User Defined Eye Movement-Based Interaction for Virtual Reality

Wen-jun Hou[1,2], Kai-xiang Chen[1,2(✉)], Hao Li[1,2], and Hu Zhou[1,2]

[1] School of Digital Media and Design Arts,
Beijing University of Posts and Telecommunications, Beijing 100876, China
noideaser@163.com
[2] Beijing Key Laboratory of Network Systems and Network Culture,
Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract. Most of the applications of eye movement-based interaction in VR are limited to blinking and gaze at present, however, gaze gestures were neglected. Therefore, the potential of eye movement-based interaction in VR is far from being realized. In addition, many scholars tried to define some special eye movements as input instructions, but these definitions are almost always empirical and neglect users' habits and cultural background. In this paper, we focus on how Chinese users interact in VR using eye movements without relying on a graphical user interface. We present a guessability study focusing on intuitive eye movement-based interaction of common commands in 30 tasks of 3 categories in VR. A total of 360 eye movements were collected from 12 users and a consensus set of eye movements in VR that best met user's cognition was obtained. This set can be applied to the design of eye movement-based interaction in VR to help designers to develop user-centered and intuitive eye movement-based interaction in VR. Meanwhile this set can be migrated to other interactive media and user interfaces, such as a Post-WIMP interface base on eye movement-based interaction, as a reference to design.

Keywords: Eye movement-based interaction · Gaze gesture · Virtual reality
Guessability · Intuitive interaction

1 Introduction

With the advent of multiple screen devices such as VR devices, interaction between human and computer has become more and more frequent and complex. Many interaction techniques in VR appearing with much challenge, most of whom have obvious disadvantages including low input bandwidth, weak adaptability and far away from natural interaction. Thanks to eye-tracking technology, eye movement-based interaction which can meet the requirements of VR interface design well is becoming more and more reliable. Nevertheless, there still exist some contradictions we still need to fix, for example, most of the applications of eye movement-based interaction in VR are limited to blinking and gaze at present, however, Gaze gestures was neglected. Therefore, the potential of eye movement-based interaction in VR is far from being realized. In

P.-L. P. Rau (Ed.): CCD 2018, LNCS 10911, pp. 18–30, 2018.
https://doi.org/10.1007/978-3-319-92141-9_2

addition, many scholars tried to define some special eye movements as input instructions, but these definitions are almost always empirical and neglected users' habits and cultural background. Therefore, this article focuses on studying intuitive eye movements that provide intuitive interaction between the real world and the VR world.

2 Related Work

2.1 Interaction in VR

The development of VR interaction technology is in its infancy, there hasn't any mature solution about how to design easy-to-use VR interaction. Different enterprises have different solutions. Some try to define VR interaction using traditional interactions, such as remote control, binding handle and touchpad. Others try to combine some new and natural interaction, such as gesture interaction and voice interaction [1].

However, these current popular VR interactions are neither natural nor easy-to-use, most of which can only be used in some special scenarios. Furthermore, most of these interaction methods simply replace keyboard-mouse operation on PC or the touch-screen operation on mobile devices. Most VR interfaces design is based on WIMP interface, but WIMP interfaces have many disadvantages that can service VR well. In a nutshell, there are still many problems when design VR interactions worth exploring.

Because of its advantages of high bandwidth, naturalness, clean, etc., eye movement-based interaction gradually began to show its heads. As early as 1993, Jacob had compared the eye movement-based interaction in VR with other three-dimensional interaction techniques and found that eye movement-based interaction is superior to other interaction techniques of pointing in most scenarios [2].

In general, if we want to improve the usability of AR interactions so that VR can reach more people, there is a way that we focus on improving or inventing new input mechanism of eye movement-based interaction for VR.

2.2 Eye Movement-Based Interaction

Studies have shown that there are three modes of eye movement, gazing, saccade and smooth pursuit [3]. Gazing is the process of aligning the foveal area of eyes with a particular object. In general, the fixation time is greater than 100 ms, which is usually 200–600 ms [4], Jacob defined that a gaze input is 1000 ms in order to avoid misuse [5]. Saccade is a rapid beating of the eyeball between two fixation points and lasts for 30 ms to 120 ms. A single saccade can cover a viewing angle of 1° to 40°, usually between 15° and 20°, with a maximum speed of 400–600°/s [6]. Hyrskykari began to use saccade as a new input type which is called gaze gesture [7]. Smooth pursuit refers to the continuous movement of the eyeball with the moving target, which is only generated during the tracking of the moving target. For rest targets, there is only eye movements. The purpose of a smooth pursuit is to keep the image of the moving target near the foveal area with a maximum speed of 30°/s [8].

Eye movement-based interaction is actually recorded through the device and identify the specific mode of eye movements as the input signal to control the specific task.

Blinking, gaze and saccade these three eye movements are usually used as an input signal that is called blinking input, gaze input and gaze gesture input in Human-computer interaction. Table 1 shows the difference between blinking input, gazing input and gaze gesture input.

Table 1. Comparison between blinking input, gazing input and gaze gesture input.

	Blinking input	Gaze input	Gaze gesture input
Parameter	Blinking duration/ Blinking frequency	Fixation duration/ Fixation field	Saccade length/ Saccade duration/ Saccade velocity
Bandwidth	Low	Lowest	High
Efficiency	Fastest	Slow	Fast
Demand for interfaces' time- space characteristics	High	Very high	Low
Naturalness	Quite natural	Natural	Not very natural
"Midas contact" problem	Appears often	Appears very often	Appears rarely

Blinking input is quick and easy. But for now, limited by the development of eye movement recognition technology, it has not been widely used. The main limitations are reflected in two points. First, the awareness system can't intelligently identify the difference between physiological and unconscious blinks. The second is that the blinking itself will affect the tracking of tracking devices. Blinking's corresponding parameters are: blinking duration, blinking frequency and so on.

Gaze input is now the most popular way of eye movement-based interaction, for a relatively simple interface, it is ideal for use. However, once the interface tasks become slightly complicated, due to its high requirements on sight stability and interface time-space characteristics, user experience will exponentially decrease, including slow efficiency, easy misuse, high cost of making mistakes, and more. Gazing's corresponding parameters are: fixation duration, fixation field and so on.

Gaze gesture input is fast. Low requirements on the time-space characteristics of the interface making gaze gesture not easy to misuse. First of all, the fastest speed of saccade up to 400°–600°/s which means that Gaze gesture input can reach 1° to 40° viewing angle within 30–120 ms which is much faster than a standard gaze input unit time 300–500 ms. Secondly, as gaze gesture input does not require the interface must be presented specific interactive controls and elements, the interactive time is also relatively high robustness and it does not necessarily require an accurate response time. Interface design will be easier and faster because of the low requirements of the interface time-space characteristics. Thirdly, because gaze gesture input is based on the sequence, it does not require a precise starting point and ending point. Therefore, for the "Midas Contact" problem, gaze gesture has more advantages than blinking and gazing. Of course, the application of gaze gesture input has not yet been matured because of its own very obvious disadvantages. For example, how to design a reasonable eye movement is a very difficult research topic. If the movement is too simple, it is easy to overlap with

unconscious eye movements, resulting in misuse; once too complicated, but also increase the user's learning costs, memory burden and cognitive load, contrary to the original intention of natural interaction. Gaze gesture's corresponding parameters are: saccade length, saccade duration, saccade velocity and so on.

Most of the applications of eye movement-based interaction in VR are limited to blinking and gaze at present, however, Gaze gestures was neglected. Therefore, the potential of eye movement-based interaction in VR is far from being realized. This article attempts to let go of ideas and allow users to decide on what type of eye movement-based interaction to use.

2.3 Intuitive Interaction

In the concept of user-centered design, intuitive design is the most important part. Cognitive psychology believes that intuitive design is the process by which people can quickly identify and deal with problems based on experience. This process is unconscious, quick and easy. Blackler's research also confirmed this process [9]. Cooper also mentioned that intuitive interface design made people quickly establish a direct connection between functions and finish tasks only by relying on the guidance from interfaces [10].

Although intuitive design helps to improve the friendliness of interaction, but there are still many limitations in the practical application process. First, whether the user-designer experience is a match. If there is a difference or misunderstanding between the user's and designer's cultural background, level of education, etc., the outcome of an intuitive design may not be truly intuitive. Second, user's experience and unconscious behavior are harder to extract and quantify and most of past research is simply a qualitative description. Most or the intuitive interaction studies comes from literature research and user interviews, so that the semantic meaning of the user's definition of input symbols may not be good because it ignores habits and cultural background of the users. So that if we want to study intuitive eye movement-based interaction in VR, we must use a more scientific and more suitable method.

The speculative method proposed by Wobbrock can solve above problem well. By building guessability and level of agreement metrics, the results of user experience are quantified and can be well used to assess the degree of instinct of action design [11]. Later, Wobbrock applied this method to the research of large touch screen interaction [12]. Due to the merits of guessability method, many researches in this field were born later. For example, Ruiz et al. Studied the guessability of gesture interactions in smartphones [13]. Vatavu et al. Studied the guessability of air gesture interactions based on television manipulations [14]. Piumsomboom et al. Studied the guessability of gestural interaction in augmented reality [15]. Japanese scholar Slipasuwanchai et al. Studied the guessability of hands and feet interaction in the game [16]. Leng et al. studied the guessability of gesture interactions in VR music applications [17].

Intuitive design is of great importance to the good application of eye movement-based interaction, especially for gaze gesture interaction which has high-bandwidth but high cognitive cost. Guessability is a good method for the research on the intuitive eye movement-based interaction which deserves to be further study.

3 Experiment

3.1 Selection of Tasks

Combined with the literature review and reference to many VR applications on the market, 30 commands and tasks were derived from typical interactions in VR. This resulted in 30 task units that were grouped into 3 categories (9 sub-categories): object control (Selection, Deselection, Movement, Rotate, Uniform scale, Editing), scene control (View point transform), and system control (Global command, Temporary command). The following Table 2 shows the list of 30 selected task units under 3 categories.

Table 2. Universal operating tasks in VR.

Category	Tasks	Task units
Object control	Selection	Single selection/Multiple selection/Select all
	Deselection	Deselection
	Movement	Move up/Move down/Move left/Move right/ Move forwards/Move backwards
	Rotate	X-axis/Y-axis/Z-axis
	Uniform scale	Scale up/Scale down
	Editing	Copy/Paste/Delete/Redo
Scene control	View point transform	Upward view point/Downward view point/ Leftward view point/Rightward view point/ Zoom in/Zoom out
System control	Global command	Open/ Close menu/ Play menu
	Temporary command	Accept/confirm Reject/cancel

3.2 Participants and Device

12 participants (6 males and 6 females) were voluntarily recruited, ranging in age from 21 to 25 years old (mean = 23.3 years old and SD = 1.44 years). All of the participants had visual acuity or corrected visual acuity of 5.0 or above. Participants must have minimal knowledge of experiencing VR eye movement-based interaction in order to avoid the impact of their prior experience on the definition of the set of eye movements. The experimental device is HTC Vive which running on the software we set up for the experiment. A video recording device was used to record what participants had said during the experiment. Experimental staff for the experimental including an operator, a recorded and a host.

3.3 Procedure

Firstly, the host introduced the basic concept of VR and eye movement-based interaction to the participants. Secondly, the host introduced the details of this experiment and all of the

30 tasks to the participants. In addition, participants were required that they do not need to take technical implementation issues such as the recognition accuracy of the eye movement into consideration, just try to imagine the most suitable eye movements that were best suited to the task. Finally, the participants were asked to wear the equipment in a sitting posture and start the experiment. Figure 1 shows the specific experimental scenario.

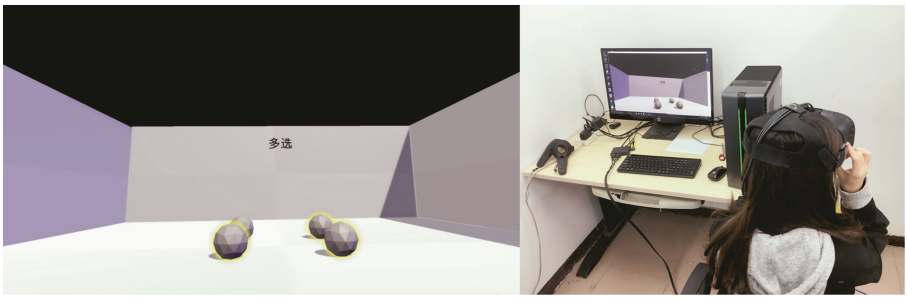

Fig. 1. The experimental scenario.

Participants started the experiment according to the Latin square experimental order in order to avoid interference caused by the legacy effect on the experimental results. For each experimental task, the VR equipment worn by participants was accompanied by the name of the task and an animation of the target scene to clearly convey the operation task to the participants. Target scenes were created by using Unity software and Steam VR, and finally applied to HTC Vive. Target scenes were made independent of any particular application such a football game which might influence the result.

During the experiment, participants were asked to define eye movements for 30 VR tasks. Meanwhile, a think-aloud protocol was used to let participants indicated the start and end of their performed eye movements and described the reason. A camera was set in front of the participant to record the experimental details, such as voice, for later analysis.

After each experiment, participants were also asked to immediately subjectively evaluate the performance of eye movements they defined from three indicators which are matching, easiness and fatigue: "The eye movement I performed is good match for its purpose"; "The gesture I performed is easy to perform"; "The gesture I performed is not tiring". The evaluation questionnaire was designed using Likert's 7-point scale, of which 7 for "very agree" and 1 for "very different agree." The entire experiment took about 40 min.

4 Results

A total of 360 eye movements were collected from 12 participants who performed 30 selected tasks. Then a consensus set of eye movements in VR that best meets user's cognition were obtained.

4.1 Designing User-Defined Eye Movements Set

Since different participants might defined different eye movements for a same task, the eye movements gotten from participants can't just put together and define the set. The eye movement symbols collected from the experiment can't simply be put together to form an action set, as different users may have different definitions of the input action. According

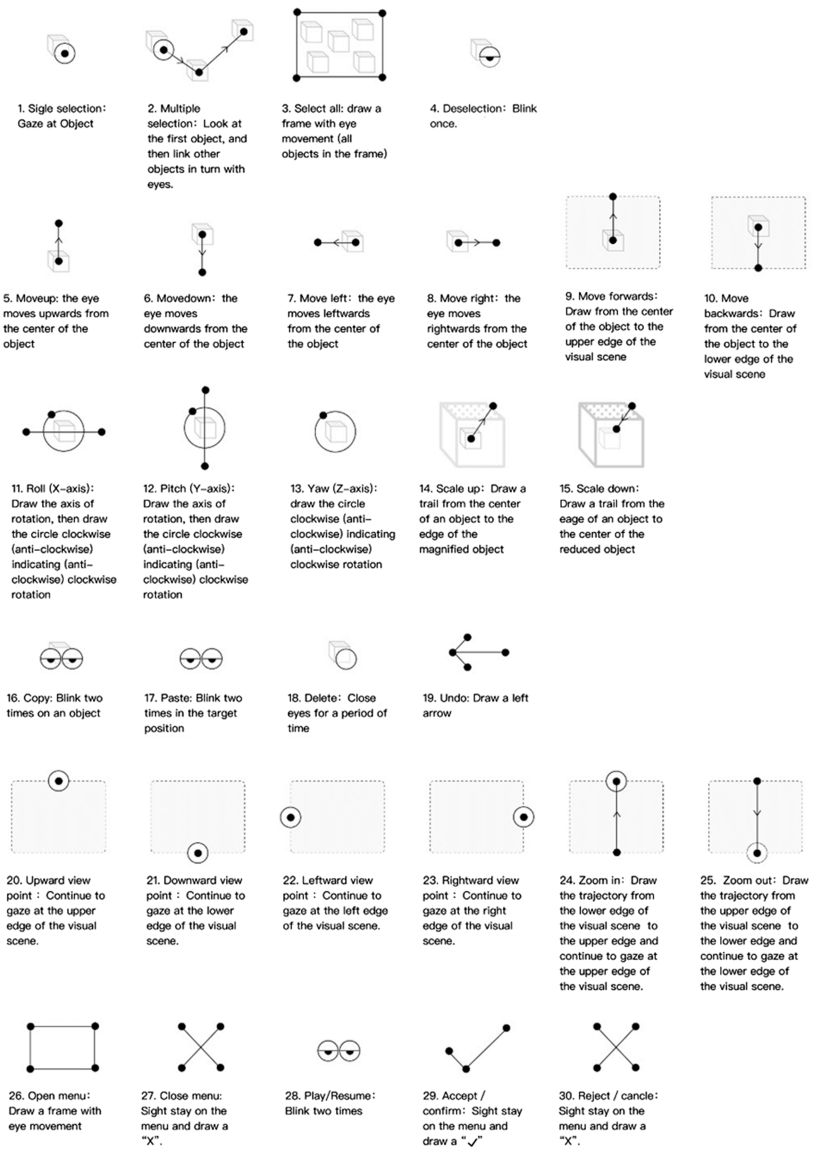

Fig. 2. The user-defined consensus gesture set for VR.

to guessability method, the action symbol with the highest frequency of occurrence is selected as the standard input symbol of the task, and its score is recorded as:

$$score = |symbols| \tag{1}$$

In Eq. 1, symbols are the appearance frequency of the standard input symbols, and the standard input symbols in each operation task are grouped together as a consensus set. Figure 2 shows the consensus set of eye-movement based interaction in VR acquired in this experiment.

In the consensus set, the blinking input accounted for 5/30 (closing eyes for a period of time were also taken for blinking input), the gaze input accounted for 8/30 and the gaze gesture input accounted for 20/30, of which there are 3 more complex gaze gestures using a circle symbol. It is noteworthy that there is a mixed input accounted for 3/30 in the consensus set.

$$G = \frac{\sum_{s \in S} |P_s|}{|P|} \cdot 100\% \tag{2}$$

In Eq. 2, G is the guessability score, P is the set of proposed symbols for all referents, and Ps is the set of proposed symbols using symbol s, which is a member of the resultant symbol set S. Figure 3 shows the guessability score for 30 tasks in descending order.

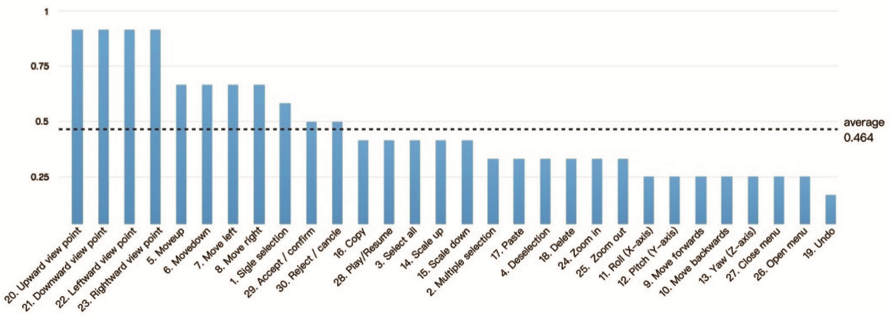

Fig. 3. Guessability score for 30 tasks in descending order.

The average guessability score for all movements in the consensus set is 46.39% (SD = 0.2292), which is relatively high. The average guessability score of object operation category was 40.35% (SD = 0.1673), of scene operation category was 72.22% (SD = 0.3012), and of system control was 38.33% (SD = 0.1264).

4.2 Level of Agreement

The agreement score was calculated by the Eq. (3) to evaluate the cognitive quality of the standard input symbols and the user group's level of awareness of the input symbols. The higher the score indicating that users can know more easily to know about what the symbols mean just by the characteristics of these symbols rather than learning these

symbols. At the same time, the cognition between users is relatively close and the scattered levels of cognitive is low.

$$A = \frac{\sum_{r \in R} \sum_{P_i \in P_r} \left(\frac{|P_i|}{|P_r|} \right)^2}{|R|} \tag{3}$$

In Eq. 3, A is the agreement score, r is a referent in the set of all referents R, P_r is the set of proposals for referent r, and P_i is a subset of identical symbols from P_r. The range of Eq. 3 is $1/|P_r| * 100\% \leq A \leq 100\%$. The lower bound is non-zero because even when all proposals disagree, each one trivially agrees with itself. Figure 4 shows the agreement score for 30 tasks in descending order.

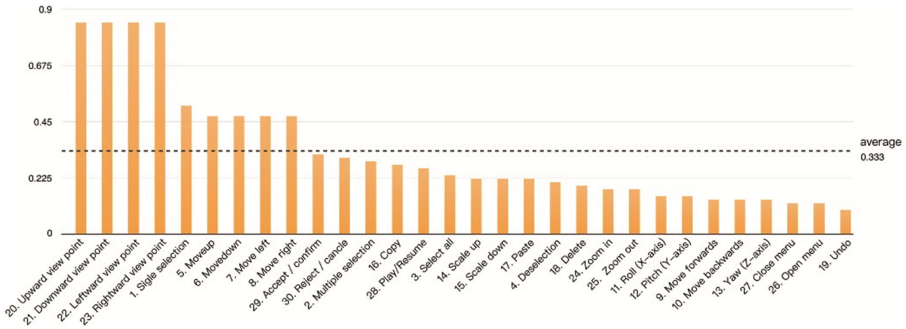

Fig. 4. Agreement score for 30 tasks in descending order.

The average agreement score for all movements in the consensus set is 33.27% (SD = 0.2369), which is relatively high. The average agreement score of object operation category was 26.83% (SD = 0.1394), of scene operation category was 62.50% (SD = 0.3442), and of system control was 22.69% (SD = 0.0972).

By comparing Fig. 2 with Fig. 3, we found that the agreement scores of single selection and multiple selection ranked higher than the guessability scores. The reason was that the eye movements defined by participants for these two tasks are more concentrated.

For single selection which was the most basic task in all VR scenarios (this task was also the leading task for most other tasks), the user's opinion divided into two groups that 7 users used gazing input while 5 users used blink input. For multiple selection, in addition to entering the consensus set of movement (look at the first object, and then link other objects in turn with eyes), the majority of users chose to repeat using the single action, which means blinking or gazing in turn.

Participants commonly find it hard to subconsciously think of appropriate movements to finish the task when the task is relevant with depth in the 3D scene. the design of these actions took more time during experiment. Judging from the results, the consistency levels of the movements such as zoom in, zoom out, move forwards, move backward, and the Yaw(Z-axis) were lower than those of the same category.

There were some abstract tasks such as “Undo” and “Open Menu” in the system control category and the object operation category. Though the agreement scores of these abstract movements were all lower than the others, the definitions were very similar to each other. This result indicating that these abstract movements may exist more appropriate definition. On the other hand, it may be that these abstract task instructions have a higher bandwidth of semantic, leading to different user preferences. High bandwidth of semantic for users provides more ways of movements design, which may be beneficial for users, it is worthy to go deeper to explore the definition of these abstract tasks.

4.3 Subjective Evaluation

In the experiment, participants were also asked to subjectively evaluate from three indicators which are matching, easiness and fatigue. For the convenience of description, we called the set which included all the standard eye movements as consensus set, and he set which included all the rest eye movements as discard set.

By comparing the consensus set and the discard set, we found that the performance of the consensus set was overall better than the other from the results of the descriptive statistics. In terms of matching indicator, the average score of consensus set is 5.61 higher than the discard set of 4.76. In terms of the easiness indicator, the average score of consensus set is 5.36 higher than the discard set of 4.88. In terms of fatigue indicator, the average score of consensus set is 5.07 higher than the discard set of 4.74. In general, the subjective evaluation of eye movements from the user-defined consensus set were better than the eye movements of the discard set (Fig. 5).

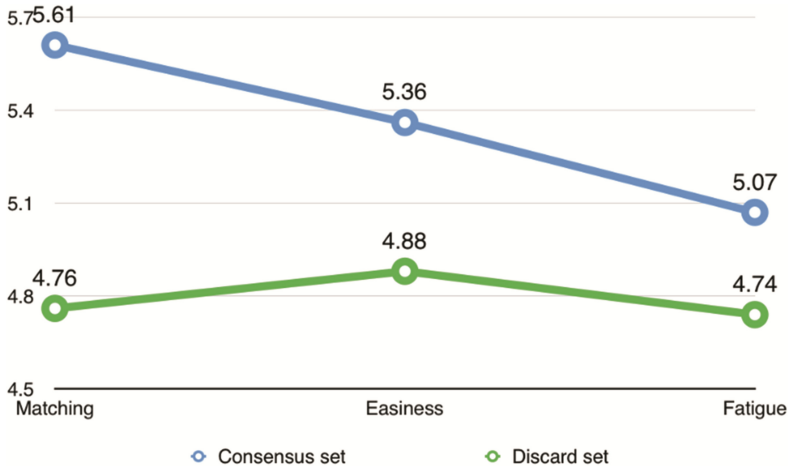

Fig. 5. Subjective evaluation of consensus set and discard set.

In addition, if the consensus set is classified into blinking, gaze, gaze gesture and mixed input according to the input type of eye movements, differences were found in all three indicators by the results of the descriptive statistics. As can be seen from the

comparison in Fig. 6, the blinking input is optimal in terms of easiness and fatigue. The gazing input is optimal in the matching indicator, which was slightly higher than the gaze gesture input. The gaze gesture input performed well in matching but both easiness and fatigues of gaze gesture were lower than blinking and gazing, suggesting that gaze gesture interaction has great potential for intuitive design. Lastly, the mixed input is far lower than non-mixed input in terms of all indicators.

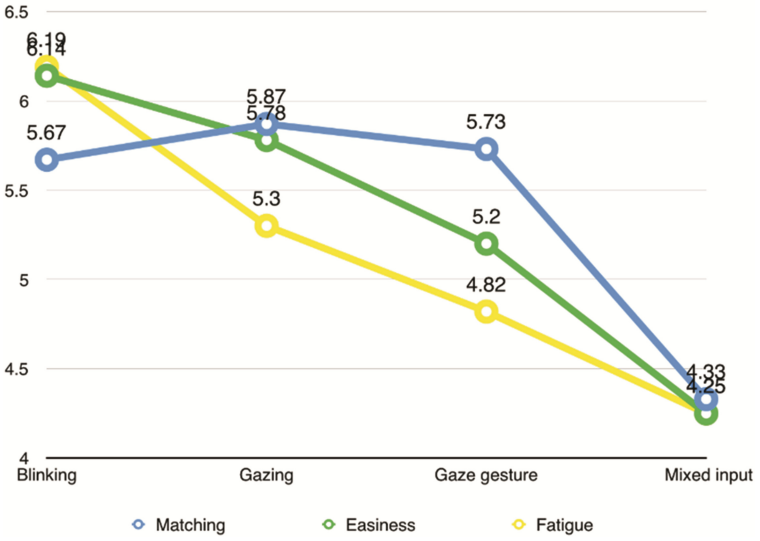

Fig. 6. Subjective evaluation of different input types.

5 Conclusion

This paper presented a guessability study focusing on intuitive eye movement-based interaction in VR. Thanks to the participants, the problem caused by the mismatch of experience between designers and users was effectively reduced, and finally a consensus set of eye movements in VR that best met user's cognition were obtained as a reference to the relevant design to help users make better use of eye movement-based interaction in VR.

From this study, we can also get these conclusions:

- The tasks that were relevant with depth in the 3D scene were more difficult to intuitive design and defined by participants.
- Participants subconsciously thought that the eye movements itself can show information about depth in 3D scene.
- Gaze gesture input for the expression of multi-dimensional information had a better overall performance.
- Abstract tasks had a higher bandwidth of semantic leading to different definitions of eye movements. These definitions are very similar to each other.

- Participants tended to repeat defining different tasks by using same simple eye movements.
- Gaze gesture input more easily led to fatigue with the eyes, so that the user always halfway changed into fuzzy input.

In the follow-up study, we will conduct a more in-depth study based on the above findings, so that the application of eye movement-based interaction in VR can be more natural and effective.

References

1. Jerald, J., LaViola Jr., J.J., Marks, R.: VR interactions. In: ACM SIGGRAPH, Los Angeles, USA, pp. 1–105. ACM Press, New York (2017)
2. Jacob, R.J.: Eye movement-based human-computer interaction techniques: toward non-command interfaces. Adv. Hum. Comput. Interact. **4**, 151–190 (1993)
3. Haber, R.N., Hershenson, M.: The Psychology of Visual Perception, 1st edn. Holt, Rinehart & Winston, Oxford (1980)
4. Young, L.R., Sheena, D.: Survey of eye movement recording methods. Behav. Res. Methods Instrum. **7**(5), 397–429 (1975)
5. Jacob, R.J.K.: What you look at is what you get: eye movement-based interaction techniques. ACM Trans. Inf. Syst. **9**(2), 152–169 (1990)
6. Goldberg, J.H., Kotval, X.P.: Computer interface evaluation using eye movements: methods and constructs. Int. J. Ind. Ergon. **24**(6), 631–645 (1999)
7. Hyrskykari, A., Istance, H., Vickers, S.: Gaze gestures or dwell-based interaction? In: Proceedings of the Symposium on Eye Tracking Research and Applications, California, USA, pp. 229–232. ACM Press, New York (2012)
8. Shioiri, S., Cavanagh, P.: Saccadic suppression of low-level motion. Vis. Res. **29**(8), 915–928 (1989)
9. Blackler, A., Popovic, V., Mahar, D.: Investigating users' intuitive interaction with complex artefacts. Appl. Ergon. **41**(1), 72–92 (2010)
10. Cooper, A., Reimann, R., Cronin, D.: About Face 3: The Essentials of Interaction Design. Wiley Publishing, Inc., Indiana (2007)
11. Wobbrock, J.O., Aung, H.H., Rothrock, B., Myers, B.A.: Maximizing the guessability of symbolic input. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, Portland, USA, pp. 1869–1872. ACM Press, New York (2005)
12. 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, pp. 1083–1092. ACM Press, New York (2009)
13. 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, BC, Canada, pp. 197–206. ACM Press, New York (2011)
14. 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, pp. 45–48. ACM Press, New York (2012)
15. Piumsomboon, T., Clark, A., Billinghurst, M., Cockburn, A.: User-defined gestures for augmented reality. In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, Paris, France, vol. 8118, pp. 955–960. ACM Press, New York (2013)

16. Silpasuwanchai, C., Ren, X.: Jump and shoot!: prioritizing primary and alternative body gestures for intense gameplay. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, Toronto, Canada, pp. 951–954. ACM Press, New York (2014)
17. Leng, H.Y., Norowi, N.M., Jantan, A.H.: A user-defined gesture set for music interaction in immersive virtual environment. In: Proceedings of the 3rd International Conference on Human-Computer Interaction and User Experience, Indonesia, pp. 44–51. ACM Press, New York (2017)