Assisting Group Activity Analysis through Hand Detection and Identification in Multiple Egocentric Videos

Nathawan Charoenkulvanich
The University of Tokyo
Tokyo, Japan
nathawan@iis.u-tokyo.ac.jp

Rie Kamikubo
The University of Tokyo
Tokyo, Japan
rkamikub@iis.u-tokyo.ac.jp

Ryo Yonetani
The University of Tokyo
Tokyo, Japan
yonetani@iis.u-tokyo.ac.jp

Yoichi Sato
The University of Tokyo
Tokyo, Japan
ysato@iis.u-tokyo.ac.jp

ABSTRACT

Research in group activity analysis has put attention to monitor the work and evaluate group and individual performance, which can be reflected towards potential improvements in future group interactions. As a new means to examine individual or joint actions in the group activity, our work investigates the potential of detecting and disambiguating hands of each person in first-person points-of-view videos. Based on the recent developments in automated hand-region extraction from videos, we develop a new multiple-egocentric-video browsing interface that gives easy access to the frames of 1) individual action when only the hands of the viewer are detected, 2) joint action when collective hands are detected, and 3) the viewer checking the others' action as only their hands are detected. We take the evaluation process to explore the effectiveness of our interface with proposed hand-related features which can help perceive actions of interests in the complex analysis of videos involving co-occurred behaviors of multiple people.

CCS CONCEPTS

• **Human-centered computing** → **User interface programming**; *Usability testing*.

KEYWORDS

Group activity analysis; hand detection; first-person video; video-based analytic.

ACM Reference Format:
Nathawan Charoenkulvanich, Rie Kamikubo, Ryo Yonetani, and Yoichi Sato. 2019. Assisting Group Activity Analysis through Hand Detection and Identification in Multiple Egocentric Videos. In *24th International Conference on Intelligent User Interfaces (IUI '19), March 17–20, 2019, Marina del Rey, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3301275.3302297

IUI '19, March 17–20, 2019, Marina del Rey, CA, USA

ACM ISBN 978-1-4503-6272-6/19/03…$15.00
https://doi.org/10.1145/3301275.3302297

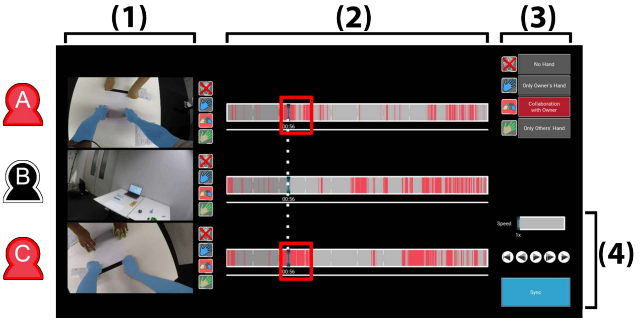

Figure 1: Our proposed interface, consisting of (1) first-person point-of-view videos captured from each person in group tasks, (2) video timelines highlighting actions of the viewer based on hand detection and identification, (3) hand-related feature buttons, and (4) video playback controls. The location of the seek bar on the parallel visualized timelines indicates e.g. the joint performance relationship of A and C in the group.

1 INTRODUCTION

Observing individual efforts in group tasks has brought insights into the quality of collective performance and the psychological evaluation of the individuals working in groups [8, 12, 15]. The authors from [12] have studied engagement levels of individuals correlated with the group size, and how the participants felt less satisfied towards the larger work-groups due to less individual contributions in the group tasks. As argued by Slavin [10, 11], individual contribution is one of the key features of successful group work, because it also facilitates team efforts in helping each other to complete the task.

To monitor how an individual engages in a group activity, researchers typically collect scenes of group task experiments using multiple fixed environmental cameras to deliberately capture the work of each participant in a group [14]. Gaze patterns [8], hand or head movements [3], and other multimodal data involving non-verbal and verbal units [6, 9, 14] serve as behavioral markers to analyze. In support of the video analysis, video browsing/annotation software such as ELAN is commonly used [17], and automatic detection of behavioral markers has been addressed [4, 16, 18].

While the researchers consider various behavioral markers in video-based group activity analysis, prior studies have taken into account hand interactions for clearly determining working states of individuals that reflect towards their engagement and contribution levels [2, 3, 12]. For example, the authors from [2] have speculated hand positions to label types of actions made by each person assigned in a group, such as when their hands are active in individually or jointly carrying out an assembly task.

In this research, we propose to explore the feasibility of detecting hands in the group task videos and see how it can impact the analysis of individual contributions in the group activity. Our work presents a designed interface (Figure 1) that incorporates an off-the-shelf deep neural network hand extractor [7] within the browsing features of multiple first-person point-of-view videos and highlights frames with specific working states, such as individual or joint actions for the hands detected, over corresponding timelines. Based on the design requirements identified from our preliminary study with two researchers who perform video-based group activity analysis, we aim to ease identification of individual and group performance from multiple videos, despite the presence of co-occurred behaviors and visually occluded scenes when multiple people are physically working together.

Towards extending the approach of the researchers in related works (e.g. [2, 12]) to examine various actions of interests performed by each person contributing to group tasks, we conducted a study to evaluate the effectiveness of our proposed interface in finding target scenes of three-person groups in assembly task videos. Our quantitative and qualitative findings demonstrated that the visualization features incorporating hand detection and identification over the use of multiple egocentric video timelines significantly ease the complexity in specifying individual actions as well as working relationships of individuals during group tasks. This contributes to offering new information layers and ease in the video-based group activity analysis.

2 INTERFACE DESIGN

2.1 Preliminary Study

To identify the requirements to design an interface that can assist the understanding of group activities, we conducted a preliminary study with two behavioral psychologists: R1) an academic researcher from the related works of [13, 14] and R2) a research and development researcher in group dynamic analysis. We had semi-structured interviews to understand procedural and technical problems in analyzing group activities. The requirements are summarized as follows:

- *Enable deliberate video capture of each individual performance that takes place in the working space.* As done in [14], a number of fixed side view cameras is usually expected to compensate for the positions of multiple people in the scenes causing visual occlusion to observe who is working or what they are working. R2 reported *"it would be nice to have a bird's eye view camera"*, but there is an additional cost to install in the new experimental setup.
- *Assist identification of individuals and their fine-grained working states while observing multiple synchronized videos.* R1 described the complexity of multiple-behavior annotation tasks to carefully speculate the micro-actions and working relationships of individuals.

2.2 Observing Multiple Egocentric Videos with Hand-Related Visualization Features

Following the aforementioned requirements, our interface leverages egocentric videos to deliberately capture individual performance in group tasks. The interface is also inspired by the design of [5], the visualization of key video frames to support the identification of actions of interests such as individual- or joint-working efforts from multiple videos.

As shown in Figure 1, our interface is composed of four main parts: (1) first-person point-of-view videos assigned to each member in group tasks, (2) corresponding video timelines for visualizing each member's working state, (3) control buttons to toggle the way of visualization, and (4) video playback controls. We show highlights over the video timelines, which supposedly indicate the scenes (examples in Figure 2) with the following actions of interests as grounded in [2]:

a) Passive action (black highlights) No hands are detected from the egocentric videos, estimating idle working states.
b) Individual action (blue) The first-person-view camera detects only the viewer's hands, estimating individual working states on tasks.
c) Joint action (red) The first-person-view camera detects collective hands of both the viewer and the others in a group, estimating joint working states on tasks.
d) Implicit action (yellow) The first-person-view camera detects hands except for the viewer's hands, estimating states of checking the others' individual or joint actions. Note, those others' actions are reflected on the other timelines with the corresponding individual or joint-action highlights.

Classifying video frames into these four features needs not only hand detection but hand identification of the camera wearer which we used as a new means to identify the action of interests for group activity analysis. Given the prior works in linking the levels of physical performance of individuals based on hand positions or movements [2, 12], we infer that automation of such behavioral descriptors may contribute to ease identification of micro-actions during the group tasks.

3 IMPLEMENTATION

In order to extract actions of individuals based on hand information, we start with the hand region segmentation and then follow with hand identity distinction. The hand region detection is based on the procedure in Khan *et al.* [7] to deal with the rapid change of background scenes in the environment within egocentric videos. The model hand detector composed of hand pixel classifiers is trained and fine-tuned with the 20 frames of hand masks for the environmental conditions within the dataset.

After a video of the probability-map is generated from the hand model detector, we best select the hand region by filtering out noise regions which are regions that have an area less than 1 percent of the video area and misdetection of face regions using OpenFace library [1]. Then, we define the viewer's hands when the egocentric

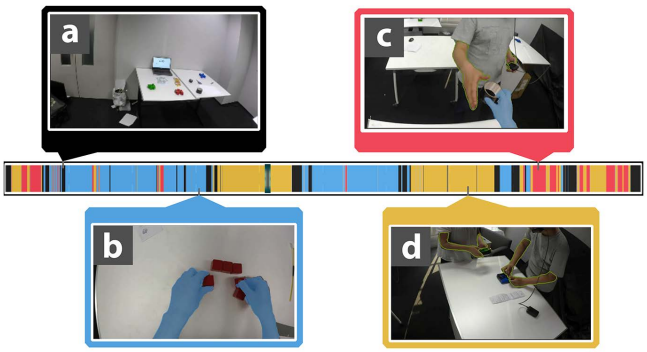

Figure 2: Sample frames visualized for each provided feature: a) no hand detected, b) only viewer's hand detected, c) both viewer and others' hands detected, and d) only others' hands detected.

view involves the perspective hand by checking if the hand's contour comes from the bottom line of the video. In the case of more than two contours on the bottom line, we select the center-most region since the perspective hands tend to be in the center of the frame. After that, we classify each frame into the four features according to the defined identity of the hands. There are some errors occurred because of the frame-by-frame hand detection. For example, when two hands' contours touch each other, this causes the system to detect as one single hand and sometimes leads to misclassification.

Vertically-aligned toggle buttons are presented for the visualization features to control the highlights on the timelines. The segmented hand regions over the first-person point-of-view videos are represented with color overlays, such as blue-filled contours for the viewer's hands, whereas green outline contours for the others' hands.

4 EVALUATION STUDY

We conducted an evaluation study which was designed to see whether the users of our proposed interface can make use of hand-related visualization features to find target scenes. These scenes from our multiple video datasets are under the following categories: 1) individual work by a specific person, 2) joint work of two or three people, 3) implicit work of a specific person checking the others' individual or joint actions, and 4) idle work, as shown in Figure 2 for examples.

Our main hypothesis involves: *the interface and its features can help the users a) identify scenes of different working states of individuals, and b) find and understand the content of individual and group performance with more ease of mind.* We validate the hypotheses through task-assigned quantitative measures and qualitative feedback of 16 participants (4 females, age: 23 - 32), which consist of graduate students in computer science and engineering fields and people from diverse industries with the frequent use of computers.

4.1 Data collection

Contents in the datasets are semi-controlled scripts of assembly tasks in a group of three workers. We had six actors divided into two groups and collected their first-person video recordings, and one dataset with two actors used in the practice tasks. Each worker has their own personal workspace to work on their task and shared workspace for assembling all parts according to the assigned goal. Main materials used for assembling are colored wooden blocks, spaghetti, marshmallow, paper, and adhesive tape. The scripts controlled the number of times of specific events for the participants to find target scenes.

4.2 Experimental Procedure

Each session lasted at most 120 minutes including the total of 4 assigned tasks to find the target scenes from multiple video datasets. The participants were first given with an explanation of the interface and the provided features, followed by the practice tasks. Next, we moved to the test tasks, in which the order of conditions to use the proposed interface or the baseline interface was counterbalanced. After finishing with all of the tasks, the participants answered questionnaires regarding the tasks conducted and responded to short interviews for additional feedback.

We designed the following four tasks for finding assigned events from synchronous multiple first-person videos, and the participants had to mark the related frames on the timeline of each presented video. We compared the results of the first three tasks between conditions to use our proposed interface and the baseline interface with no visualization. We recorded the selection logs on the last task when the participants freely utilized the features according to their preference.

Task 1 The participants were asked to find four scenes of joint actions, corresponding to the events when a worker gave specified objects to another worker. They enabled only the "joint action" feature to visualize the frames of detected collective hands in the condition to use the proposed interface.

Task 2 The participants were asked to find all of the scenes of individual actions, corresponding to the events when a worker is assembling blocks in their own working space. They enabled only the "individual action" feature in the condition to use the proposed interface.

Task 3 The participants were asked to find specific scenes of a certain worker observing actions of other members individually assembling blocks in the working space. They enabled the "implicit action" and "individual action" features in the condition to use the proposed interface.

Task 4 The participants were asked to search for the specific events with the freedom to choose the provided features. They were encouraged to use the features as much as possible. Similar to the first three tasks, the assigned events involved: one worker is picking up some spaghetti from the table, two workers are sticking paper together in the shared space, and one worker is observing two workers sticking paper in the shared space.

4.3 Evaluation Measures

4.3.1 Task Completion Time. We expected that shorter time can be one of the signs for ease of task completion, so we compared the time usage between an unassisted and assisted interface quantitatively by pairwise t-tests.

4.3.2 Questionnaire. After finishing each task from 1 to 3, the participants were given with the question of "How do you rate the ease of completing the task?" in the seven-point scales (found very difficult = 1, very easy = 7). Then, we investigated the ease of the task completion between an unassisted and assisted interface quantitatively by using the Wilcoxon signed-rank test.

In the last task, we asked the participants "Which features did you find useful in finding the target events?" The participants were allowed to answer with multiple selections.

4.3.3 User observation and feedback. We observed the participants' task performance on how they used each feature to find the events and the controls on the interface. After they 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 reflect back the ease of using the features in the assisted interface to complete given tasks, how they found each hand-related actions useful, and suggestions for function or design improvements.

5 RESULTS

5.1 Statistical results

Table 1 shows the statistical results compared between unassisted and assisted interfaces regarding average time completion and the ease of task completion ratings for each task. The pairwise t-scores displayed no significant differences in the average time completion ($p = 0.61$, 0.42, and 0.18 respectively for each task). However, the result of the rating scores exhibited significant differences based on the Wilcoxon signed-rank test ($p = 0.0008$, 0.0004, 0.0004 respectively for each task).

In the selection of features to use in task 4, we found the preference of having the hands of the individual actions detected which displayed the useful cues in the task of finding individual work with 100% agreement (16/16 participants). Likewise, with 100% agreement (16/16), detection of collective hands for joint actions was found useful in finding scenes of joint-working efforts. To find scenes of a person checking the others' working performance, the participants, however, responded less useful of the feature detecting only the hands of the others (11/16). They had to combine with individual or joint action features to fully make sure that the corresponding scenes include the others' active actions.

5.2 User feedback and observation

Overall feedback from the participants confirmed that the visualization of hand-related working states in the timelines offer them ease as well as confidence in finding target events. One participant told us that *"The highlighted part gave me the confidence that I have gone through all important parts which need to be checked thoroughly."* The visualization also reduced the cognitive demands of the participants by limiting the search area from the timelines of multiple videos as one of the participants said that "*I can go straight to the highlighted part and scan the smaller search area.*"

Table 1: Average time completion and the score of ease of completing the task with standard deviation for task 1, 2 and 3

	Unassisted: Average (Std)	*Assisted: Average (Std)*
Task completion time		
Task 1	301.06 (190.73)	268.75 (147.97)
Task 2	326.06 (174.13)	279.31 (135.04)
Task 3	337.56 (190.66)	258.00 (121.40)
Ease of completing the task		
Task 1	2.81 (1.29)	5.56 (1.00)
Task 2	3.19 (1.33)	6.06 (0.75)
Task 3	2.44 (1.27)	5.44 (1.17)

The parallel timelines of multiple videos with hand-related visualization features were found effective to see the relationships of individuals working in a group. They said that *"Each button helps to figure out the specified task, especially for task including more than one person."* The participants can identify action-related individuals based on the synchronous frames indicating the correlation of actions.

In terms of the usability of the designed interface, the highlighted frames were regarded as useful for the long interval actions because it was clear to see the visualization. The participants rarely noticed the short frames indicating a short-span individual or joint actions. One participant mentioned that "*I can find long interval time of action and mark their start and end point easier thanks to the area of highlighted color.*" At the same time, a large amount of displayed visualization data on the timeline also had negative impacts, as one participant mentioned that *"I felt confused and could not concentrate because there is too much information shown at the same time."*

6 DISCUSSION

Our study aims to investigate whether our supportive user interface can decrease the task load for grasping working states of individuals in cooperative work. We have received the results of significant differences in the ease of task completion. The visualization for egocentric video frames can act as a guidepost, and let the participants feel less frustrating even when they miss some scenes. They can go back and forth through the whole content over the timelines with more confidence.

Based on the questionnaire and feedback, we can confirm that the proposed features incorporating hand detection and identification are useful for finding actions of interests. Detection of individual hands and collective hands was especially preferred for identifying *individual* and *joint* actions respectively. Detection of the *implicit* action was not always working due to its functional design and accuracy limitations. We also received feedback on the parallel timelines of the interface, which were found to be effective in identifying joint action related members.

While we collected positive comments about the visualized timelines, the design of showing highlights further needs usability improvements. For example, we need to avoid the users from overlooking important scenes caused by focusing mostly on the frames with accentuated highlights. We also observed how the users failed to notice short-interval target actions due to frame-by-frame highlights and overcrowded visualized information. These issues could be resolved by improving the look of the interface such as through context-aware zoom-in timeline or increasing accuracy of the visualization.

Current system techniques are limited in distinguishing the hand's identity based on detected frame-by-frame hand information. This sometimes leads to errors in classifying actions of interests from the cluster of hands, especially for short-interval actions. Furthermore, a quarter of participants suggested having more depth in the features by detecting the object of manipulation for limiting the search area.

7 CONCLUSION

We investigate the potential of leveraging detected hand-related actions on the synchronous multiple first-person videos interface for expanding the scope of analysis techniques of the group activity analysis. Our evaluation study findings support how our hand-related features, especially for detecting the individual and joint actions, are useful for identifying certain group activity events. Future work will be to improve the functional elements for the accuracy of hand detection, as well as design components of visualization features to enable obtaining more specific information such as with object manipulation. We believe our investigation can open a new area in the complex behavioral analysis of observing both individual and group performance.

8 ACKNOWLEDGEMENT

We would like to show our gratitude to Dr. Noriko Suzuki of Kyoto University who provided expertise regarding the prior works that supported our study.

REFERENCES

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

[2] Mutlu Cukurova, Rose Luckin, Manolis Mavrikis, and Eva Millán. 2017. Machine and Human Observable Differences in Groups' Collaborative Problem-Solving Behaviours. In *European Conference on Technology Enhanced Learning*. Springer, 17–29.

[3] Mutlu Cukurova, Rose Luckin, Eva Millán, and Manolis Mavrikis. 2018. The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. *Computers & Education* 116 (2018), 93–109.

[4] Keita Higuchi, Soichiro Matsuda, Rie Kamikubo, Takuya Enomoto, Yusuke Sugano, Junichi Yamamoto, and Yoichi Sato. 2018. 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, 571–582.

[5] Keita Higuchi, Ryo Yonetani, and Yoichi Sato. 2017. EgoScanning: Quickly Scanning First-Person Videos with Egocentric Elastic Timelines. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6536–6546.

[6] Hayley Hung and Daniel Gatica-Perez. 2010. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia* 12, 6 (2010), 563–575.

[7] Aisha Urooj Khan and Ali Borji. 2018. Analysis of Hand Segmentation in the Wild. *arXiv preprint arXiv:1803.03317* (2018).

[8] Catharine Oertel and Giampiero Salvi. 2013. 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, 99–106.

[9] Haruka Shoda, Koshi Nishimoto, Noriko Suzuki, Mamiko Sakata, and Noriko Ito. 2016. Creativity Comes from Interaction. In *International Conference on Human Interface and the Management of Information*. Springer, 336–345.

[10] Robert E Slavin. 1991. Synthesis of research of cooperative learning. *Educational leadership* 48, 5 (1991), 71–82.

[11] Robert E Slavin. 1996. Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary educational psychology* 21, 1 (1996), 43–69.

[12] Noriko Suzuki, Mayuka Imashiro, Mamiko Sakata, and Michiya Yamamoto. 2017. The Effects of Group Size in the Furniture Assembly Task. In *International Conference on Human Interface and the Management of Information*. Springer, 623–632.

[13] Noriko Suzuki, Tosirou Kamiya, Ichiro Umata, Sadanori Ito, Shoichiro Iwasawa, Mamiko Sakata, and Katsunori Shimohara. 2013. Detection of division of labor in multiparty collaboration. In *International Conference on Human Interface and the Management of Information*. Springer, 362–371.

[14] Noriko Suzuki, Ichiro Umata, Toshiro Kamiya, Sadanori Ito, Shoichiro Iwasawa, Naomi Inoue, Tomoji Toriyama, and Kiyoshi Kogure. 2007. 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.

[15] Elora C Voyles, Sarah F Bailey, and Amanda M Durik. 2015. New pieces of the jigsaw classroom: increasing accountability to reduce social loafing in student group projects. *The New School Psychology Bulletin* 13, 1 (2015), 11–20.

[16] Isaac Wang, Pradyumna Narayana, Jesse Smith, Bruce Draper, Ross Beveridge, and Jaime Ruiz. 2018. EASEL: Easy Automatic Segmentation Event Labeler. In *23rd International Conference on Intelligent User Interfaces*. ACM, 595–599.

[17] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*. 1556–1559.

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