Pointing, Pairing and Grouping Gesture Recognition in Virtual Reality

Author(s):
Gorobets, Valentina; Merkle, Cecily; Kunz, Andreas

Publication date:
2022

Permanent link:
https://doi.org/10.3929/ethz-b-000556099

Originally published in:
Lecture Notes in Computer Science 13341, https://doi.org/10.1007/978-3-031-08648-9_36

Pointing, Pairing and Grouping Gesture Recognition in Virtual Reality

Valentina Gorobets(✉), Cecily Merkle, and Andreas Kunz

Swiss Federal Institute of Technology, Zurich, Switzerland
{gorobets,kunz}@iwf.mavt.ethz.ch
https://www.icvr.ethz.ch/index_EN

Abstract. During a team discussion, participants frequently perform pointing, pairing, or grouping gestures on artifacts on a whiteboard. While the content of the whiteboard is accessible to the blind and visually impaired people, the referring deictic gestures are not. This paper thus introduces an improved algorithm to detect such gestures and to classify them. Since deictic gestures such as pointing, pairing and grouping are performed by sighted users only, we used a VR environment for the development of the gesture recognition algorithm and for the subsequent user studies.

Keywords: Virtual reality · Gesture recognition · Non-verbal communication · Deictic gestures · Integration

1 Introduction

Lively discussions among people heavily rely on non-verbal communication, such as deictic gestures, facial expressions, body poses etc. Following Mehrabian and Ferris [11], such non-verbal communication could make up to 55% of the overall information exchange. Among these non-verbal communication elements, deictic gestures become particularly important when referring to common artifacts in a team meeting such as a whiteboard. Deictic gestures are used to support information exchange and are intuitively performed and understood by sighted people. However, blind and visually impaired people (BVIP) can not access these deictic gestures, and thus there is a need for detection and interpretation as stated by Kane et al. [7]. While detecting and interpreting gestures in an easy task for human being, machines need sophisticated algorithms to reliably detect gestures and to avoid erroneous output to the BVIP.

This paper thus introduces an improved algorithm for gesture detection. For this, the paper is structured as follows: Sect. 2 introduced previous work in this field, followed by Sect. 3, in which our algorithm is explained more in detail, together with the technical setup and a description of our user study. The remainder of the paper gives a statistical evaluation in Sect. 4, before we discuss the results in Sect. 5. Finally, we conclude our paper with a summary and an outlook on future work in Sect. 6.

This work was commonly funded by DFG, FWF, and SNF under No. 211500647.

K. Miesenberger et al. (Eds.): ICCHP-AAATE 2022, LNCS 13341, pp. 313–320, 2022.
https://doi.org/10.1007/978-3-031-08648-9_36

2 Related Work

Tracking and interpreting deictic gestures has been researched since many years. Research on gestures can be basically divided into 2D and 3D gestures with regard to the interaction space. A comprehensive overview on gestures is given by van den Hoven and and Mazalek [6]. Gesture detection is researched in various application fields. Hofemann et al. [5] for instance use pointing gestures to instruct a robot which part to pick from a table, while pointing gestures were studied in a student-tutor relationship by Sathayanarayana et al. [14]. Besides possible applications of detecting pointing gestures, the pointing accuracy was also intensively studied, e.g., in [1,4].

For detecting hand gestures and interpreting them, recent research employs deep learning algorithms that can detect from incoming video streams. However, training such deep learning networks requires a large data set of annotated gestures, which is a time consuming procedure [13]. Such deep learning networks were again applied to human-robot interaction [12], but also to meeting environments [3] or classroom settings [10,15].

However, there is only little work related to gestures in team meetings, and how to analyze different kinds of gestures to be output to BVIP. A first approach was introduced by Kunz et al. [8] who detected pointing gestures on artifacts on a horizontal workspace. Later, pointing gestures on artifacts on a vertical screen were detected using an Microsoft Kinect depth cam [2], and recently an HTC Vive tracking system was used to track hand gestures and to distinguish them into pointing, pairing, and grouping as the most relevant ones for referring to objects (cards) on a whiteboard [9].

3 Methodology

Our work builds upon the work from Liechti et al. [9] who detected and distinguished three different gestures on a whiteboard: pointing, pairing, and grouping. While this work proved that a distinction of deictic gestures based on tracking signals is in general possible, the accuracy was rather low. This was due to wrongly detected artifacts by the algorithm, position dependency of the pointing person with regard to the whiteboard, and a virtual pointing ray defined by the user's forearm.

Although the envisioned application of our algorithm is to detect deictic gestures by sighted persons in a team meeting together with visually impaired people, we completed the optimization of the algorithm in a Virtual Environment (VE). A VE allows for more replicable conditions in the user study (see also Sect. 3.3).

3.1 Technical Setup

Our technical setup consists of the HTC Vive Pro head-mounted display (HMD) and two HTC Vive trackers. The HMD is used for displaying the VE only, while

its tracking capabilities were not used. Instead, a tracker was attached to the user's head, while the other was attached to the user's wrist. Unlike in [9], the virtual pointing ray is thus calculated from the user's eye position (the top of the head minus an offset) and the user's hand, which eventually is the more precise approach.

3.2 Detection Algorithms

The performed gestures by the user will generate a virtual pointing ray that invisibly intersects with the virtual whiteboard. Thus an invisible trajectory is drawn on the whiteboard. The trajectory consists of points that are generated with a certain spawn rate while pointing, and then interconnected with straight lines. This trajectory will be evaluated by our algorithm and then categorized as "pairing" (indicate a connection between two cards), "grouping" (indicating a cluster of multiple cards), or "pointing" (indicating a single card) (see also Fig. 1).

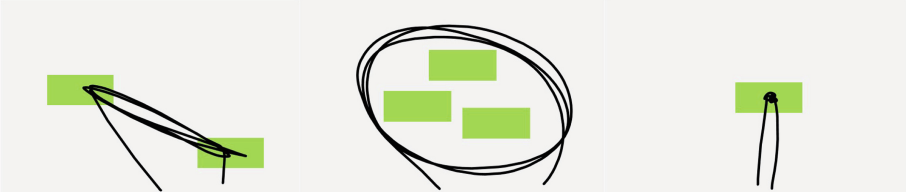
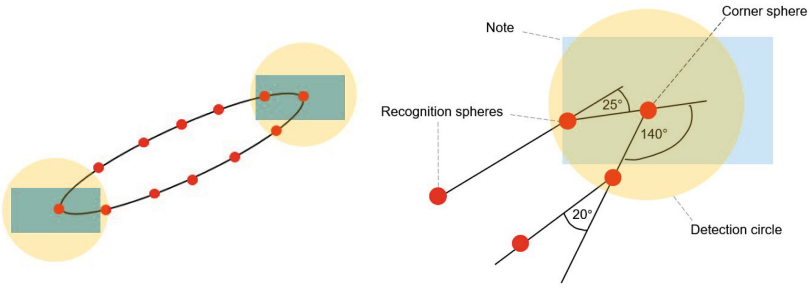


Fig. 1. Pairing, grouping and pointing gestures gestures to be detected.

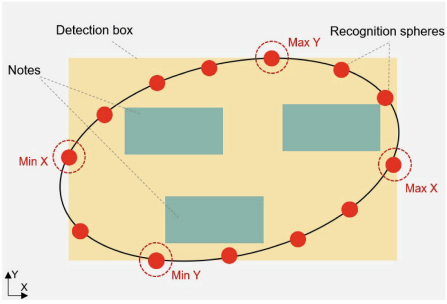
Detection of Pairing Gesture. A pairing gesture is assumed to have an elliptic-shaped trajectory as shown in Fig. 1 (left) and that the regions close to the elliptic focal points determine the two cards to be paired (see Fig. 2a (left)). In order to determine these regions of the ellipse, the curvature used. The ellipse is discretized by the so-called "recognition spheres", that are generated by a given spawn rate while the user is gesturing (see Fig. 2a (right)). Subsequent spheres are interconnected by straight lines, and the angle between two subsequent lines is determined. The two lines that include the maximum angle are defined by in total three points, from which they one in common. This point is the "corner sphere" and supposed to be on a card. The radius of this sphere is then virtually increased to cope with imprecise pointing of the user. Thus, the detection of pairing gestures can be tuned by two parameters: the threshold for the angular change, and the radius of the corner sphere. The values were empirically determined in a separate pilot study to be $\alpha = 100°$ and $r = 0.3\,\text{m}$.

Detection of Grouping Gestures. In case α is always below the set threshold, then a grouping gesture is assumed. For the grouping gesture detection, all sphere

positions are recorded, and the extreme values of the x- and y-coordinates define four coordinates of a boundary box, in which the grouped notes re-assumed (see Fig. 2b).



(a) Pairing gesture to be detected.



(b) Grouping gesture to be detected.

Fig. 2. Detection of pairing and grouping gestures.

Detection of Pointing Gestures. The pointing gesture on a note is simply detected in a temporal manner. If a note is hit by subsequent detection spheres for more than 2 s, it will be detected as pointing.

3.3 Virtual Environment (VE) Design

For creating the VE and interactions within it, we used Unity game engine. Three different zones (see Fig. 3) represent three distance classes in relation to the whiteboard. We wanted to investigate if the distance or the position of the user affects the recognition algorithm and if it does, then in which manner. The top-down view of the VE illustrates the distances from the center of the whiteboard to the center of each zone: the center of the cyan zone is located 2.5 m from the whiteboard, the center of the yellow zone is 1 m away, the center of the magenta zone is 3.3 m. The yellow zone is used to perform gestures next

to the whiteboard, the cyan zone represents further distance, and the magenta zone was used to test gestures that are performed from the side. The virtual whiteboard is represented by the user interface proposed by [8] in our VE.

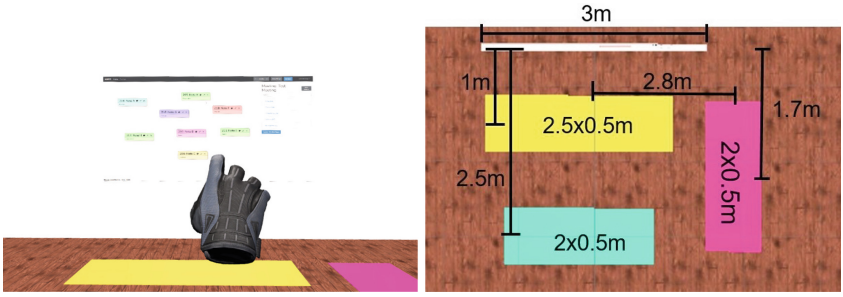


Fig. 3. User's perspective in the HMD and top-down view on the distance zones. (Color figure online)

3.4 Experiment Design

To test and compare the improvements of our algorithm compared to [9], we designed our experiment in the following manner. First, we gave a short introduction to the participants and collected the data obtained from the initial questionnaires. Second, they were asked to perform a particular gesture, following the oral instructions given by the experimenter. Therefore, we consider those instructions as the ground truth for the recognition algorithm.

Participants. We conducted the user study with 30 right-handed participants from 19 to 31 years old. There were 4 females and 26 males with normal or corrected-to-normal vision. We didn't include blind and visually people because the intention of this user study is to test the recognition algorithm we proposed, which can be done with the sighted people only.

4 Results

There are three possible outcomes: the algorithm recognized the gesture correctly, incorrectly or it didn't recognize anything while the gesture was performed.

4.1 Overall Recognition Ratio

Each participant performs 45 gestures during the experiment: 15 pointing, 15 pairing, and 15 grouping gestures. This gives a total number of 450 gestures of each type. Therefore, the total number of the performed gestures among all 30 participants is 1350. Our algorithm correctly recognized 44% of all gestures, 30% of the gestures were not recognized, and 26% were recognized wrongly. The results were then compared to the ones from [9] (see Fig. 4).

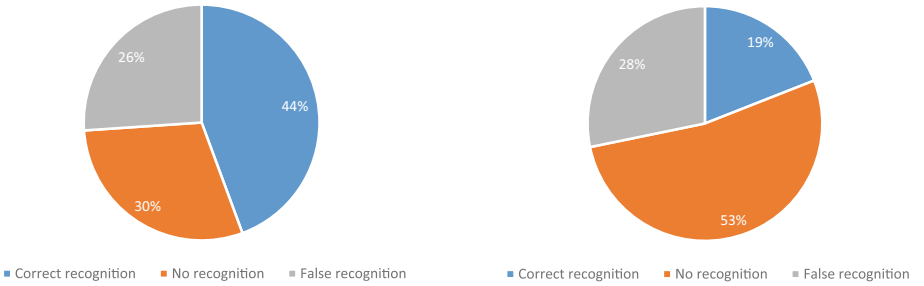


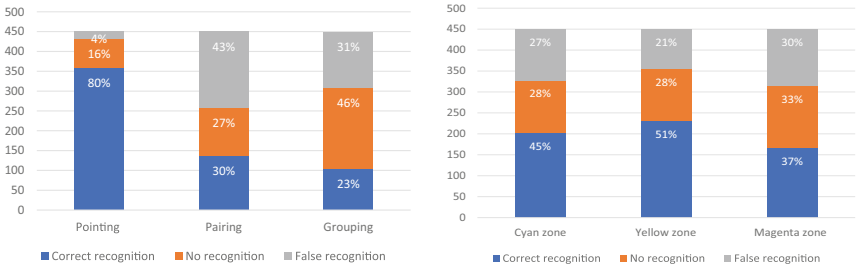
Fig. 4. Side by side comparison of the recognition ratio obtained with our algorithm (on the left) and the algorithm (on the right) from [9].

4.2 Recognition Ratio Based on the Gesture Type

As some of the gestures are easier to recognize than others, we decided to investigate how the recognition ratio depends on the gesture type. We assumed that the pointing gesture will have the highest ratio as it is performed with only one note. Figure 5a gives an overview of the recognition ratio for pointing, pairing and grouping gestures.

4.3 Recognition Ratio Based on the Distance Zones

As it was discussed in Sect. 3.3, we tested our algorithm in three different zones to study if the user position affects the recognition algorithm. The results are presented on Fig. 5b. As it can be observed from Fig. 5b, gesturing from the side has the lowest recognition rate of 37% among all three zones. Zones in which participants stand in front of the whiteboard show better recognition results: 45% when the gestures are performed from a longer distance and 51% when participants stand right next to the whiteboard.



(a) Recognition ratio based on the gesture type.

(b) Recognition ratio based on the distance zones.

Fig. 5. Recognition ratio for various study conditions.

5 Discussion

False recognition for pairing and grouping gestures (see Fig. 5a) was usually caused by the inclusion of additional notes that were close to the notes involved in the performed gesture. Another reason for the low recognition rate for the pairing gesture is caused by the user behavior. We assumed that during pairing two notes participants will do it repeatedly. However, during the experiment we observed that some participants were pairing two notes by pointing at them successively. In this case, pairing gesture was either not recognized, or recognized false. Pointing was also not recognized if pointing gestures were performed very fast, since the time threshold for the start of the pointing gesture recognition was not reached.

Our algorithm performs worse for the gesturing from the side. However, it could also be caused by the inaccurate gesturing of the participants. Due to their position, it is more difficult to "aim" and perform an accurate gesture when they face the whiteboard from the side.

6 Summary and Outlook

In this paper, we described an approach for the pointing, pairing and grouping gestures recognition. We investigated the overall recognition ratio for all performed gestures, and how it changes depending on the gesture type and the position to the whiteboard.

For future work, we will decouple the results of the recognition algorithm from the performance of human gesturing. Such decoupling will allow us to test the initial accuracy of the algorithm without considering differences caused by different gesturing behavior. The next steps will also include transferring this algorithm to the real life scenario by creating a virtual twin of the real setup and using only two trackers without the HMD to recognise the performed gestures. Next, an interface to communicate gesturing information to blind and visually impaired people will be implemented.

References

1. Akkil, D., Isokoski, P.: Accuracy of interpreting pointing gestures in egocentric view. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, September 2016. https://doi.org/10.1145/2971648.2971687
2. Dhingra, N., Valli, E., Kunz, A.: Recognition and localisation of pointing gestures using a RGB-D camera. In: Stephanidis, C., Antona, M. (eds.) HCII 2020. CCIS, vol. 1224, pp. 205–212. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50726-8_27
3. Hassink, N., Schopman, M.: Gesture recognition in a meeting environment. Master's thesis, University of Twente (2006)

4. Herbort, O., Krause, L.-M., Kunde, W.: Perspective determines the production and interpretation of pointing gestures. Psychon. Bull. Rev. **28**(2), 641–648 (2020). https://doi.org/10.3758/s13423-020-01823-7
5. Hofemann, N., Fritsch, J., Sagerer, G.: Recognition of deictic gestures with context. In: Rasmussen, C.E., Bülthoff, H.H., Schölkopf, B., Giese, M.A. (eds.) DAGM 2004. LNCS, vol. 3175, pp. 334–341. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-28649-3_41
6. van den Hoven, E., Mazalek, A.: Grasping gestures: gesturing with physical artifacts. AI EDAM **25**(3), 255–271 (2011)
7. Kane, S.K., Wobbrock, J.O., Ladner, R.E.: Usable gestures for blind people: understanding preference and performance. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 413–422. ACM, New York (2011). https://doi.org/10.1145/1978942.1979001
8. Kunz, A., Alavi, A., Sinn, P.: Integrating pointing gesture detection for enhancing brainstorming meetings using Kinect and pixelsense. Procedia CIRP **25**, 205–212 (2014)
9. Liechti, S., Dhingra, N., Kunz, A.: Detection and localisation of pointing, pairing and grouping gestures for brainstorming meeting applications. In: Stephanidis, C., Antona, M., Ntoa, S. (eds.) HCII 2021. CCIS, vol. 1420, pp. 22–29. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78642-7_4
10. Liu, T., Chen, Z., Wang, X.: Automatic instructional pointing gesture recognition by machine learning in the intelligent learning environment. In: Proceedings of the 2019 4th International Conference on Distance Education and Learning, pp. 153–157 (2019)
11. Mehrabian, A., Ferris, S.R.: Inference of attitudes from nonverbal communication in two channels. J. Consult. Psychol. **31**(3), 248–252 (1967). https://doi.org/10.1037/h0024648
12. Pizzuto, G., Cangelosi, A.: Exploring deep models for comprehension of deictic gesture-word combinations in cognitive robotics. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE (2019)
13. Ripperda, J., Drijvers, L., Holler, J.: Speeding up the detection of non-iconic and iconic gestures (SPUDNIG): a toolkit for the automatic detection of hand movements and gestures in video data. Behav. Res. Methods **52**(4), 1783–1794 (2020). https://doi.org/10.3758/s13428-020-01350-2
14. Sathayanarayana, S., et al.: Towards automated understanding of student-tutor interactions using visual deictic gestures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 474–481 (2014)
15. Wang, J., Liu, T., Wang, X.: Human hand gesture recognition with convolutional neural networks for k-12 double-teachers instruction mode classroom. Infrared Phys. Technol. **111**, 103464 (2020)