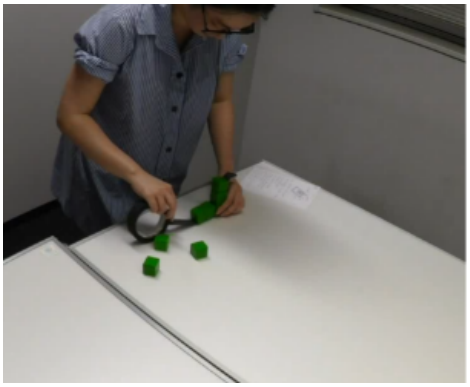

FIGURE 6.8: Example of a worker is observing another worker assembling blocks on his or her private space

6.2.4 Task 4 (Usefulness)

Finally, the last task is divided into three events which are one frame of one worker pickup spaghetti from resource table, one interval of two workers sticking paper together at the shared space, and one interval of one worker who acted as observer for two workers working on sticking paper at the shared space. The participants are asked to search for these events with the freedom to choose the provided features. They are encouraged to use the cue buttons as much as possible, and we record the selection history for analyzing the usefulness of the provided cues in various situations.

6.3 Evaluation Measures

6.3.1 Task Completion time

We expected that shorter completion time can be one of a sign for ease of task completion, so we measured a task completion time for each task.

FIGURE 6.9: (left) Example of pickup spaghetti. (middle) Example of two workers is sticking paper together in a shared space. (right) Example of one worker is observing another two workers are sticking paper at shared space together

Therefore, we compared the time usage between an unassisted interface and assisted interface quantitatively by non-pairwise t-scores.

6.3.2 Questionnaire

After finishing each task in the experiment, we let the participant answer the questionnaire which related to the task. The questionnaires for task 1 to task 3 are seven-point scale. The questions given to the participant are "How do you rate the ease of completing the task?" (very difficult = 1, normal = 4, and very easy = 7) per each task. Then, we investigated the ease of the task completion between an unassisted interface and assisted interface quantitatively by using the Wilcoxon signed-rank test.

In the last task, we asked the participants to explore the cues that they found useful for finding the target scenes. The scenes were assigned similarly as the prior tasks above for the following actions: 1) individual action, 2) collaborative action, and 3) observer action. The questionnaires given to the participant are "Which features did you find useful in the target event?". The participants are allowed to answer with multiple selections or select one as not found useful feature.

6.3.3 User observation and feedback

During the time participants worked on the assigned task, we observed them how they used each feature to find the events, and took notes on how they used the controller in the interface. After a participant completed all of the assigned tasks, we let them fill additional comments in the questionnaire form and discuss verbally. While filling the additional feedback, the participants were encouraged to consider the usefulness of features in assisted interfaces, how they found each cue useful and how easy to complete given tasks, and suggestion for function or design aspects for improving the interface.

6.4 Results

6.4.1 Statistical results

The statistical results compared between unassisted and assisted interfaces are shown in Figure 6.10 regarding average time completion per task and the ease of task completion for each task. The average and standard deviation of time completion for each task are shown in Table 6.2. As well as the average and standard deviation score from the questionnaire about the ease of task completion are shown in Table 6.3.

		Average (seconds)	Standard deviation
Task 1	Unassisted	289.08	163.25
	Assisted	302.75	154.72
Task 2	Unassisted	359.42	186.45
	Assisted	300.42	146.96
Task 3	Unassisted	354.08	206.31
	Assisted	273.17	133.11

TABLE 6.2: Average time completion and standard deviation for task 1, 2 and 3

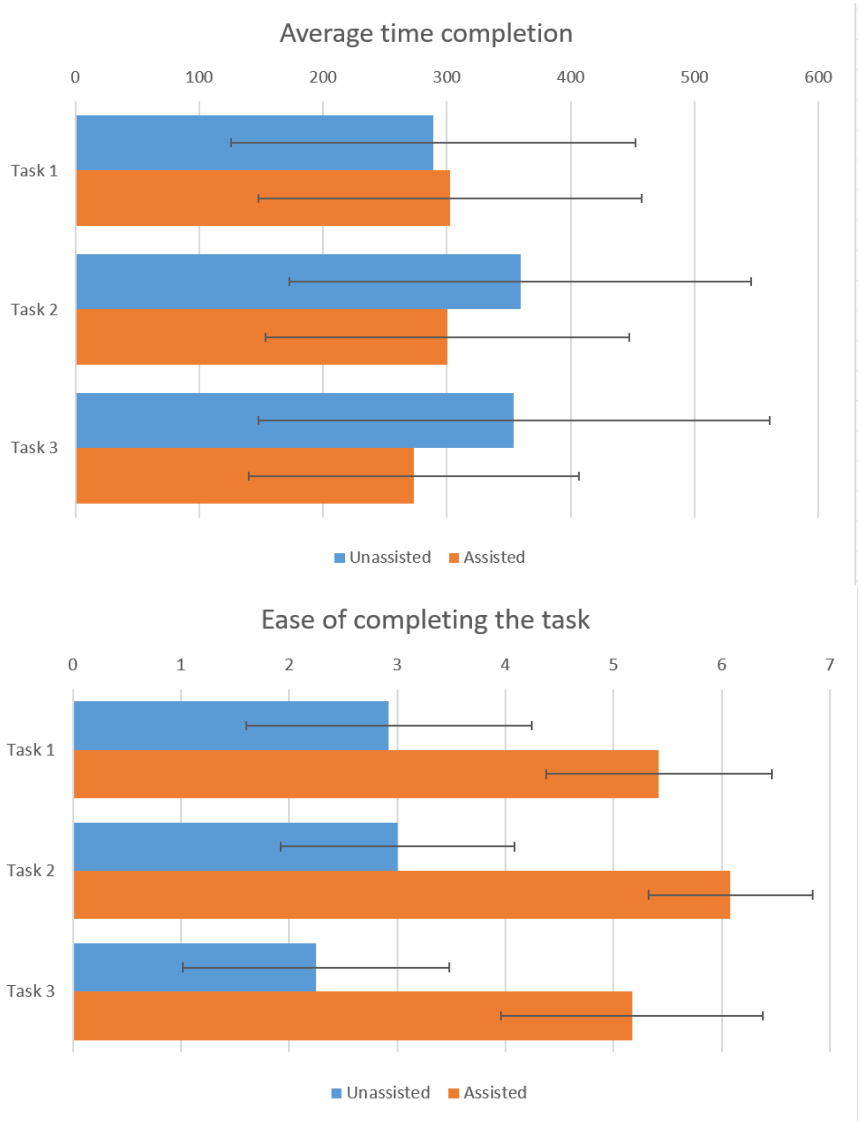

FIGURE 6.10: Statistical results a)completion time for each task, b) the ease of completing each task

The non-pairwise t-scores displayed no significant differences in the average time completion with task 1 ($p = 0.86$), task 2 ($p = 0.43$), and task 3 ($p = 0.27$) respectively to each task. However, when comparing the usability scale for task ease between two conditions, the Wilcoxon signed-rank test exhibited significant differences for each task respectively ($p = 0.0055$, $p = 0.0020$, and $p = 0.0019$). This showed that the ease of completing the task with the assisted condition is higher than that of unassisted. We specifically saw the

		Average score	Standard deviation
Task 1	Unassisted	2.92	1.32
	Assisted	5.42	1.04
Task 2	Unassisted	3.00	1.08
	Assisted	6.08	0.76
Task 3	Unassisted	2.25	1.23
	Assisted	5.17	1.21

TABLE 6.3: Average score of ease of task and standard deviation for task 1, 2 and 3

effect in task 1 because its average time performance in the unassisted interface was slightly faster than the assisted but led to a higher ease in task completion.

6.4.2 User observation and feedback

The result of the questionnaire from task 4 is shown in Figure 6.11. From the result, there is an outstanding cue for each task that the participants found it useful. In the task of finding a worker pickup spaghetti from the resource table, the owner's hand cue is found useful with 100% agreement (12/12). While the no hand cue and other's hand cue is found useful too as an additional data in some participants. In the same way, finding the interval of time which 2 workers are sticking paper together has a unanimous answer that collaboration cue is helpful for this task, and some participant also considers the owner's hand cue and the other's hand cue as an additional information for more precise search. However, in the task of finding an observer observed the 2 workers are sticking paper together have two outstanding answers instead of one which is only other's hand cue and collaboration cue.

Afterward, we get the feedback from the participant which can be categorized based on the visualization of hand cues, the design of the user interface and the subjective understanding of given tasks.

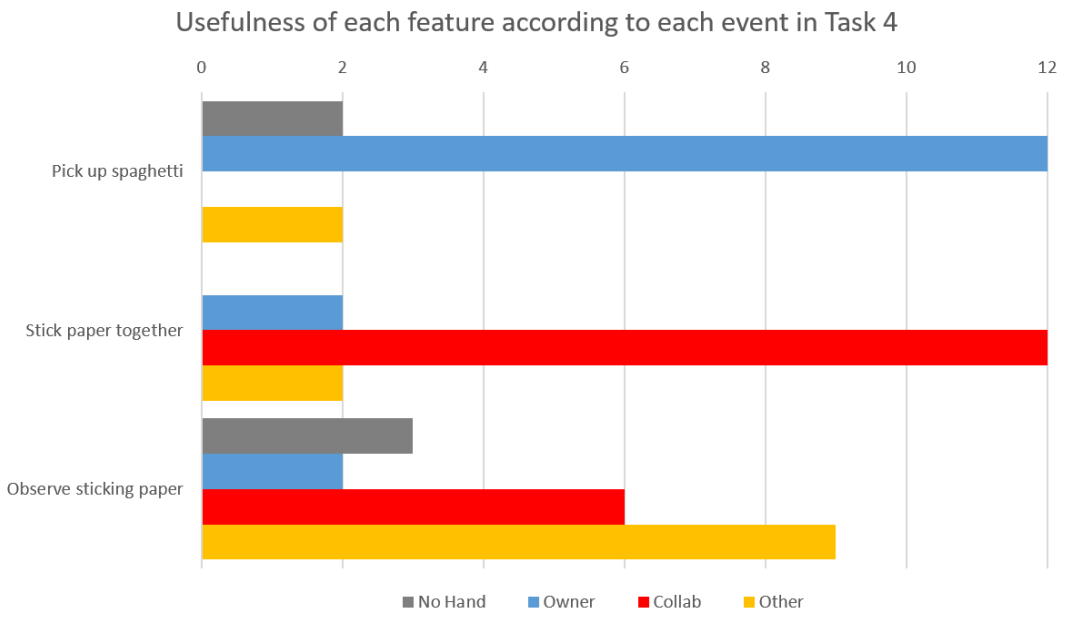

FIGURE 6.11: Usefulness ratio respond to each events in Task 4

Feedback about the hand cue visualization

We observed that the participant was able to use the visualization of hand cues in designed interface for accomplishing the assigned task as they said "*The highlighted part gives me the confidence that I have gone through all important part which needs to be checked thoroughly,*" "*I can go straight to the highlighted part and limit the search area to smaller area,*" "*Each button help to figure out the specified task, especially for task include more than one person,*" and "*I can find the interval time of action easier thanks to the area of highlighted color on the timeline.*"

While we have a lot of positive feedback, some participants showed concern about the accuracy of visualization data and large amount of visualization data on the timeline as they said "*The visualization information already helps a lot in finding the specified event, but it will be even more helpful if the accuracy is higher,*" and "*I felt confused and could not concentrate much on the video contents because there are too many videos and visualization information which need to pay attention at the same time.*"

For the new feature suggestion, many participants suggest having object detection within the system for enhancing the proposed cues. Some of them only mentioned object detection, but some of them suggested for finding an object only when the hand is detected.

Feedback about the design of interface

After listening to the feedback about hand cue, we moved on asking for comment about the interface design. Several of them said that they were not used to watch the first-person video, and in addition to watching multiple videos at once made them confused and did not know where to start looking. Most of them said that they like the design of video controller which allowed them to move through video with different size step. They also like the visualization of hand cue that shows on the parallel timeline because it helps them recognized the action-related worker easier.

In the same time, the participants also suggested some design points that they think the interface should have. There are a number of participants suggest that the visualization may utilize the intensity of color to represent the area coverage percentage of hand area detected. Furthermore, one participant said that "*The overall of data visualization is good for understanding the flow of events, but I want to be able to zoom up for specific part*". There is also a concern of scroll direction of a mouse which we use if for forwarding or backward 1 second. Some of the participants normally use the opposite direction of the system's scrolling direction which made them felt confused, so they decided to use only keyboard and button on the interface.

Feedback about the given task

In task 3 which required the participants to look for the worker who acted as an observer, several participant mention that it is difficult to notice that a worker is observing others as one of them said "*There is a time an actor observed*

another worker by glancing for a while resulted in less movement of their head, I have to observe from the other worker's view instead to be able to notice that the worker was observing another worker."

6.5 Discussion

6.5.1 Effectiveness

Even though the result of the average time completion especially from task 1 which the participant finish slightly faster in the unassisted condition, the result of ease of task completion in assisted condition is much higher due to their perception affect their mental effort. Searching for specific action within 3 egocentric videos of 4 minute-long is a laborious task because they can only brute-force through all of the video content, while the assisted one can act as a guide for them and they feel more at ease and less frustrating when they missed some scene with normal playback.

According to the feedback, the participant can feel more confident to go through the whole timeline in the video with a faster speed, and more comfortable to come back and forth. They also gave the feedback that the provided hand cues are meaningful for indicating these target scenes.

6.5.2 Usefulness of only the owner's hand detection feature for finding the interval time of individual work

According to the statistic result of questionnaire and the feedback from participant in task 2 (finding interval time of personal assembling blocks for each worker) and task 4.1 (finding a worker pickup spaghetti from resource table), the only owner's hand detection can provide the point of time which such personal activity should be located in and also provide the length of time that the activity has been done. The personal work in task 1 and task 4.1

can also be differentiate by the length of interval since the assembling task is a long time action, while the pickup action is an action that happens in a short time between other action especially between walking action which shows up as no hand detection because the flow likely to walk to the resource table, then locating the object and pick up object, and then go back to their personal space. Therefore, some participants also use another cue as an additional information, but all of them agreed that the only owner's hand appearance inside the first-person video is really helpful for finding personal working activity.

6.5.3 Usefulness of both owner's hand and other's hand detected for indicating collaboration work

In a similar way, the both owner's hand and other's hand detected cue are all agreed from their questionnaire that it is useful for finding collaboration action such in task 1 (finding handing object event) and task 4.2 (finding two workers sticking paper together at the shared space). Analogously, the handing object and sticking paper can also be distinguished by the length of interval time, since handing object is a short time action and working on the same area sticking paper together is a long time action. The participant also makes use of locations of visualization on the parallel timeline to figure out that at that location should be the collaborative action by having the similar characteristic of a pinkish red area between parallel timeline.

6.5.4 Usefulness of only other's hand detection for finding the observation state of worker

The usefulness of only other's hand detection is not that strong for finding the observation state of worker due to its accuracy and noisy information, but most of the participant found that this cue is useful when the worker is really

on a focus action especially when they moving in to close distance to the one they take an interest to. Therefore, when the only other's hand detection failed to detect, some of the participants can adapt to look at the additional information like only owner's hand detection or both side's hand detection depend on what the workers are observed working on.

6.5.5 Limitation

During the experiment, the participant mentioned that even with the provided data is there, they sometimes still need to go through the whole sequence when it is a short time action. This result is caused by the complexity of scenes which contain multiple hands overlapping together and the current system techniques could not distinguish the hand's identity from the cluster of hand. Furthermore, the participant reported that the interacted object detection within the hand area will be very useful for decrease the search area according to the current existing cues.

Chapter 7

Conclusion

We explored a computer-vision assisted detection which incorporates the hand detection scheme. The evaluations were conducted using multiple first-person videos which are recorded in the same working timeline for each collaborator. Then, the interface was designed with the objective to investigate the potential of hand cue in hope of expanding the scope of analysis techniques of the collaborative behavior analysis. Furthermore, we investigated the effects of the provided data to assess collaboration activities analysis. As the result of conducted user studies have shown, we get the confirmation that different kind of hand cues, especially the owner of the perspective hand-related, are useful for identifying certain actions. We are convinced that this investigation can lead to a new area of collaborative behavioral analysis.

With the potential of using hand cue within the egocentric video for collaborative work analysis, we believe that the interesting direction for future work of this research is to provide more hand cues by specifying the object-related activity so that more information can be obtained. In addition, we might also apply the technique from [25] to increase the accuracy of hand detection in the preprocessed stage.

Bibliography

[1] Noriko Suzuki et al. "The Effects of Group Size in the Furniture Assembly Task". In: *International Conference on Human Interface and the Management of Information*. Springer. 2017, pp. 623–632.

[2] Noriko Suzuki et al. "Detection of division of labor in multiparty collaboration". In: *International Conference on Human Interface and the Management of Information*. Springer. 2013, pp. 362–371.

[3] Noriko Suzuki et al. "Nonverbal behaviors in cooperative work: a case study of successful and unsuccessful team". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 29. 29. 2007.

[4] Mutlu Cukurova et al. "The NISPI framework: Analysing collaborative problem-solving from students' physical interactions". In: *Computers & Education* 116 (2018), pp. 93–109.

[5] Mutlu Cukurova et al. "Machine and Human Observable Differences in Groups' Collaborative Problem-Solving Behaviours". In: *European Conference on Technology Enhanced Learning*. Springer. 2017, pp. 17–29.

[6] Peter Wittenburg et al. "ELAN: a professional framework for multimodality research". In: *5th International Conference on Language Resources and Evaluation (LREC 2006)*. 2006, pp. 1556–1559.

[7] Kåre Sjölander and Jonas Beskow. "Wavesurfer-an open source speech tool". In: *Sixth International Conference on Spoken Language Processing*. 2000.

[8] *ELAN: A professional framework for multimodality research*. https://tla.mpi.nl/wp-content/uploads/2017/02/Screen_ELAN_494.png. Accessed: 2018-07-04.

[9] Gregorio Convertino et al. "Articulating common ground in cooperative work: content and process". In: *proceedings of the SIGCHI conference on human factors in computing systems*. ACM. 2008, pp. 1637–1646.

[10] Nathaniel Blanchard et al. "Automatic Classification of Question & Answer Discourse Segments from Teacher's Speech in Classrooms." In: *International Educational Data Mining Society* (2015).

[11] Bryan L Pellom and John HL Hansen. "Automatic segmentation of speech recorded in unknown noisy channel characteristics". In: *Speech Communication* 25.1-3 (1998), pp. 97–116.

[12] Steve Cassidy and Jonathan Harrington. "Emu: An enhanced hierarchical speech data management system". In: *Proceedings of the Sixth Australian International Conference on Speech Science and Technology*. 1996, pp. 361–366.

[13] Keita Higuchi et al. "Visualizing Gaze Direction to Support Video Coding of Social Attention for Children with Autism Spectrum Disorder". In: *23rd International Conference on Intelligent User Interfaces*. ACM. 2018, pp. 571–582.

[14] Rie Kamikubo et al. "Rapid Prototyping of Accessible Interfaces With Gaze-Contingent Tunnel Vision Simulation". In: *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM. 2017, pp. 387–388.

[15] Jeffrey F Cohn et al. "Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding". In: *Psychophysiology* 36.1 (1999), pp. 35–43.

[16] Isaac Wang et al. "EASEL: Easy Automatic Segmentation Event Labeler". In: *23rd International Conference on Intelligent User Interfaces*. ACM. 2018, pp. 595–599.

[17] L Chen et al. "(2017). Hotspots Detection for Machine Operation in Egocentric Vision. In 2017 15th IAPR International Conference on Machine Vision Applications (MVA). Institute of Electrical and Electronics Engineers (IEEE)." In: ().

[18] Shao Huang et al. "Egocentric Hand Detection Via Dynamic Region Growing". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14.1 (2017), p. 10.

[19] Kankana Roy, Aparna Mohanty, and Rajiv R Sahay. "Deep Learning Based Hand Detection in Cluttered Environment Using Skin Segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 640–649.

[20] Catharine Oertel and Giampiero Salvi. "A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue". In: *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM. 2013, pp. 99–106.

[21] Hayley Hung and Daniel Gatica-Perez. "Estimating cohesion in small groups using audio-visual nonverbal behavior". In: *IEEE Transactions on Multimedia* 12.6 (2010), pp. 563–575.

[22] Massimo Zancanaro, Bruno Lepri, and Fabio Pianesi. "Automatic detection of group functional roles in face to face interactions". In: *Proceedings of the 8th international conference on Multimodal interfaces*. ACM. 2006, pp. 28–34.

[23] Minjie Cai, Kris M Kitani, and Yoichi Sato. "An ego-vision system for hand grasp analysis". In: *IEEE Transactions on Human-Machine Systems* 47.4 (2017), pp. 524–535.

[24] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. *OpenFace: A general-purpose face recognition library with mobile applications*. Tech. rep. CMU-CS-16-118, CMU School of Computer Science, 2016.

[25] Xiaoming Deng et al. "Joint Hand Detection and Rotation Estimation Using CNN". In: *IEEE Transactions on Image Processing* 27.4 (2018), pp. 1888–1900.