·Article·

Gesture-based target acquisition in virtual and augmented reality

Yukang YAN[1,2], Xin YI[1,2*], Chun YU[1,2,3,4], Yuanchun SHI[1,2,3,4]

1. *Department of Computer Science and Technology, Tsinghua University, Beijing* 100084*, China*
2. *Beijing Key Lab of Networked Multimedia, Beijing* 100084*, China*
3. *Key Laboratory of Pervasive Computing, Ministry of Education, Beijing* 100084*, China*
4. *Global Innovation eXchange Institute, Tsinghua University, Beijing* 100084*, China*

*** Corresponding author,** yixin@tsinghua.edu.cn

Received: 25 January 2019 **Accepted:** 18 April 2019

Supported by the Natural Science Foundation of China (61521002, 61572276, 61672314); the National Science Foundation Graduate Research Fellowship (DGE−1256082); the National Social Science Fund of China (15CG147); Beijing Key Lab of Networked Multimedia.

Citation: Yukang YAN, Xin YI, Chun YU, Yuanchun SHI. Gesture-based target acquisition in virtual and augmented reality. Virtual Reality & Intelligent Hardware, 2019, 1(3): 276—289
DOI: 10.3724/SP.J.2096-5796.2019.0007

Abstract **Background** Gesture is a basic interaction channel that is frequently used by humans to communicate in daily life. In this paper, we explore to use gesture-based approaches for target acquisition in virtual and augmented reality. A typical process of gesture-based target acquisition is: when a user intends to acquire a target, she performs a gesture with her hands, head or other parts of the body, the computer senses and recognizes the gesture and infers the most possible target. **Methods** We build *mental model* and *behavior model* of the user to study two key parts of the interaction process. *Mental model* describes how user thinks up a gesture for acquiring a target, and can be the intuitive mapping between gestures and targets. *Behavior model* describes how user moves the body parts to perform the gestures, and the relationship between the gesture that user intends to perform and signals that computer senses. **Results** In this paper, we present and discuss three pieces of research that focus on the mental model and behavior model of gesture-based target acquisition in VR and AR. **Conclusions** We show that leveraging these two models, interaction experience and performance can be improved in VR and AR environments.

Keywords Gesture-based interaction; Mental model, Behavior model; Virtual reality; Augmented reality

1 Introduction

Virtual reality is growing to be an important platform for various types of scenarios (e.g., games[1], training[2] and education[3]), where target selection is a common and basic interaction task. For example, a game player often picks up game props, a mechanic might need several tools (e.g., a hammer). Currently, target selection tasks require users to switch to the menu interface and select the target out of a list of candidates. This method requires a series of manipulations including invoking the menu, choosing the target category, scanning the items in the menus until the user pinpoint the desired one, and finally getting back to the ongoing task. This interaction process could be time-consuming and distracting, especially when users are new to the interface or when the target is buried deeply in a hierarchical menu[4].

Compared to selecting targets through menus, gesture-based interaction can be a potential solution to simplify the searching process. Gesture-based approach has the advantage of enabling eyes-free and direct input[5] and does not require an additional interface. Eyes-free input refers to the interaction that does not require the participation of users' visual focus on the input devices, e.g., typing on the keyboards to input words while looking at the screen. In this paper, we explore to design intuitive hand gestures to retrieve virtual objects, head-based gestures to trigger control commands, and using eyes-free pointing gestures to acquire targets in the interaction space around the body. A key to gesture-based target selection is designing the mapping between the gestures and the targets, which greatly influence the learnability of the gestures. An ideal mapping should satisfy several criteria: it should be easy to discover and memorize[6,7], be easy to perform, be consistent with the acquired experience of users[8], gains high consensus across users[9], and be associated with high productivity and low error. To meet these criteria, we study how users build connections between gestures and targets, which we refer to as users' *mental model*. Based on the results, we leverage users' acquired interaction experience and sense of proprioception in the gesture-target mapping design.

Related researches on gesture mapping design applied two major approaches. One way is to specially design the appearance of the target to suggest the assigned gesture. The gestures were mapped to the shapes, colors, motions of the targets, or directly overlaid onto the targets[8,10–13]. In these cases, the gestures were cued by the appearance of the targets, and thus are easier to discover and remember[13]. However, using these techniques, users need to observe the targets to find the gestures, which introduces a high visual load. The second but more widely used solution is the user-defined approach, which was first introduced by Wobbrock et al.[9] to design gestures for interaction on an interactive surface. This approach portrays the effect of a command (e.g., to delete an item), and then asks a group of users to design their own gestures to issue this command. The gesture with the highest consensus will be assigned to the command. In this way, the elicited command-gesture mappings reflect daily behaviors and experience of users, which results in a more contextual connection between gestures and commands[9]. The approach has been successfully applied to many areas[14–17]. In this paper, we follow the approach of user-defined gesture to elicit gesture-target mappings that users feel most intuitive.

In addition to intuitive mappings, the recognition of the input gestures is another challenge to gesture-based target selection approaches. In the process of performing gestures, user controls her body parts to mimic the gesture in her mind. However, limited by the motor control accuracy, the performed gesture is usually not exactly the same as the desired one. Meanwhile, the system senses the performed gesture through several information channels (e.g., camera and initial sensors). The sensing devices also introduce noise and errors due to limited sensing accuracy. What we need to achieve is to detect and recognize the user's intended gesture in spite of these noises and errors, and accurately return the target that the user intends to select. We refer this part to be the understanding of user's *behavior model*. To achieve this goal, we sample target positions in the interaction space around users, collect selecting data and regress the position offset to help predict the desired target of the user.

In the process of performing gestures, user's control accuracy relies on a number of factors, and two important factors among them are spatial memory and proprioception[18]. Spatial memory is the part of memory that is responsible for recording information about different locations and the spatial relations between objects[19]. It can help users efficiently retrieve positions of targets[20] in acquisition tasks. Previous work studied the ability and effectiveness of users to build the spatial memory, both in 2D[21] and 3D[20] spaces. In addition, proprioceptive feedback is important for human's movement control[18]. Proprioception is the sense of position and orientation of one's body parts with respect to each other[22]. With the help of proprioception, users could perform eyes-free acquisitions of the targets on various platforms. For example, Face-Touch[23] visualizes targets in the virtual world via a VR headset, and enables users to select them by tapping onto the back of the headset at the according position to the target in their view. Although users cannot see their hands or the target location on the headset in the real world, they can still estimate the target location and reach to the vicinity of it without aiming. Similarly, with the help of proprioception, users can perform target selection on a remote screen (Air Pointing[18]), select different directions by orienting a mobile device (VirtualShelves[22,24]), or control the body posture as an input modality (Pose-IO[25], FootGesture[26]).

The process of gesture-based target acquisition is as followed: User has a desired target as the intention while she may have other ongoing tasks. Considering the intention, tasks and the current environment as the context, user will decide a gesture to acquire the target. By controlling her hand, head or other body parts, the user performs the gesture. The computer senses the gesture in different ways, including using visual-based, inertial sensor-based sensing techniques. Limited by the accuracy of user's movement control and the sensing accuracy, the sensed gesture data is not exactly the intended gesture in user's mind. In most cases, it can be represented as a set of gesture candidates with different estimated possibility to match the intention. Within this set, computer will extract temporal, spatial and frequency features and run algorithm to recognize the original performed gesture and user's intention. As with many intelligent input algorithms, the input prediction model that estimates the possibility of different candidate gestures should be trained based on user data. As we will show in this paper, the mental model and behavior model of different users share some similar characteristics, but are distinct from each other in other aspects. Therefore, to achieve reliable performance in real use, the algorithm should not only consider patterns emerged from the data of a number of different users, but also be able to continously adapt to each individual user during usage. In practice, the training process demands a varying size of training data, depending on the specified error tolerance, the signal-noise ratio and the tasks itself. For example, as we will show in Section 2.1, using data from 12 participants and a list of top-5 candidates, users could reach an accuracy over 94% when performing object retrieval by grasping gesture. Through this process, the two models we described above play very important role. *Mental model* describes how users choose a gesture given intention, task and context. *Behavior model* reveals the connection between the performed gesture and the sensed gesture. Leveraging these models, we can improve the understanding of the user's intention and then provide the intended target more efficiently and more accurately.

Based on the two models, we developed three novel target selection approaches in VR and AR. First, we designed intuitive grasping gestures to retrieve virtual objects in VR through a gesture elicitation experiment[27]. Evaluation results showed that novice user successfully retrieve targets with accuracy of 75.51% without any training. Second, we designed head movement based gesture to trigger command on AR devices. Through participatory design process, we generated nine head gestures and assigned them to trigger basic commands (e.g., select, drag and drop). This approach supported controlling the device in a hands-free way. Third, we studied how user acquire targets in the interaction space around the body without turning head to look at them[28]. By analyzing the distribution of acquisition points, we generated the connection between the acquisition points and the desired target positions and then improved the acquisition accuracy.

2 Materials and method

To explore users' mental model of gesture mapping, we applied participatory design process in two cases:

(1) eliciting grasping gestures of objects for object retrieval tasks in VR; (2) eliciting head movement based gestures to support hands-free control of AR devices. To study users' behavior model of performing gestures, we analyzed users' target acquisition behaviors and model the connection between their acquisition points and desired target positions.

2.1 Object retrieval by grasping gestures in VR

In reality, the gestures that we used to grasp or manipulate different objects are adapted to their different shapes, sizes, and usages. For example, to grasp a mug, we often adapt our gestures to be "hook" shaped for its ring-based handler. Based on this observation, we aim to explore how users think up the grasping gestures of virtual objects and whether they can achieve consistent mappings from gestures to the objects. So we conduct this gesture elicitation experiment to probe the consistency and intuitiveness of the grasping gesture mappings.

2.1.1 Gesture elicitation

We recruited 20 participants (14M/6F) from a local campus. They aged from 20 to 27 (AVG=23.6). The task of participants was to recall grasping gestures once given an object name. We used two cameras to record the gestures by taking pictures from the front and side views. Forty-nine different, as listed in Table 1 objects were tested and 980 object-gesture pairs were collected. Authors merged the same object-gesture pairs into 140 distinct object-gesture mappings. We applied two main metrics to measure the consistency of the mappings. One is the number of gestures that were mapped to each object. The results showed that 18/49 objects were mapped to only one unique gesture by all the participants and all objects were mapped to no more than five gestures. The other one was the agreement score proposed by Wobbrock[9]. In our study, the average agreement score was 0.68 (SD=0.27) and 36/49 objects achieved a score of no less than 0.50, which could be regarded as indicators of robust proposals[17]. All these results proved that the mappings of grasping gestures achieved very high consistency across users. Based on the results, we also build taxonomy of the elicited gestures, as shown by Figure 1.

Table 1 The object list for the study, which was divided into six groups, and objects in the same group could appear in the same scenarios

Scenarios	Object lists
Office	book, briefcase, eraser, mouse, keyboard, pen, scissor, stapler
Game Weapons	binocular, bow, dagger, grenade, handgun, rifle, shield, spear, sword
Sports	barbell, basketball, badminton racket, cue, golf club, javelin, ski stick, shot, skipping rope
Electronics	camera, flash drive, headphone, interphone microphone, phone, remote control
Home	bowl, broom, comb, glasses, mug, perfume, toothbrush, umbrella, watch
Food	apple, banana, beer, hamburger, popsicle, watermelon

2.1.2 Evaluation

We used Perception Neuron, a MEMS (Micro-Electro-Mechanical System) to record gesture data of users. We asked users to perform all the gestures and recorded positions of fourteen joints (except for the carpometacarpal joint of the thumb) of five fingers relative to the palm, the position and orientation of two palms for 40 frames for each gesture. In total, we obtained 12 participants×101 gestures (chosen out of 140 gestures)×2 rounds×40 frames = 96960 frames of data. Using these data, we implemented an SVM-based classifier. Leave-Two-Out validation showed the offline classification accuracy was 70.96% (SD=9.25%)

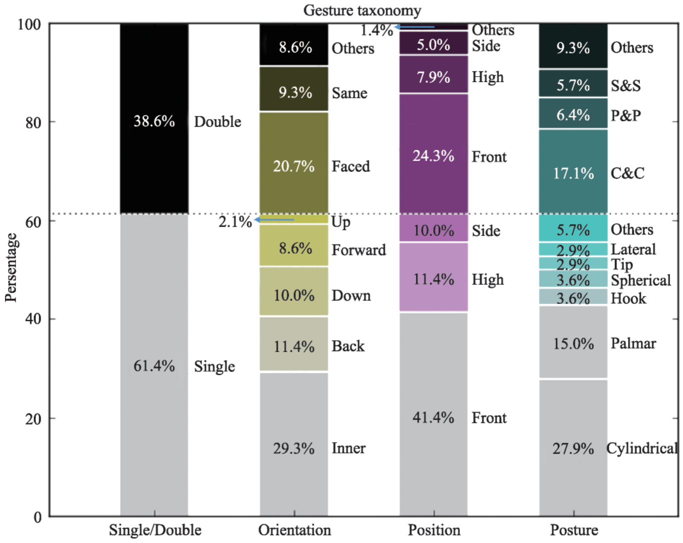

Figure 1　Distribution of the gestures that users designed in this study in four dimensions of the taxonomy. “Single/ Double” refers to whether user performs the gesture with both of the hands or only one of them. “Orientation” refers to the orientation of the hand palm while performing the gesture. “Position” refers to the relative position of the hand to the main body, e.g., on two sides or in the front. “Posture” refers to the hand shape while performing the gesture.

for when the target was exactly most possible candidate, 89.65% (SD=6.39%) for when the target was among top three candidates and 95.05% (SD=4.56%) for top five candidates.

We also tested the actual object retrieval tasks using the obtained gesture classifier. To test the discoverability of the gesture-object mappings, we recruited twelve new participants (8M/4F) who did not participate in the previous experiment. These participants were aged from 21 to 25 (AVG=23.1). The task of participants was to perform a gesture to retrieve the target objects. The tested objects were the same 49 objects listed in Table 1. We implemented the experiment platform using Unity engine, which is shown in Figure 2. Target name, participant's current input gesture and the recommended candidates with the highest likelihood were visualized in this interface. We arranged three sessions in this experiment: discovery, learning and recall sessions. In discovery session, we only inform participants the grasping metaphor and let them recall their own grasping gestures to retrieve the target objects. In learning session, we show participants the gestures that were supported by the system and they took time to practice to learn the gestures and then performed them in the retrieval tasks. The gestures were exactly the 101 chosen gestures in *Gesture Elicitation*. For the objects with more than one assigned gestures, we showed all the gestures that can be used to retrieve them. In recall session, they came back to the lab after a week and perform the retrieval tasks again. On average, the discovery and recall session took 15min and the learning session took 30min to complete. Results showed that in discovery session, participants could discover the exact gestures for 40% of the target objects. In learning and recall sessions, given a list of five most possible candidates, participants could successfully retrieve 94.50% and 93.20% of the objects using our approach. Participants also commented that they thought this approach to be interesting and intuitive.

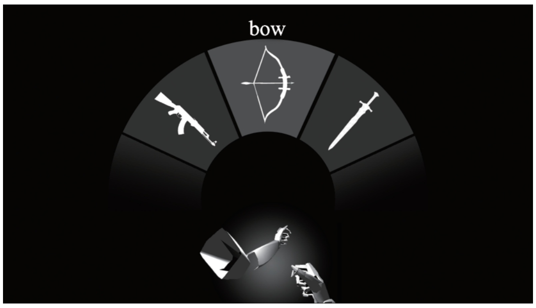

Figure 2　The user interface of the experiment, showing the task (bow), object candidates (bow, rifle, and sword) and current gesture (the visualization of arms and hand).

2.2　Command selection by head gestures in AR

Current state-of-the-art head mounted AR devices (e.g., Hololens) mostly require users to trigger commands (e.g., select, main menu) by mid-air hand gestures. However, there are a number of situations where users' hands are occupied, e.g., while writing notes with a pen in the hand. In these cases, we explored to use head movement based gestures to control the AR devices in a hands-free way. However, different to hand gesture, users are less familiar with performing head gestures and some of the head gestures may be easy to confuse with unintentional head movements. So in this research, our goal was to probe the *mental model* of users about how they will design the head gestures to be intuitive and how they will avoid the confusions between head gestures and unintentional head movements from their own perspective. To address these challenges, we went through gesture exploration and design processes. As a result, we generated a set of nine intuitive and distinguishable head gestures to trigger basic control commands on HMD AR devices.

2.2.1 Gesture space exploration

Previous research has explored to use head movements as a human-computer interaction channel[29], including wheelchair control[30,31] for users with limited hand or arm mobility and target selection tasks on desktop[32,33] and mobile devices[34] for able-bodied users. However, they required users to perform pre-defined head movements to trigger different functions. In this research, we first explore the whole usable gesture space and then elicited intuitive gesture-command mappings from end users.

In the exploration process, we recruited sixteen participants (12M/4F) from a local campus. Their average age was 24.44 (SD=1.90). The task of participants was to propose usable head gestures that met two design goals, which were intuitive to perform and distinguishable from unintentional head movements. We showed a cursor and its recent trajectory of 500ms to help them observe the amplitude and direction of the head rotations and movements. In total, we collected 210 head gesture instances. Based on the results, we summarized the gesture taxonomy, which is listed in Table 2. We also elicited design inspiration and strategies for avoiding false positives. The design inspiration included *Act like using hands* and *Transfer daily experience*. Users proposed to perform gestures with their heads as if using their hands, e.g., to raise the head fast towards the upper right corner to mimic throwing objects away with the right hand; and to transfer acquired experience of performing head gestures, e.g., to lean the head to the shoulder when user

Table 2 The head gesture taxonomy that we summarized from the results

Movement	Lower or Raise	Lower or raise the head along x axis
	Tilt	Rotate the head along y axis
	Rotate	Rotate the head along z axis
	Stretch	Stretch the neck and move the head to different directions rotating
	Dwell	Stop the head movement for a short duration
Trajectory	Directional	Move the head to different directions
	Shape	Use head to draw geometrical shapes, e.g., circle
	Character	Use head to write characters or numbers
Flow	Delimiter	To perform a head gesture at the start and the end as the delimiter to switch the mode
	Repetition	Repeat a head gesture for more than one times
	Reverse	Perform a head gesture and then reverse it
Nature	Transfer	Use the head movement to mimic the hand gestures
	Existence	Use the head gestures that already exist, e.g., nodding
	Infrequent	Actions Use the head movements that were rarely performed in daily life
	Large amplitude	Enlarge the amplitude of daily head movements

The dimensions include movement category, movement trajectory, gesture flow and nature of the design. The x, y, z axes are illustrated in Figure 4.

needs a rest. The strategies included *Infrequent actions* (performing the actions that were infrequent in daily life), *Repeat it twice*, *Draw strokes* (e.g., triangles), *Delimiter gestures* and *Forth and back*.

2.2.2 Participatory gesture design

In the design process, we adopted the participatory design and conducted a gesture elicitation experiment with users. We recruited sixteen participants (10M / 6F) for this experiment, with an average age of 25.56 (SD=2.97). The task of participants was to design a head gesture to trigger a target command. We referred to the command set of a state-of-the-art AR headset (Hololens), which included nine basic commands: *Drag*, *Hold*, *Home* (return to the main menu), *Scroll up/down*, *Select*, *Double Tap*, *Zoom in/ out*. For each command, we first showed its effect on the Hololens by recorded screen videos. Participant watched this video through Hololens and after he or she confirmed to understand the effect, sufficient time was given to them to design the related head gesture. In total, we collected 267 head gestures that were reduced to 80 distinct gesture-command pairs after merging. We also applied agreement score[9] to measure the consistency of the mapping. Figure 3 shows the results. The consistency of the mappings were relatively low compared to previous gesture elicita-tion experiments[16,17,35]. This reflected that users had less experience with head gestures. Based on participants' most popular proposals and several refinements, we gene-rated the final gesture set, which is shown in Figure 4.

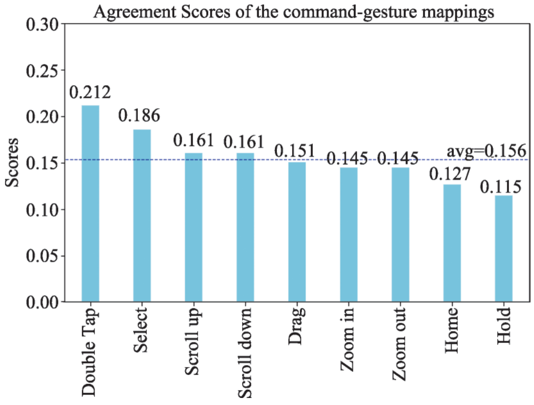

Figure 3 The agreement scores for the head gestures that participants designed for each command.

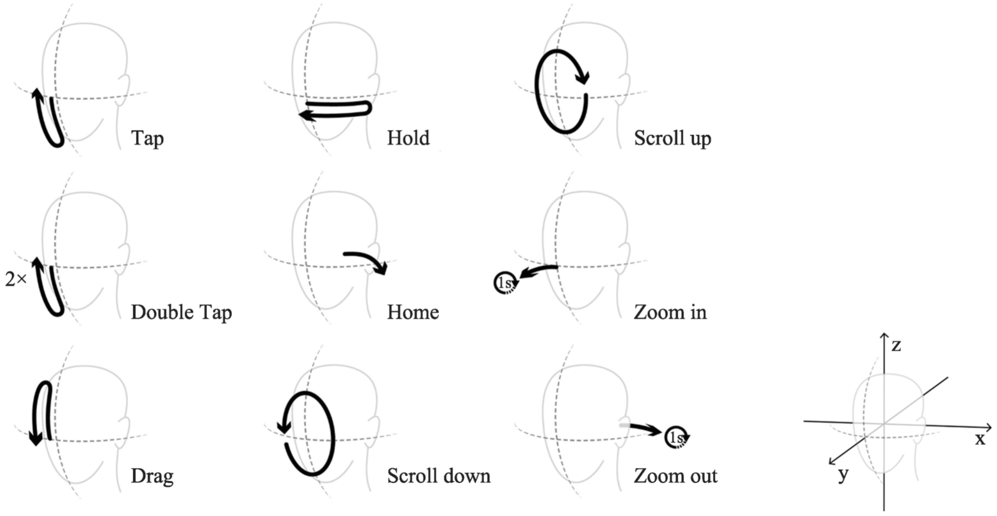

Figure 4 The final design of the HeadGesture set for the nine commands. The movement of head is indicated by the arrows. "2×" represents the repeating of the action for twice; "1s" is an illustration for a dwell.

2.3 Eyes-free target acquisition in VR

Two pieces of research above studied user's *mental model* on how to design intuitive gestures for target selection tasks. This research focus on both user's *mental model* and *behavior model* in a specific situation which is to acquire targets in the interaction space around body in an eyes-free way. Currently, acquiring an object in VR is eyes-engaged: Even the object is near, user has to turn the head to the direction and visually locate the target before acquiring it. However, in physical world, users have the ability to acquire targets without eye's participation[36]. By leveraging the spatial memory and proprioception, people can reach for an object in an eyes-free way (e.g., a driver reaches gear stick while driving). However, during this process, the control accuracy of user will be lower compared to eyes-engaged way and their acquisition points may have significant offsets. So in this research, we sampled target positions to study the acquisition behavior of users while performing the acquiring gestures.

2.3.1 Subjective acceptance

Before testing the control accuracy and speed of the acquisition, we decided to first probe the subjective acceptance of users. The aim was to test the confidence level of target acquisition with different levels of difficulty in an eyes-free test condition. The difficulty is controlled by the distance between objects and the target positions. We recruited twelve participants (8M/4F) from a local campus. They were between the ages of 22 and 26 (AVG=24.2, SD=1.34), with average arm length of 67.95cm (SD=3.22). The task of participants was to acquire the target at different position around the body and at the condition of different distances between the target and its surrounding distracting objects. Participants were required to point at the targets "as accurately as possible" and acquired the target by pointing at the target with the controller and pressing the controller trigger. Sixty positions were evenly sampled in the angular space, which were arranged at twelve levels from −180° to 180° horizontally and five levels from −60° to 60° vertically. The target distance started at 5cm and users could enlarge it 1cm at a time until they felt the distance enough for accurate acquisitions. Figure 5 shows the overall results of the subjective acceptance of eyes-free target acquisition. RM-ANOVA tests showed that the vertical angle and horizontal angle of the target position significantly affected the minimum distance between targets respectively ($F_{4,44}$=5.173, *p*=0.002; $F_{11,121}$= 31.451, *p*<0.001). This was also consistent with subjective feedback of users. "*When I lifted my arm, the jitters limited my accuracy.*" [P1] "*The positions in the rear were very difficult to reach and the postures were uncomfortable.*" [P5]

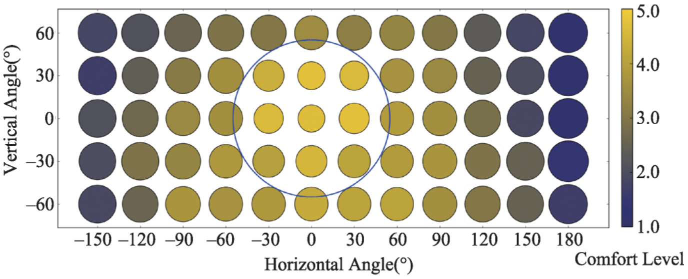

Figure 5 The circles summarize the comfort level (color) and the minimum distance between targets (radius) for different target positions (centers) when users acquired them.

2.3.2 Control accuracy

In this experiment, we study user's *behavior model* of performing gestures to acquire targets at different positions. We recruited 24 participants (20M/4F) from a local campus, aged from 20 to 26 (AVG=23.21, SD=1.64). The task of participants was to rotate to the instructed *direction* and acquire the targets at the informed *position*. The target positions were the same 60 positions as tested in *Subjective Acceptance* experiment. The twelve directions were the twelve horizontal angles selected for the target positions. For each *direction*, 60 target positions were tested. Acquisition tasks in different directions were to test how well users rebuilt their sense of proprioception after rotations. They were given enough time to get familiar with the 60 target positions before the experiment. All 60 targets were shown during practicing, but none is visible in the test sessions. The order of target position and rotation direction was randomized for each participant. The experiment setting is illustrated in Figure 6. In total, we collected 24 participants × 60 positions × 12 directions=17280 trials of target acquisitions. By data processing, we removed 1.12% of the acquisitions which acquisition offsets were out of the range of three times deviation from the averaged values.

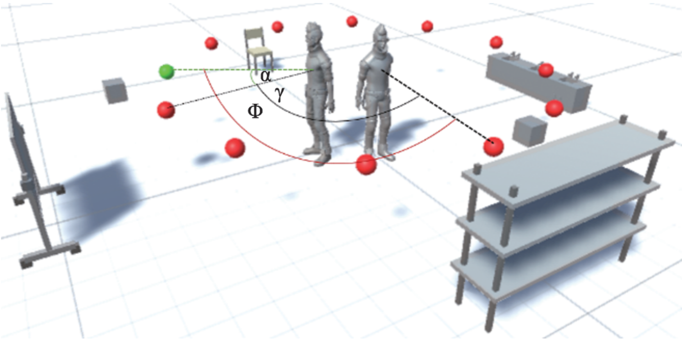

Figure 6 The concept of the experiment settings. The furniture indicates the virtual surroundings of the participants. The red spheres indicate the twelve directions that the participants rotate to, and the green sphere indicates the positioned target.

The overall results is shown by Figure 7. Sixty circles summarize the acquisition points of different target positions, which centers are the average acquisition positions and the radii are their standard deviations. The blue lines show the average offsets of participants' acquisition points from the intended target positions. We run RM-ANOVA to analyze the effects of the target positions (horizontal and vertical angles) and the number of rotations on the acquisition offsets. The results show that horizontal angle and vertical angle of the target position significantly affected the acquisition offset ($F_{11,253}$=95.48, p<0.001; $F_{4,92}$= 22.76, p<0.001). The vertical offsets increased symmetrically as the horizontal angle changed from 0 degree to both sides (180 and −180 degrees). This was because when targets were located to two sides, users need to abduct the shoulder at a large angle to acquire them, which limited the range they could raise them arm to and therefore they reached lower positions. The number of rotations also significantly effected acquisition offsets ($F_{11,253}$=70.46, p<0.001). This result indicates that the ability of participants to reintegrate proprioception decreased with the number of rotations. Post hoc tests showed that the offset of the first acquisition was significantly smaller than the others (all p<0.001), which showed that their reintegration ability dropped most significantly after the first rotation.

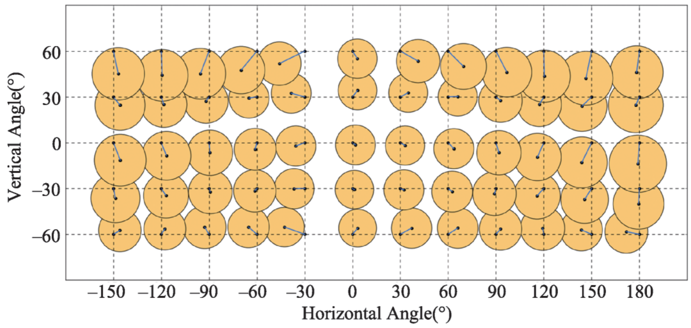

Figure 7 Summary of the main results of this experiment. The centers of the circles are the averaged positions of twelve acquisitions of 24 users and radiuses are the standard deviations. The blue lines visualize the offsets of the average positions from the target's actual positions. All coordinates and lengths are converted to angles in degrees.

Based on the acquisition data of 60 evenly sampled target positions, we interpolated the acquisition offsets and standard deviations of the whole space. Figure 8 shows the interpolation of the standard deviations. As standard deviation reflected the closeness of the acquisition points and the acquisition accuracy of the participants at different target positions, user interface designers could refer to this result to arrange the target (e.g., icon) position for eyes-free acquisitions. We also interpolated the offset model of participants. Using this model, we improved the accuracy of selecting from 60 target positions from 74.99% to 78.17%.

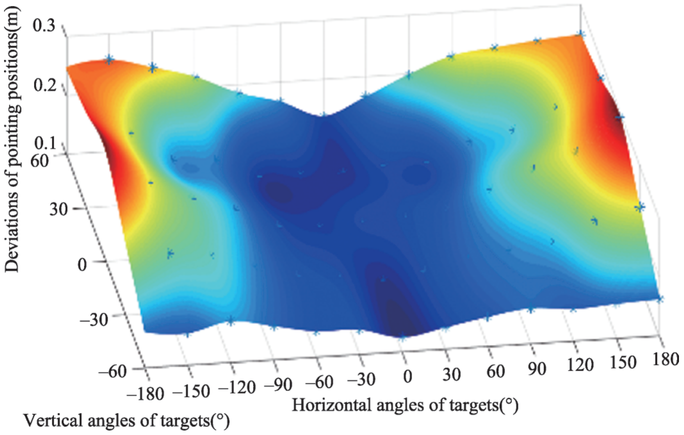

Figure 8 The interpolation of the standard deviations of the target acquisition points on the whole surface, visualized into a heat map.

3 Discussion

3.1 Mental model of gesture design

We studied the mental model of users in three research. Mental model determined how users design and recall gesture-object mappings. This help us elicit the most intuitive mappings which can reduce the learning effort of users and improve the discoverability of the designed gestures. Compared to previous participatory design of gestures[9], we supported many-to-one mapping between gestures and the target objects. This design is based on the finding that for some objects, users' mental models on the corresponding gesture were significantly different. In addition, our results reflected the acceptance level of performing gestures to acquire targets. In the eyes-free manner, users would refuse to use the approach when the acquisition task was too difficult (the target was too close to distracting objects). So it is of great value to probe the mental model, otherwise the designed gestures would never be adopted or discovered by users. However, it is too complicated and involves many factors, including users' acquired experience and personality. So in our research, we focused on the general model of most users and in the future, we can further study it in more detailed level.

3.2 Behavior model of acquiring targets

After users come up with a gesture to acquire the target, he performs it by controlling his body to complete the anticipatory movements. However, as the movements introduce noises and errors, we compute the behavior model to deal with the mapping between the intended targets and the senses gestures. As our research results showed that there are specific patterns of users' behavior changes when targets were located at different positions. Leveraging these patterns, we could either improve the acquisition accuracy or to guide the design of user interface. Compared to related work[18,22–25], this is the first study to test the usability and feasibility of eyes-free target acquisition in virtual reality environment. As we showed, this model was also affected by other factors, including the number of rotations of the user. After several times of rotation, especially the first rotation, users lost their reintegration of the proprioception and spatial sense and finally created larger acquisition offsets. By adding more and more factors into consideration, the behavior model will be more powerful in predicting the intended targets of users.

4 Limitation and future work

We present three studies on the mental model and behavior model of users performing gestures. However, there are several factors that were not evaluated in this research and should be studied in the future. One factor is the modality of the gesture interaction. We tested the gesture interaction with hands and heads independently, but we did not study how combined modalities will affect the mental models of users. As our results showed, when designing head gestures, users will transfer their experience of hand gestures but also created more unique gestures for head movements. This connection and other effects that may be introduced by other modalities (e. g., foot and gaze) will be tested in the future. Another factor was the interaction effect between the mental model and the behavior model. With different mental models, users understand gestures in different ways with different metaphors. Will this affect how well and accurately they perform a gesture? In this research, we studied the models separately, but it would be of value to test interaction effects in the future.

5 Conclusions

In this paper, we present and discuss three pieces of research on the gesture-based target acquisition in VR and AR. We studied two key models, *mental model* and *behavior model* of user, in the interaction process to better understanding user's intention and improves the efficiency and accuracy of the target acquisition tasks. By leveraging these models, we realized three useful target acquisition approaches in VR and AR. Our approaches enable user to perform intuitive gestures with the hand and head to acquire target objects in VR and AR. The evaluation results also provide implication for designing the target layout of the user interface and for developing the gesture recognition algorithm. Our results showed that when the hand and head are used for gesture input, a mental and behavioral model can help interpret users' input intentions from an ambiguous data set. We plan our future work to explore more input modalities and the interactions between mental model and behavior model, which would help complement the limitation of this work.

References

1 Geoffrey M. Davis. Virtual reality game method and apparatus. US Patent, 5423554, 1995-06-13

2 Gallagher A G, Ritter E M, Champion H, Higgins G, Fried M P, Moses G, Smith C D, Satava R M. Virtual reality simulation for the operating room: Proficiency-based training as a paradigm shift in surgical skills training. Annals of Surgery, 2005, 241(2): 364–372

3 Kaufmann H, Schmalstieg D, Wagner M. Construct 3D: A virtual reality application for mathematics and geometry education. Education and Information Technologies, 2000, 5(4): 263–276
DOI:10.1023/A:1012049406877

4 Bowman D A, Wingrave C A. Design and evaluation of menu systems for immersive virtual environments. In: Proceedings IEEE Virtual Reality 2001, Yokohama, Japan, 2001, 149–156
DOI:10.1109/VR.2001.913781

5 Baudel T, Beaudouin-Lafon M. Charade: remote control of objects using free-hand gestures. Communications of the

ACM, 1993, 36(7): 28–35
DOI:10.1145/159544.159562

6 Nacenta M A, Kamber Y, Qiang Y, Kristensson P O. Memorability of pre-designed and user-defined gesture sets. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Paris, France: ACM, 2013, 1099–1108
DOI:10.1145/2470654.2466142

7 Wagner J, Lecolinet E, Selker T. Multi-finger chords for hand-held tablets: recognizable and memorable. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Toronto, Canada, ACM, 2014, 2883–2892
DOI:10.1145/2556288.2556958

8 Kulshreshth A, Joseph J. LaViola J. Exploring the usefulness of finger-based 3D gesture menu selection. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Toronto, Canada, ACM, 2014, 1093–1102
DOI:10.1145/2556288.2557122

9 Wobbrock J O, Morris M R, Wilson A D. User-defined gestures for surface computing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Boston, USA, ACM, 2009, 1083–1092
DOI:10.1145/1518701.1518866

10 Bragdon A, KoH-S. Gesture select: acquiring remote targets on large displays without pointing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Vancouver, Canada, ACM, 2011, 187–196
DOI:10.1145/1978942.1978970

11 Carter M, Velloso E, Downs J, Sellen A, O'Hara K, Vetere F. PathSync. Multi-User Gestural Interaction with Touchless Rhythmic Path Mimicry. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. San Jose, USA, ACM, 2016, 3415–3427
DOI:10.1145/2858036.2858284

12 Esteves A, Velloso E, Bulling A, Gellersen H. Orbits: Gaze Interaction for Smart Watches using Smooth Pursuit Eye Movements. In: Proceedings of the 28th Annual ACM Symposium on User Interface Software&Technology. Charlotte, USA, ACM, 2015, 457–466
DOI:10.1145/2807442.2807499

13 Yatani K, Partridge K, Bern M, Newman M W. Escape: a target selection technique using visually-cued gestures. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Florence, Italy, ACM, 2008, 285–294
DOI:10.1145/1357054.1357104.

14 Shimon S S A, Lutton C, Xu Z, Morrison-Smith S, Boucher C, Ruiz J. Exploring Non-touchscreen Gestures for Smartwatches. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. San Jose, USA, ACM, 2016, 3822–3833
DOI:10.1145/2858036.2858385

15 Kray C, Nesbitt D, Dawson J, Rohs M. User-defined gestures for connecting mobile phones, public displays, and tabletops. In: Proceedings of the 12th international conference on Human computer interaction with mobile devices and services. Lisbon, Portugal, ACM, 2010, 239–248
DOI:10.1145/1851600.1851640

16 Ruiz J, Li Y, Lank E. User-defined motion gestures for mobile interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Vancouver, Canada, ACM, 2011, 197–206
DOI:10.1145/1978942.1978971

17 Vatavu R-D: User-defined gestures for free-hand TV control. In: Proceedings of the 10th European Conference on Interactive TV and Video. Berlin, Germany, ACM, 2012, 45–48
DOI:10.1145/2325616.2325626

18 Cockburn A, Quinn P, Gutwin C, Ramos G, Looser J. Air pointing: Design and evaluation of spatial target acquisition with and without visual feedback. International Journal of Human-Computer Studies, 2011, 69(6): 401–414
DOI:10.1016/j.ijhcs.2011.02.005

19 Johnson E A. A study of the effects of immersion on short-term spatial memory. Purdue Polytechnic Masters Theses.

Purdue University. 2010

20 Cockburn A, McKenzie B. Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Minneapolis, USA, ACM, 2002, 203−210
DOI:10.1145/503376.503413

21 Gutwin C, Cockburn A, Gough N. A Field Experiment of Spatially-Stable Overviews for Document Navigation. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. Denver, USA, ACM, 2017, 5905−5916
DOI:10.1145/3025453.3025905

22 Li F C Y, Dearman D, Truong K N. Leveraging proprioception to make mobile phones more accessible to users with visual impairments. In: Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility. Orlando, USA, ACM, 2010, 187−194
DOI:10.1145/1878803.1878837

23 Gugenheimer J, Dobbelstein D, Winkler C, Haas G, Rukzio E. FaceTouch: Enabling Touch Interaction in Display Fixed UIs for Mobile Virtual Reality. In: Proceedings of the 29th Annual Symposium on User Interface Software and Technology. Tokyo, Japan, ACM, 2016, 49−60
DOI:10.1145/2984511.2984576

24 Li F C Y, Dearman D, Truong K N. Virtual shelves: interactions with orientation aware devices. In: Proceedings of the 22nd annual ACM symposium on User interface software and technology. Victoria, Canada, ACM, 2009, 125−128
DOI:10.1145/1622176.1622200

25 Lopes P, Ion W A, Mueller D, Hoffmann P, Jonell P, Baudisch P. Proprioceptive interaction. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. New York, USA, ACM, 2015, 939−948
DOI:10.1145/2702123.2702461

26 Scott J, Dearman D, Yatani K, Truong K N. Sensing foot gestures from the pocket. In: Proceedings of the 23nd annual ACM symposium on User interface software and technology. New York, USA, ACM, 2010, 199−208
DOI:10.1145/1866029.1866063

27 Yan Y, Yu C, Ma X, Yi X, Sun K, Shi Y. VirtualGrasp: Leveraging Experience of Interacting with Physical Objects to Facilitate Digital Object Retrieval. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Montreal QC, Canada, ACM, 2018, 1−13
DOI:10.1145/3173574.3173652

28 Yan Y, Yu C, Ma X, Huang S, Iqbal H, Shi Y. Eyes-Free Target Acquisition in Interaction Space around the Body for Virtual Reality. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Montreal QC, Canada, ACM, 2018, 1−13
DOI:10.1145/3173574.3173616

29 LoPresti E, Brienza D M, Angelo J, Gilbertson L, Sakai J. Neck range of motion and use of computer head controls. In: Proceedings of the Fourth International ACM Conference on Assistive Technologies. ACM, New York, USA, 2000, 121−128
DOI:10.1145/354324.354352

30 Jia P, Hu H H, Lu T, Yuan K. Head gesture recognition for hands-free control of an intelligent wheelchair. Industrial Robot: an International Journal, 2007, 34(1): 60−68
DOI:10.1108/01439910710718469

31 Craig D A, Nguyen H T. Wireless real-time head movement system using a personal digital assistant (PDA) for control of a power wheelchair. 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference. Shanghai, China, 2006, 6235−6238
DOI:10.1109/IEMBS.2005.1615921

32 Gorodnichy D O, Roth G. Nouse ′use your nose as a mouse′ perceptual vision technology for hands-free games and interfaces. Image and Vision Computing, 2004, 22(12): 931−942
DOI:10.1016/j.imavis.2004.03.021

33 Varona J, Manresa-Yee C, Perales F J. Hands-free vision-based interface for computer accessibility. Journal of Network

and Computer Applications, 2008, 31(4): 357–374
DOI:10.1016/j.jnca.2008.03.003

34 Crossan A, McGill M, Brewster S, Murray-Smith R. Head tilting for interaction in mobile contexts. In: Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services. Bonn, Germany, ACM, 2009, 1–10
DOI:10.1145/1613858.1613866

35 Piumsomboon T, Clark A, Billinghurst M, Cockburn A. User-Defined Gestures for Augmented Reality. In: Human-Computer Interaction–INTERACT 2013. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, 282–299
DOI:10.1007/978-3-642-40480-1_18

36 Edge D, Blackwell A F. Peripheral tangible interaction by analytic design. In: Proceedings of the 3rd International Conference on Tangible and Embedded Interaction. Cambridge, United Kingdom, ACM, 2009, 69–76
DOI:10.1145/1517664.1517687