HandVis: Visualized Gesture Support for Remote Cross-Lingual Communication

Kuan-Yu Lin
Dept. of Computer Science
National Tsing Hua University
Hsinchu, 30013 Taiwan
s104062552@m104.nthu.edu.tw

Chien-Tung Lai
Dept. of Computer Science
National Tsing Hua University
Hsinchu, 30013 Taiwan
jacklai5505@gmail.com

Seraphina Yong
Dept. of Computer Science
University of Chicago
Chicago, Illinois 60637, USA
seraphina@uchicago.edu

Hao-Chuan Wang
Dept. of Computer Science
National Tsing Hua University
Hsinchu, 30013 Taiwan
haochuan@cs.nthu.edu.tw

Shuo-Ping Wang
Institute of Information Systems and Applications
National Tsing Hua University
Hsinchu, 30013 Taiwan
s102062507@m102.nthu.edu.tw

CHI'16 Extended Abstracts, May 07-12, 2016, San Jose, CA, USA
ACM 978-1-4503-4082-3/16/05.
http://dx.doi.org/10.1145/2851581.2892431

Abstract
Effective communication between those who are not fluent in a non-native language can potentially be quite difficult. The common language selected to be used throughout an exchange can encumber those who might not speak it as proficiently as others. Remote communication further heightens the difficulty since less channels are available for communication. We introduce HandVis, a video conferencing interface that visualizes elements of hand gesture, such as trajectory and amount. Gesture is intended to be a communicative tool that can compensate for language deficits. The results of a user study indicate how HandVis can be utilized constructively by less-proficient speakers during cross-lingual communication.

Author Keywords
Computer-mediated communication system; motion sensing; cross-lingual communication; enhanced videoconferencing

ACM Classification Keywords
H.5.3. Group and Organization Interfaces: Computer-supported cooperative work;

Introduction

As members of work groups and organizations are increasingly distributed to different locations, video conferencing tools are used on a regular basis for remote communication. However, it tends to be a less efficient medium than face-to-face (F2F), such as resulting in reduction in message understanding [10] and decision making quality [5].

Media richness theory (MRT) states F2F communication is a richer medium than video conferencing [3][4]. In video conferencing, interlocutors may experience less social cues and social presence than in F2F communication as the number and types of channels available for transmitting information are inferior to F2F. Video conferencing with mixed language proficiency (e.g., cross-lingual workgroups) adds an additional layer of difficulty for the users involved, since both linguistic difficulties and the loss of communication cues must be dealt with.

Mutual understanding is often an issue during cross-lingual communication, since individuals who are not very proficient in a language may not be capable of expressing everything they want to say using that language. Under these circumstances, gestures may be useful for communicating concepts that cannot be exchanged properly when linguistic difficulties are experienced. Gestures are known to be applied as a compensatory channel of communication when language deficits are present, and are also a powerful tool for discussing abstractions [5][7].

To mitigate potential difficulties caused by both a mismatch in language fluency and interaction over computer-mediated communication (CMC), we present a video conferencing interface called HandVis, which uses gesture visualization in video conferencing to heighten user awareness of gestures. This may encourage users to communicate using more gestures and improve the quality of such an interaction. In addition to increasing overall gesture, we also wish to increase iconic gesture use in particular. Iconic gestures (representing object features, actions, and spatial relations) have a communicative purpose, serving to aid the partner's understanding [1][5].

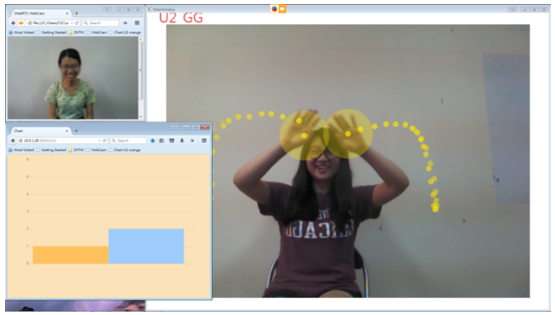

Figure 1: In the HandVis interface, the top left window represents first-person video and the right-side window shows the interlocutor's video stream with hand gesture visualization. The bottom-left window shows hand movement speed of both users in a bar chart. The chart's background color indicates the user's own bar color.

Visualization-Augmented Awareness and Use of Gesture

HandVis functions similarly to a standard video conferencing application with some supplementary features; the hand areas of users on-screen are overlaid with a translucent, bright color and the paths of users' hands are traced. HandVis integrates a live-update visualization of both users' gesture movement speed, which appears as a pair of fluctuating bars in each user's application window. (Figure 1)

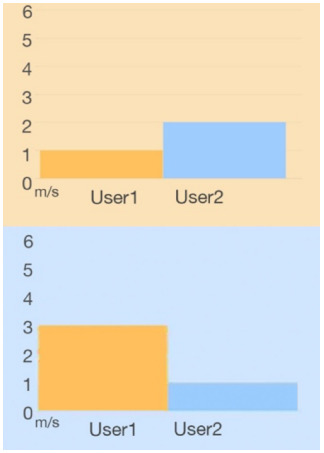

Figure 2: Bar charts of gesture speed.

Figure 3: The images of our gesture visualization shown in the pilot survey. It shows comparison between different colors.

The highlighting and motion tracing of a user's hand areas are intended to improve gesture awareness, expression, and understanding. Gestures may become more salient to the user with the addition of this visual emphasis. Saliency detection, as part of the pre-attentive component of human attention, causes it to be drawn to features like intensities and edges [8]. The neon color and clean edges of these visual cues contrast with footage to draw users' attention. The fluctuating bars are designed to encourage consciousness of making gesture, which may boost the amount of iconic, as well as non-iconic, gesture.

In a laboratory study, we investigated the effectiveness of this interface on mixed-language proficiency user pairs by having subjects perform brainstorming tasks together over a spectrum of communication media (audio, HandVis, and F2F communication). Iconic gesture amount and level of understanding were measured. We hypothesized that HandVis could be as conducive to effective communication of mixed-proficiency users as F2F communication, considering that the visualization and highlighting of gesture may make it a more useful channel for communication, compensating for the limited amount of visible social cues available in regular video conferencing [3].

In this initial analysis, we found that HandVis and F2F support message understanding during group brainstorming equally well, and that HandVis helped less proficient users (e.g., non-native English speakers) to speak up and contribute more, equalizing the level of participation between proficient and less-proficient speakers of a common language.

Visualization of Hand Gestures

Design Question

Previous research indicates that greater amount of cues available in communication supports more complicated tasks [3]. Though video conferencing supports richer social cueing than other CMC modalities like text messaging, there are still some restrictions. For example, delay and limited resolution of video streaming may obstruct communication. Therefore, in F2F communication, the level of social presence is higher than in video conferencing. But F2F communication requires co-location, which is impeded by geographical constraints. Therefore, to aim for better quality video communication, we consider the following design question: *can we use visualization effects like hand movement tracking to enhance existing social cues in video conferencing, to improve user comprehension similar to F2F and successfully collaborate on more complicated tasks?*

Prototyping

To address the issue, we present HandVis, which represents one's hand gesture in two ways. Both of these visualizations are displayed on our interface at the same time. Firstly, the position and trace of a user's hand would be highlighted on the video of the user. The system also measures the speed of each movement and represents the speeds as bar-charts next to the video. (Figure 1 and Figure 2.)

To track hand gestures and show the visualization, we use Microsoft Kinect to implement our system. The Kinect motion sensor is capable of tracing not only users' hand movement, but also other features such as head position and face feature points at the rate of 30 frames per second. This study only concerns users'

Figure 4: These images compare differing amounts of tracking dots.

Figure 5: These images vary in dot size, ranging between 15, 25, 35, 50 pixels.

Figure 6: These images vary the radii of the highlighting circles.

hand gestures, so we capture hand movement using Kinect sensor and its SDK, which is provided by Microsoft Developer Network (MSDN) in our system.

Hand Gesture Highlighting

We employed a pilot survey to collect naïve users' judgments on alternative visual designs of gesture highlighting (See Figure 3, 4, 5, and 6). Based on the results, we decided to represent the previous 20 points of hand position to trace hand movement and track 30 points per second. Tracking point radius is 15 pixels, and 90 pixels for current hand position. Both current hand position and trace colors are orange.

Speed of Hand Movement

To let users recognize each gesture's speed, HandVis also provides a dynamic bar chart to display both users' current hand motion speed. The chart background color corresponds to the user's bar color. (Figure 2.)

Evaluation Study

We conducted a within-subject user study with two conditions to evaluate the effects of our prototype and examine our hypotheses. The two conditions under comparison were HandVis and F2F. Since we are interested in mixed-proficiency communication, we arranged the participants into pairs. Each pair consisted of a fluent speaker and a non-fluent speaker of English. To control the conversation topic, we employed three brainstorming tasks used by a previous study [10] and asked users to discuss them during the experiment. Conditions and task order was counterbalanced.

Although participants were in the same room during the experiment, in the HandVis condition we used two 42-inch displays to display the interface and separate the participants, so they could not see each other directly. The experimental setting simulated video conferencing (see Figure 7). We connected users' Kinect sensors to their partner's computer, so the skeleton data and RGB video of each user can be sent to their partner's computer directly, without any kind of network. This minimizes instability or delay in the streaming. To sync the video streams, we used the webcam connected to the user's computer. Each user could then see both their own and interlocutor's video stream simultaneously. (Figure 7)

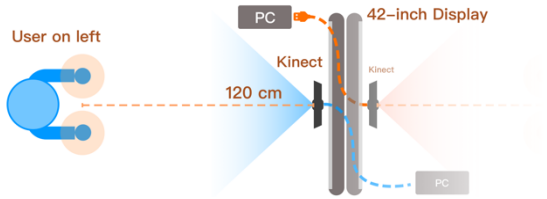

Figure 7: The configuration of our "simulated video conferencing" experiment. The Kinect of a participant was wired to the computer of this participant's interlocutor in order to prevent delay or bandwidth issues during video streaming.

Procedure

We recruited 12 naïve subjects (6 males, 6 females) to participate in our study, grouped into 6 pairs. Each pair consisted of one English-fluent speaker and one English second-language (ESL) speaker. For our study, the requirements for being a fluent speaker participant were the following: The participant should have lived in an English-speaking country for more than 7 years consecutively, and use >50% English in their work environments and daily lives. To ensure that the level of ESL speakers would be sufficient for a brainstorming discussion, we required that the participants should have passed any English examination which we

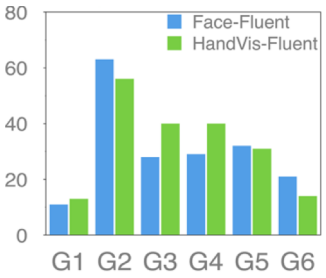

Figure 8: Number of iconic gestures by fluent speakers. Blue bars represent the F2F condition, and green bars represent the HandVis condition.

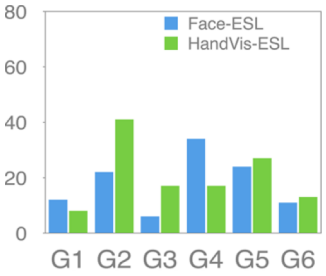

Figure 9: Number of iconic gestures by non-fluent speakers. Yellow bars represent F2F condition, and orange bars represent HandVis. We focus on G2 and G3 in our interviews, because they made significantly more iconic gestures in the HandVis condition.

considered equivalent to score a minimum of 80/120 on the TOEFL iBT test. Participants were assigned to their pairs randomly. All participants reported in the study questionnaire that they were not acquainted with their partner prior to our study.

Each pair was asked to discuss three brainstorming tasks. The instructions encouraged participants to discuss benefits, drawbacks and general implications of a hypothetical situation. The tasks are: all humans "having an extra thumb", "having an extra eye" and "having wings" in the year 2020, which have been used in previous research [10]. The discussion period for each brainstorming task was timed to last for 10 minutes.

After each brainstorming session, participants completed a questionnaire. We also conducted face-to-face interviews with participants after all tasks were completed and asked them to compare/contrast the two media based on their user experience.

Measures

Speech Amount

We segmented the transcript of each interaction into utterances and labeled each utterance with its speaker. We then counted the number of utterances by individual participants in a brainstorming session. The number of sentences contributed by each user was converted to a percentage ratio, representing the speech amount of a user relative to their partner's in the same brainstorming session.

Iconic Gesture

Two coders were recruited to count occurrences of iconic gesture in the video of each brainstorming session. Transcripts of all videos were produced and converted to subtitles. Coders watched videos with subtitles and marked words or phrases where users made iconic gestures matching the language context. Subjects sometimes made gestures shortly before or after the corresponding language; coders marked the words which triggered the gesture, or vice versa. Both coders watched all videos independently, and achieved satisfactory inter-coder agreement (Kappa=.98)

Understanding of Messages

After each task, each subject completed a self-report questionnaire with a set of four 5-point Likert scale questions for assessing the perceived level of message coherence people exchanged during the session. A sample question is *"my partner always clearly expressed his/her thoughts"*.

Results

We found that the two conditions resulted in similar levels of perceived message clarity ($F<1$, n.s.). This provided some initial support that HandVis can help mixed-proficiency pairs communicate as well as they do during F2F.

Also, some participants appeared to make more iconic gestures in the HandVis condition (see Figure 8 and 9). In particular, we observed that the non-fluent, low-proficiency speakers in groups G2 and G3 gestured a lot more in HandVis than in F2F. Respectively, G2, G3 gesture amounts were 41,17 for HandVis and 22,6 for F2F. While to fluent speakers, the influence of F2F and HandVis on the number of iconic gestures seems to be similar (see Figure 8).

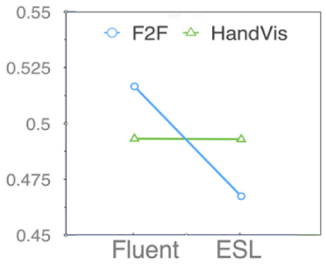

Figure 10: Amount of speech by individuals of different English proficiency when using different media. With HandVis, ESL speakers increased their speech amount to the level of fluent speakers.

What's very interesting is that there was a significant interaction effect between language proficiency (fluent, non-fluent) and condition (F2F, HandVis) on individual's speech amount, as identified by a mixed-model ANOVA analysis that accounted for pairing of individuals and repeated measures, F [1,10.86] = *9.00, p<.05*.

Previous work found that English-fluent speakers tended to be more talkative than ESL participants [9]. With HandVis, ESL participants turned out to speak more (M = 0.50, SD = 0.01) than they did in the F2F condition (M = 0.47, SD = 0.01). Most importantly, there was no difference in speech amount for English-fluent speakers (M = 0.50, SD = 0.01) and ESL (M = 0.50, SD = 0.01) under the HandVis condition (see Figure 10). Using HandVis increased non-fluent speakers' speech amount, equalizing the participation in the brainstorming activity between fluent and non-fluent individuals working together.

Interviews

From the interview, some participants did use HandVis in a beneficial way. As participant G3-B reported, when she noticed that the bar-chart indicated that her partner had a lower level of gesture movement speed, she would explain more about the idea.

G2-A told us that when he observed that the bar-chart showed higher level of gesture movement speed from his partner, he would make more gestures to compete with her.

Some participants reported that the HandVis visualization disturbed the conversation, as she could not focus well on the video stream, hand tracing, and visualization of gesture quantities simultaneously.

Communicating in a second language already requires more cognitive resources than doing so in one's native language, and the additional stress of having to process supplemental information could result in a backfiring effect.

Discussion

The result of the questionnaire indicated no difference in understanding between the F2F and HandVis conditions. With visualized gesture support, a video conferencing tool for remote communication can be as effective as co-located, face-to-face communication.

What's more surprising is that HandVis equalized the speech amount between fluent and non-fluent speakers. English-fluent speakers can easily dominate the conversation, and our tool has shown usefulness on avoiding the occurrence of this sort of dominance, which is important to many group activities.

To further evaluate the impact of HandVis, we will conduct further studies such as comparing HandVis to video and audio conditions. With hand movement visualization, people can "draw" things in the air using their hands. We can further explore this design space by adding more features (e.g., color selector, eraser) for gesture-augmented communication.

Acknowledgement

The work is supported in part by Ministry of Science and Technology, Taiwan, R.O.C. (102-2221-E-007-073-MY3 and 103-2218-E-007-017-MY3) and Microsoft Research Asia UR project.

References

1. Geoffrey Beattie and Heather Shovelton. 2002. An experimental investigation of some properties of individual iconic gestures that mediate their communicative power. *British Journal of Psychology* 93, 2 (May 2002), 179–192. DOI=http://dx.doi.org/10.1348/000712602162526

2. Herb H. Clark & Susan Brennan. 1991. Grounding in Communications. *Perspectives on Socially Shared Cognition*, Washington DC: APA, 127-149.

3. Richard L. Daft & Robert, H. Lengel. 1986. Organizational information requirements, media richness and structural design. *Management Science*, 32, 5 (May 1986), 554-571.

4. Alan R. Dennis & Joseph S. Valacich. 1999. Rethinking media richness: Towards a theory of media synchronicity. *Proceedings of the 32rd Hawaii International Conference on System Sciences.*

5. Susan R. Fussell, Leslie D. Setlock. 2012. Multicultural teams. In *Leadership in science and technology: A reference handbook*. 255-264. Thousand Oaks, CA: SAGE Publications, Inc. DOI=http://dx.doi.org/10.4135/9781412994231.n29

6. Spencer D. Kelly, Sarah M. Manning, and Sabrina Rodak. 2008. Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education. *Language and Linguistics Compass* 2, 4 (July 2008), 569–588. DOI=http://dx.doi.org/10.1111/j.1749-818X.2008.00067.x

7. Susan Goldin-Meadow. 2015. From action to abstraction: Gesture as a mechanism of change. *Developmental Review* 38 (Dec 2015), 167–184. DOI=http://dx.doi.org/10.1016/j.dr.2015.07.00

8. Anne M. Treisman and Garry Gelade. 1980. A feature-integration theory of attention. *Cognitive Psychology* 12, 1 (January 1980), 97–136. http://dx.doi.org/10.1016/0010-0285(80)90005-5

9. Hao-Chuan Wang, Susan F. Fussell, and Leslie D. Setlock. 2009. Cultural difference and adaptation of communication styles in computer-mediated group brainstorming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09). ACM, New York, NY, USA, 669-678. DOI=http://dx.doi.org/10.1145/1518701.1518806

10. Hao-Chuan Wang and Chien-Tung Lai. 2014. Kinect-taped communication: using motion sensing to study gesture use and similarity in face-to-face and computer-mediated brainstorming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14). ACM, New York, NY, USA, 3205-3214. DOI=http://dx.doi.org/10.1145/2556288.2557060