Toward Automated Analysis of Communication Mirroring

Kumiko Hosogoe, Miyu Nakano
Faculty of Social Welfare,
Iwate Prefectural University
Takizawa, Japan
email: hosogoe@iwate-pu.ac.jp, g221r007@s.iwate-pu.ac.jp

Oky Dicky Ardiansyah Prima, Yuta Ono
Graduate School of Software and Information Science,
Iwate Prefectural University
Takizawa, Japan
email: prima@iwate-pu.ac.jp, g231q005@s.iwate-pu.ac.jp

***Abstract*—Mirroring is a technique in which someone unconsciously reflects other people's behavior, such as gestures and facial expressions. This technique can help us to build rapport with others by making communication more effective and reflective. Due to the developments in computer vision, human behavior observation based on vision cameras has become viable. Moreover, the wide spread of omnidirectional cameras has made it possible to observe more people at the same time, making it easier to analyze face-to-face conversation scenes. In this study, we propose a framework to perform a time series analysis based on Dynamic Time Warping (DTW) to determine whether communication mirroring has been established. The framework uses human pose estimation techniques to track hand gestures of two persons during a conversation. The framework will detect all gestures that are similar to the trained gestures. Our experiments show that the DTW was able to detect mirroring acts having distinct gestures. However, detections of similarities of weak gestures are challenging.**

Keywords-communication mirroring; communication mirroring; human pose estimation; DTW; 360 degree camera.

I. INTRODUCTION

Gestures are forms of nonverbal communication that use body movements instead of words. These movements include the hand, head, or other parts of the body. Gestures enable to produce more intuitive communication and flexible interactions. Using gestures to reflect the behavior of the talking partner can create a strong connection during the conversation. These techniques are called mirroring, from a communication perspective. Mirroring can improve rapport with others. It happens very naturally when people are talking [1]. The ability to mirror others nonverbally facilitates empathy. Although many people have an ability to empathize with others, only a few people have excellent natural empathy.

Many studies have been conducted in conversation scene analyses. Most of them are multimodal, using cues produced by audio and video recordings [2]. Traditionally, analyzing this data involves works of manual coding of the repeating behavior, its duration, and response latency. BECO2, an integrated behavior coding system, has been widely used in universities in Japan to train students to perform behavior analysis [3]. With this system, observers can record and analyze the occurrence and duration of behaviors by pressing keyboard keys corresponding to those in each category.

Audio and video processing techniques have contributed to performing conversation scene analyses effectively. On the one hand, audio processing can reveal non-verbal behaviors, including utterances, acts of stressing, and speaking rate based on the speech signals. On the other hand, video processing can measure facial information and gestures. Otsuka and Araki [4] introduce Voice Activity Detection (VAD) to determine the presence/absence of utterances from speech signals. An omnidirectional camera-microphone system was used to partition an input audio stream into homogeneous segments according to the speaker's identity as being captured by the camera. This diarization is vital to identify different speakers' turns in a conversation.

Advanced computer vision applications have enabled the estimation of human posture from a single image [5][6]. These approaches are more flexible than those based on a depth camera. Depending on the Field of View (FOV) of the cameras, human posture from multiple targets can be effectively measured [7]. With an omnidirectional camera, it is possible to take photos of people as if it were taken from the front, even if they talk to face each other. Derivation of human posture from images opens up new ways to recognize gestures. The extracted gestures can be used to identify specific responses during a conversation. However, even with the same gestures, the length and speed of these gestures are different. Dynamic Time Warping (DTW) is a promising technique that can measure the similarity between two temporal sequences of gestures [8]. It enables a non-linear mapping of one signal to another by minimizing the distance between them.

This study proposes a framework for automatically detecting the presence of mirroring motion in hand gestures during a conversation between two people using an omnidirectional camera. The framework converts the video image obtained from the omnidirectional camera into a panoramic image and extracts the posture information of the conversation participants from the video. The Graphical User Interface (GUI) allows observers to select gestures as training data. The training data is used to estimate similarity to gestures found in the observed data. Detection of communication mirroring is done by thresholding the similarity of hand gestures between participants.

This paper is organized as follows. Section II describes related works on behavior coding and gesture analysis based on computer vision techniques. Section III introduces our proposed framework to analyze the presence of communication mirroring. Section IV describes our experimental setting and its results. Finally, Section V presents our conclusions.

 ISBN: 978-1-61208-761-0

II. Related Work

Traditional methods to measure nonverbal behavior rely on the manual coding system. Since there are a lot of ambiguities on judging a particular action, many observers tend to perform in an inconsistent manner, thus degrading the quality of the measurements. Jaana et al. [8] developed an automated behavior analysis system using a single omnidirectional camera. This system analyzed facial expressions, head nodding, utterances based on facial landmarks extracted from facial images captured by the camera. Schneider et al. [9] proposed a gesture recognition system using human postures obtained from a single camera, using a human pose estimation framework, OpenPose [5]. This system utilized DTW to classify the time-series data. Although this method yielded promising results, it has limitations in case of attempting to classify more similar gestures. For general-purpose gesture recognition, the Gesture Recognition Toolkit (GRT) is a cross-platform open-source C++ library designed to make real-time machine learning and gesture recognition more accessible [10].

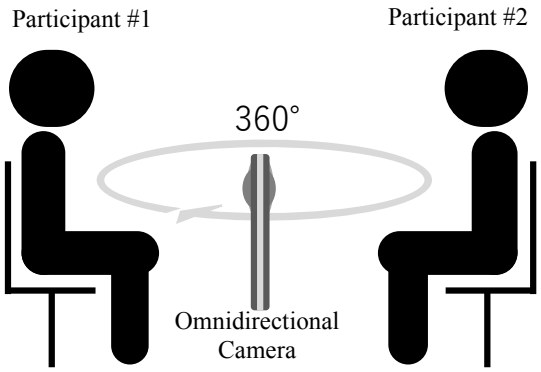

Figure 1. Experimental setting.

III. Communication Mirroring Detection

We build a framework to detect communication mirroring that occurs in a conversation scene, as shown in Figure 1. The omnidirectional camera captures an image of the whole bodies of the two communication participants.

A. Panoramic Image Projection

The omnidirectional camera produces two fisheye images to represent a 360° image. These images are combined and warped into a panoramic image, as shown in Figure 2, so that the information in the image can be interpreted directly. The panoramic image displays a 360° image as a rectangular image.

B. Pose Estimation

We use the OpenPose framework [5] to estimate the body posture of the communication participants. The skeletal information of the hands is extracted for the analysis. This information consists of six body joints, such as the wrist, elbow, shoulder, and neck. We normalize each joint to achieve translation and scale invariance [8]. Normalization is done by taking the neck joint as the origin of the coordinate system and subtracting this coordinate from all other joints.

C. Training

We provide a GUI to allow observers to select the typical movement of the participant, which can be considered as a gesture. The automatic data trimming will remove images that

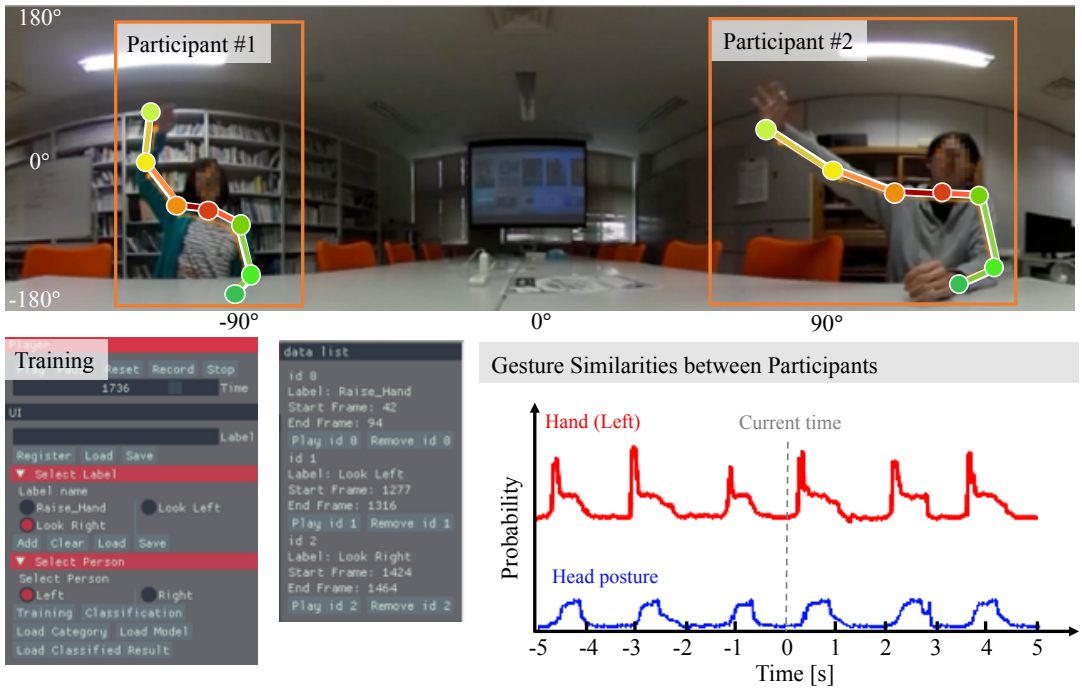

Figure 2. The user interface of our framework.

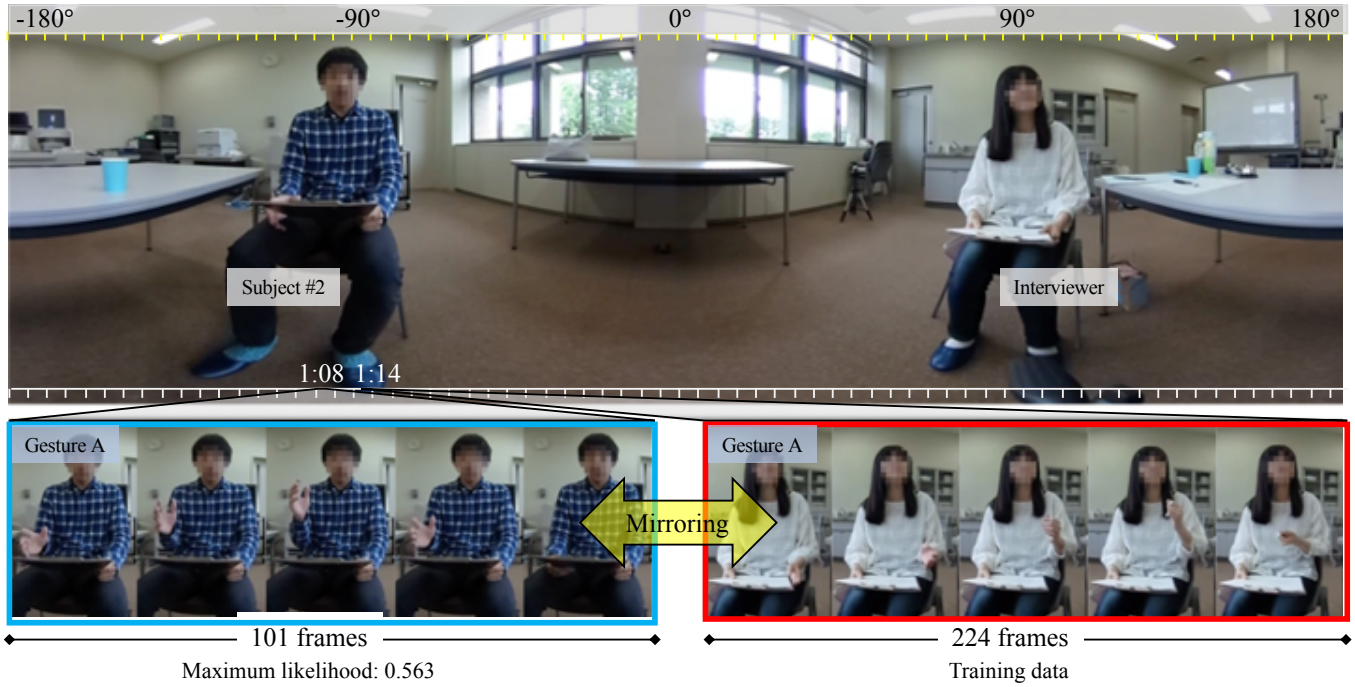

Figure 3. The successful case of similar gestures detected in Experiment 1.

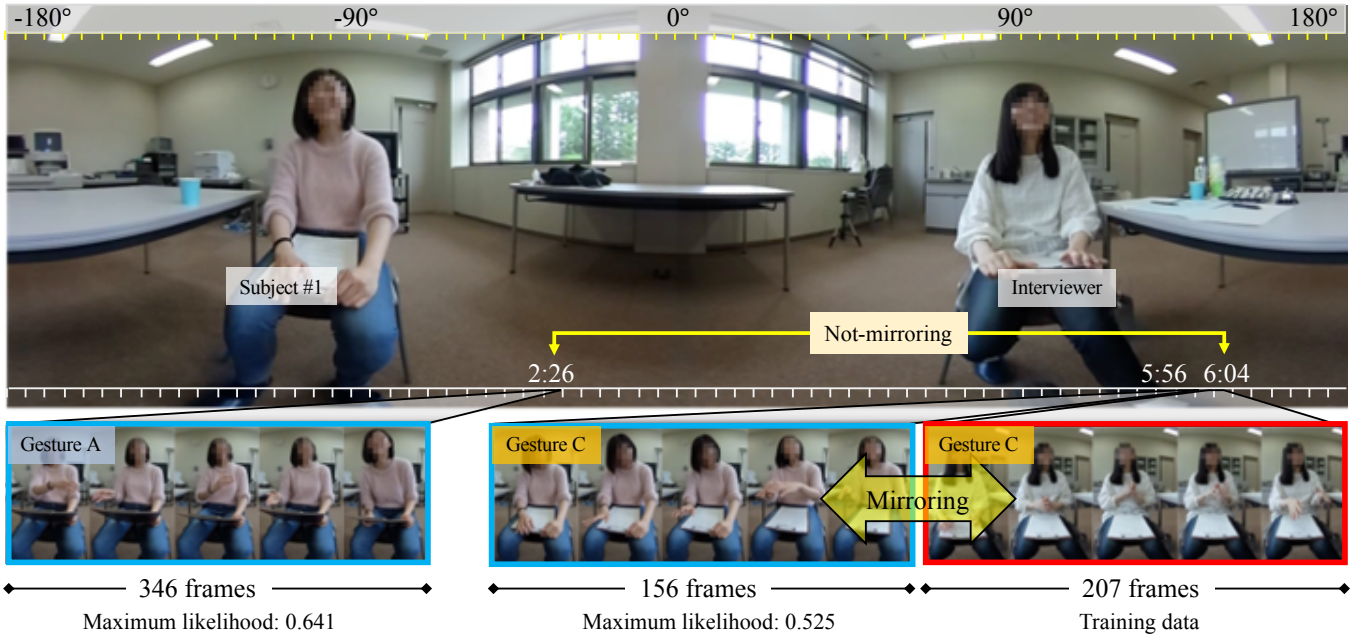

Figure 4. An example of weak gestures that failed to classify.

do not contain body movements from the start to the end of the training.

D. Detection

We apply maximum likelihood from the warping distance to estimate the similarity between the training data created from a participant and the hand movement data of another participant in a conversation. Here, we present a graphical representation to determine the maximum likelihood threshold. Finally, we use the threshold in order to detect gestures that resemble the training data.

IV. Experiment and Result

In this study, two experiments were conducted, involving two subjects and an interviewer. All participants in the experiments agreed to participate and signed the consent forms to allow their data to be used in publications of this research. Each subject was interviewed face to face, at which

 ISBN: 978-1-61208-761-0

TABLE I. MAXIMUM LIKELIHOOD BETWEEN GESTURES OF THE SUBJECT AND THE INTERVIEWER FOR EXPERIMENT 1.

No.	Subject's Gesture	Interviewer's Gesture	Maximum Likelihood
1.	A	A	0.563
2.	B	B	0.688
3.	C	C	0.660
4.	D	D	0.558

Accuracy: 100%

TABLE II. MAXIMUM LIKELIHOOD BETWEEN GESTURES OF THE SUBJECT AND THE INTERVIEWER FOR EXPERIMENT 2.

No.	Subject's Gesture	Interviewer's Gesture	Maximum Likelihood
1.	A	C	0.641
2.	B	B	0.582
3.	C	C	0.525

Accuracy: 67%

point the interviewer randomly imitated the subject's gestures approximately five seconds after recognizing the subject's gesture. Subjects had not previously been informed that the interviewer imitated his or her gestures during the interview. We used RICOH THETA S to record the interview scenes in 1,920×1080 pixels panoramic image frames. Skeletal information was obtained from each frame by using the OpenPose framework [5].

A. Experiment 1

The interviewer imitated four subject's hand gestures: A, B, C, and D. This data was trained using the GRT [10] to predict the corresponding gestures performed by the subject. Table I shows the maximum likelihood between the gestures of the subject and the interviewer. All gestures performed by the interviewer have the maximum likelihood with the corresponding subject's gesture. A successful case of similar gestures detected in this experiment is shown in Figure 3.

B. Experiment 2

During the interview, the subject was excitedly speaking and performing subtle gestures. The interviewer imitated three of them (A, B, and C) and trained this data using the GRT [10]. However, gesture A performed by the interviewer did not resemble that of the subject. Table II shows the maximum likelihood between the gestures of the subject and the interviewer. Gesture A of the subject was determined as Gesture C performed by the interviewer, as shown in Figure 4.

From the above results, we have shown that our proposed framework for detecting the presence of mirroring motion in hand gestures yielded promising results. Some subtle gestures, however, were difficult to classify. In our experiment, we did not make a detailed analysis to optimize the warping window size for the DTW calculation [11]. This issue could be addressed in future research to improve the results.

V. CONCLUSION

In this study, we have proposed a framework to determine whether communication mirroring has been established from the recorded video scenes. Our experiments show that the DTW was able to detect mirroring acts having distinct gestures. The proposed framework will provide a new indication for developing an integrated behavioral analysis system that will enable the assessment of communication mirroring.

ACKNOWLEDGMENT

We thank the faculty of Social Welfare, Iwate Prefectural University, Japan, for funding this project.

REFERENCES

[1] Z. Jiang-Yuan and G. Wei, "Who Is Controlling the Interaction? The Effect of Nonverbal Mirroring on Teacher-Student Rapport," US-China Education Review, A(7), pp. 662–669, 2012.

[2] D. Jayagopi, D. Sanchez-Cortes, K. Otsuka, J. Yamato, and D. Gatica-Perez, "Linking speaking and looking behavior patterns with group composition, perception, and performance," ICMI'12 - Proceedings of the ACM International Conference on Multimodal Interaction, pp. 433–440, 2012.

[3] Behavior coding system, DKH Co. Ltd., https://www.dkh.co.jp/product/behavior_coding_system/ [retrieved: February 20, 2020]

[4] K. Otsuka and S. Araki, "Audio-visual technology for conversation scene analysis," NTT Technical Review, 7(2), pp. 1-9, 2009.

[5] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," Computer Vision and Pattern Recognition, pp. 1-9, 2017.

[6] R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense Human Pose Estimation In The Wild," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7297–7306, 2018.

[7] D. Scaramuzza, "Omnidirectional camera," In Encyclopedia of Computer Vision, 2012.

[8] Y. Jaana, O. D. A. Prima, T. Imabuchi, H. Ito, and K. Hosogoe, "The development of automated behavior analysis Software," Proc. SPIE 9443, Sixth International Conference on Graphic and Image Processing (ICGIP), pp. 1-5, 2014.

[9] P. Schneider, R. Memmesheimer, I. Kramer, and D. Paulus, "Gesture Recognition in RGB Videos Using Human Body Keypoints and Dynamic Time Warping," Lecture Notes in Computer Science, vol. 11531, pp. 281–293, 2019.

[10] N. Gillian and J. A. Paradiso, "The gesture recognition toolkit," Journal of Machine Learning Research, 15, pp. 3483–3487, 2014.

[11] C. A. Ratanamahatana and E. Keogh, "Everything you know about Dynamic Time Warping is Wrong," In Proceedings of the 3rd Workshop on Mining Temporal and Sequential Data, pp. 1–11, 2004.

 ISBN: 978-1-61208-761-0