Quality of Interaction Arising from Augmented Reality Content: A Comprehensive Study

Seongjean Kim*, Jinwoo Kim*, and Sanghoon Lee*†
* Department of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea
† Department of Radiology, College of Medicine, Yonsei University, Seoul, South Korea
E-mail: {jin.k, jw09191, slee}@yonsei.ac.kr, Tel: +82-2-2123-7734

***Abstract*—Current augmented reality (AR) head-mounted displays (HMDs) have rapidly developed with the users' requirement to expand the real-world experience to the virtual world. To bridge the real and virtual space, mid-air hand gestures have been regarded as a de-facto interaction method for AR systems. However, providing natural interaction is still limited and measuring the quality of interaction (QoI) has received little attention. In this research, we explore a comprehensive study for perceptive QoI in AR HMDs, focusing on frequently used object interaction tasks. To quantitatively analyze the degree of QoI, we develop an *AR-QoI* database (e.g., select, translation, and rotation) with 72 content scenes that include mutually independent attributes. A novel protocol for QoI evaluation was designed to collect robust subjective opinions in conjunction with object data from 32 participants. Through a systematic formative study, we identify challenges that the user might face when interacting with unfamiliar motion. Moreover, we discover meaningful relationships between various combinations of interaction types and the degree of QoI by clustering scene attributes. Notably, the constructed dataset contains a number of ground-truth labels that correspond to each AR scene. Through rigorous statistics evaluation, we demonstrate that our framework is reasonable for measuring QoI.**

I. Introduction

To maximize the sense of reality, visual content is developed for more expressive environments such as augmented, virtual, and mixed reality (AR, VR, and MR). Among them, the AR industry is rapidly emerging, and user expect an increasingly high quality of interaction (QoI) for AR services. To bridge real and virtual space, the hand gesture is regarded as one of the most commonly accepted interaction methods in AR environment [1]–[4]. Therefore, various hand gesture interaction techniques such as retargeting [5], redirection [6], psudo-haptics [7], and control to display [8] are incorporated into AR services. Despite this success, humans still unsatisfy current AR QoI system due to unwanted distrubances including computational delay and unfamiliar interaction, and this is regarded as a hindrance to the development of the related industry.

Improving interaction performance has been studied in the field of computer vision area to enhance model accuracy (hand detection, tracking, and joint estimation) or to reduce computational cost (hand mesh rendering). On the other hand, measuring QoI from the user's perspective has received little attention. In this research, we explore comprehensive studies quantifying the user's interaction satisfaction in AR contents which is an essential factor to provide an improved system.

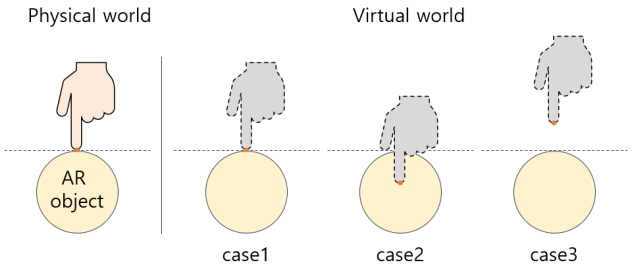


Fig. 1. Overview of the interaction cases regarding the physical and virtual movements with differently remapped virtual hands.

However, these studies have not been benefited from the availability of sizeable labeled datasets. To this end, we develop a new and large database that is capable of exploring dominant factors that lead to reduced QoI.

This research begins with a formative study to examine the challenges that users might face when interacting with AR HMDs. As shown in Fig. 1, inaccurate capturing of hand gestures from real-space to virtual-space makes interaction tasks much more challenging, resulting in insufficient experience in AR environments. More specifically, if the virtual hand is reconstructed forward of the physical hand, the perceptive interaction range will increase (case2), and vice versa, the interaction range will decrease (case3). The discrepancy between the real-world and virtual-world interaction provokes self-contributing movements, which are unnecessary movements to fully control virtual hand behavior [9]–[14]. Therefore, we hypothesize that mapping virtual hands correctly can provide sufficient QoI [15], [16]. Based on this assumption, we propose a comprehensive study to find interaction attributions that hinder motion remapping from real space to virtual space.

To achieve our goal, we first examine any interaction gestures in AR HMDs which may cause in-use difficulties to users. After that, we categorize the types of hand gesture interactions according to the complication levels based on the subjects' comments, our observations, and interviews. The synthetic AR content produced contains 72 contents with different interaction types that correspond to various gesture poses. We design a novel QoI evaluation protocol to obtain self-diagnosis responses using questionnaires. Moreover, we obtain a subjective QoI score for each AR content using a

TABLE I
LIST OF CONTENT AND THE CORRESPONDING ATTRIBUTES IN AR QOI DATASET

Content index	Interaction type	Hand gesture type	Position of virtual object	Interaction with reference	Content index	Interaction type	Hand gesture type	Position of virtual object	Interaction with reference	Content index	Interaction type	Hand gesture type	Position of virtual object	Interaction with reference
C001	Select	Index	Close + Predictable	O	C025	Translation	Pinch (With open fingers)	Close + Predictable	O	C049	Rotation	Pinch (With open fingers)	Close + Predictable	O
C002				X	C026				X	C050				X
C003			Close + Unpredictable	O	C027			Close + Unpredictable	O	C051			Close + Unpredictable	O
C004				X	C028				X	C052				X
C005			Far + Predictable	O	C029			Far + Predictable	O	C053			Far + Predictable	O
C006				X	C030				X	C054				X
C007			Far + Unpredictable	O	C031			Far + Unpredictable	O	C055			Far + Unpredictable	O
C008				X	C032				X	C056				X
C009		Index + middle	Close + Predictable	O	C033		Pinch (With close fingers)	Close + Predictable	O	C057		Pinch (With close fingers)	Close + Predictable	O
C010				X	C034				X	C058				X
C011			Close + Unpredictable	O	C035			Close + Unpredictable	O	C059			Close + Unpredictable	O
C012				X	C036				X	C060				X
C013			Far + Predictable	O	C037			Far + Predictable	O	C061			Far + Predictable	O
C014				X	C038				X	C062				X
C015			Far + Unpredictable	O	C039			Far + Unpredictable	O	C063			Far + Unpredictable	O
C016				X	C040				X	C064				X
C017		Full hand	Close + Predictable	O	C041		Grab	Close + Predictable	O	C065		Grab	Close + Predictable	O
C018				X	C042				X	C066				X
C019			Close + Unpredictable	O	C043			Close + Unpredictable	O	C067			Close + Unpredictable	O
C020				X	C044				X	C068				X
C021			Far + Predictable	O	C045			Far + Predictable	O	C069			Far + Predictable	O
C022				X	C046				X	C070				X
C023			Far + Unpredictable	O	C047			Far + Unpredictable	O	C071			Far + Unpredictable	O
C024				X	C048				X	C072				X

graphical user interface during the rating procedure. Based on data acquired from 32 subjects, we perform statistical tests to identify a quantitative relationship between various factors, QoI, and its severity. To verify the superiority of the constructed dataset, we use QoI prediction schemes. We report the predictive performance in terms of the correlation to the subject's scores.

II. FORMATIVE STUDY

We could not find any prior work that has focussed on measuring the QoI considering both hand gesture types and interaction type in AR HMD scenarios. To guide our QoI dataset construction, we carried out a formative study to figure out the in-use difficulties during user interaction with virtual objects according to mutually independent attributes.

A. Formative Study: Method

The 23 inexperienced participants (satisfying the subject criteria recommended in [17]) were recruited for the formative study. All participants were of ages ranging from 22 to 34 years and were screened for normal visual acuity on the Landolt chart. For the study, each participant was exposed to a variety of mid-air hand interaction tasks (e.g., manipulating virtual objects in MRTK HandInteractionExample scene). After the tutorial, participants interacted with the AR-HMD evaluation following a thinking-aloud protocol [1]. They were asked to record what they found, what challenges they had, and which interaction attributes led to hinder the AR immersion.

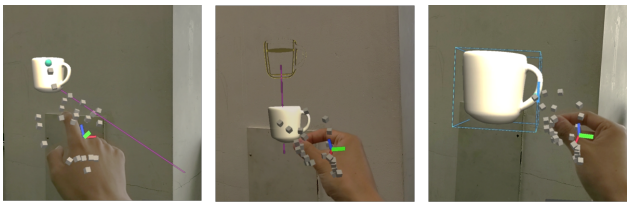
B. Formative Study: Findings

We found 3 main attributes that were extrapolated from participants' comments, our observation during each formative study, and post-questionnaires.

1) Interaction types: During the study, participants were exposed to well-used interaction gestures in AR contents which are provided as a standard in the HoloLens device (e.g., bloom, ready, tap, hold, and drag). Using a variety of gestures, they conducted AR interaction scenarios including object selection, translation, and rotation. In most cases, perceptually natural interactions were recorded in the scenarios involving simple types of interactions. On the other hand, scenarios involving complex types of interaction led to discomfort feelings and questioned the ability to work with AR devices. This finding led us to hypothesize that reconstruction mismatch between virtual and real hands could occur more in complex interaction scenarios.

2) Distance to virtual objects: From the study, participants responded that it would be helpful for interactable virtual objects to maintain QoI if they are positioned in a balanced place. In particular, when the virtual object was located close to the display, participants indicated that the device could not provide a natural QoI because of the interaction boundary awareness issue caused by limited device geometry. On the contrary, participants reported discomfort when the virtual object was located farther than the users' arm length. By positioning virtual objects at a proper place, we could prevent participants' hands from colliding with real-world objects or going out of the device geometry.

3) Predictable interaction: We found another important factor through formative study. Participants tended to report favorable experiences when faceing predictable object interaction situations. In particular, when virtual objects are crucially located and interaction types are predictable, participants felt the interactions more natural. This was reasonable because the mechanism controlling body balance in the brain is much more stable when they estimate predictable body motion.



Select Translation Rotation

Fig. 2. Visualization of representative example of AR content.

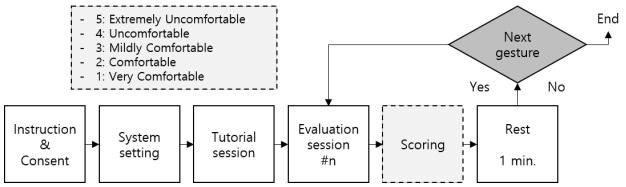


Fig. 3. Designed protocol for subjective QoI evaluation.

III. Dataset Construction

Findings from the formative study guided us to construct the novel dataset to measure QoI in AR environments. The main scenario of the generated AR content was developed and run using the Unity3D engine. We chose three well-used interaction types: select, translation, and rotation. Using this, our dataset was constructed according to the core mutually independent attributes described below. Furthermore, we set the background of all scenes as empty spaces so that the subjects can focus more on virtual object interaction. The factors used in content construction are summarized in Table I and we visualize the representative AR content in Fig. 2.

A. Hand Gesture Type

This condition provides the user to interact with virtual objects in a variety of hand poses, ranging from micro-gesture to macro-gesture. In the selecting case, we employ three hand gestures including index finger, index+middle finger, and full hand. For each content, subjects choose the virtual object only using the designated fingers. For both translation and rotation cases, pinch (with open fingers), pinch (with close fingers), and grab are utilized for the hand gesture types. To distinguish each hand gesture type, we utilize the OpenCV hand gesture recognition library.

B. Position of Virtual Object

In each scenario, the virtual objects are positioned based on the subject's arm length, referring to the formative study. To do this, we measure the arm length of each subject and record the AR environment before conducting a subjective evaluation. The device then automatically displays the virtual object inside or outside of this criterion. Furthermore, we reflect predictable interaction factors discussed in the formative study. This condition is implemented by allowing the virtual objects that appear in each scenario to be determined or to appear in a random location.

C. Interaction with Reference

We construct a dataset to ensure that the guidelines effectively provide natural interaction quality. This is achieved by providing references that allow users to easily interact with virtual objects. Therefore, participants execute a reaching movement with visual feedback of the interaction direction. By doing so, we can figure out that users are aware of the tracked interaction area and record plausible interaction scores. On the content table, 'O' and 'X' represents whether the reference is provided or not.

IV. Subjective Assessment

A. Subjective Assessment: Protocol

Both objective data (i.e., rendered hand data, scene parameters, and interaction duration) and subjective data (opinions on the degree of QoI) are collected when the subjects conduct the constructed AR contents. For this analysis, we design a novel evaluation protocol as shown in Fig. 3.

Before the evaluation, we provided tutorial content introducing the evaluation process and the degrees of QoI. During the interaction, six AR scenes were shown, broadly covering the range of AR factors. This process helped to record more reliable data in two ways. First, participants became familiar with the interaction methods used in the experiment, which helped to avoid potentially unwanted manipulation errors. Second, the tutorial led participants to normalize subjective opinion scores within the overall AR scene due to human psychological expectations [18].

Following the tutorial session, the full assessment session consists of three parts: a rest session, a display session, and an evaluation session. Rest periods of 1 min duration were set to minimize any accumulated feeling of AR sickness. In each session, an AR sequence was displayed. Depending on the user's proficiency, the display session's time was differently recorded. After finishing each interaction in content, participants asked to record subjective QoI scores using a 5-points Likert scale marked as follows: Extremely Uncomfortable (5), Uncomfortable (4), Mildly Comfortable (3), Comfortable (2), and Very Comfortable (1).

B. Subjective Assessment: Analysis

After subjective evaluation, we validate the subjective score obtained from each content through statistical analysis. To this end, we first compute the mean opinion score (MOS) represented as:

$$d_k = \frac{\sum_{j=1}^{N} s_{jk}}{N}, \tag{1}$$

where N is the number of subjects and s_{jk} is the score delivered by subject j on the content k.

To measure the statistical reliability of the predicted data, we computed the confidence interval(CI) on the obtained MOS. Using the MOS, the CI of $100 \times (1-\alpha)\%$ is computed using the interval estimation:

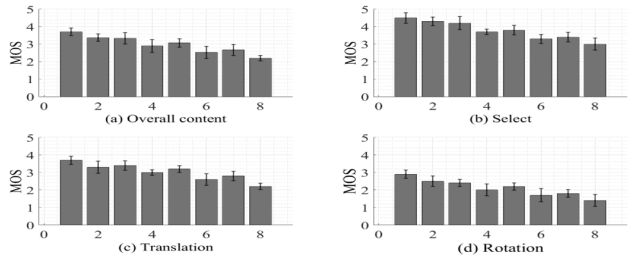


Fig. 4. MOS of content for (a) Overall, (b) Select, (c) Translation and (d) Rotation

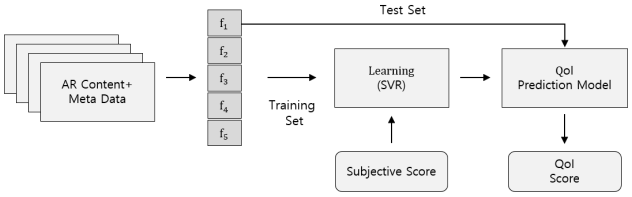


Fig. 5. Framework of objective assessment.

$$CI = d_k \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{N}}, \tag{2}$$

where d_k is the computed MOS of each content, σ is the standard deviation, and $Z_{\alpha/2}$ is the $z-$score for $\alpha/2$. For this research, we set $\alpha = 0.05$ in accordance with a confidence level of 95%, and $N = 32$ as to the number of subjects.

Fig. 4 shows the content MOS averaged for each interaction with a 95% confidence interval. Each bar represents the average result of the three hand gesture type for the 8 contents. By constructing the dataset using the independent variables from the formative study, we have achieved in distributing the subjective scores to the 5 scores without being biased towards a specific score.

V. Objective Assessment

After obtaining the level of QoI score from the evaluation session, this subjective score is used as the ground-truth label of the corresponding AR scene for supervision of the QoI prediction model. To show how this constructed dataset can quantify the level of QoI, we designed simple features by adopting both meta data from HMDs device and recorded videos. These features are regressed onto the subjective score by using support vector regression (SVR).

A. Objective Assessment: Method

The overall processing of the objective assessment is depicted in Fig. 5. As we mentioned above, differently remapped virtual hands are an important contributing factor of QoI. In this context, to extract representative features that reflect real and virtual mismatch in the AR environment, we use an optical flow algorithm [19] and hand mesh position represented in cartesian coordinates.

Let $m_{n,t}$ be the motion vector of the n^{th} pixel at the t^{th} frame. The first motion feature f_1 is the average motion magnitude obtained by spatiotemporal mean pooling of $m_{n,t}$ in the AR content:

$$f_1 = \frac{1}{T \cdot N} \sum_t \sum_n |m_{n,t}|, \tag{3}$$

where N and T are the total number of motion vectors and frames, respectively. Moreover, the variance of motion magnitude is extracted to represent the holistic distribution of motion vector for QoI measure, which is defined as:

$$f_2 = \frac{1}{T \cdot N} \sum_t \sum_n (|m_{n,t} - f_1|)^2. \tag{4}$$

After we define motion factors, the relative difference between real and virtual positions is represented using the recorded hand position ordinate in word coordinate. Motivated by previous literature [20], we compute flow field in each frame as follow:

$$f_3 = u_{n,t} - \frac{I_{x,t}u_{n,t} + I_{y,t}v_{n,t} + I_{z,t}}{\lambda + I_x^2 + I_y^2 + I_z^2} I_{x,t}, \tag{5}$$

$$f_4 = v_{n,t} - \frac{I_{x,t}u_{n,t} + I_{y,t}v_{n,t} + I_{z,t}}{\lambda + I_x^2 + I_y^2 + I_z^2} I_{y,t}, \tag{6}$$

$$f_5 = \frac{I_{x,t}u_{n,t} + I_{y,t}v_{n,t} + I_{z,t}}{\lambda + I_x^2 + I_y^2 + I_z^2} I_{z,t}, \tag{7}$$

where $I_{k,t}$ is relative difference calculated by $k(t) - k(t-1)$ and $k = x, y, z$, and $u_{n,t}$ and $v_{n,t}$ are the subset of motion factor which is obtained as $m_{n,t} = \sum(u_{n,t}^2 + v_{n,t}^2)$.

Using the above-defined features, we learn the SVR along with the subjective QoI score. At each trial, the randomly chosen training set consists of 80% of the constructed dataset with data-label pairs, and the test set consists of the remaining 20% data-label pairs.

B. Objective Assessment: Result

Predictive performance was evaluated using Spearman's rank-order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) relative to the QoI score. Table II shows the performance of the tested models in terms of SRCC and PLCC. To better understand the contributions of each extracted feature, we compared the performance of the subset using the various feature combination. Although simple features are extracted, the prediction model shows reasonable performance. Note that when $f_1 - f_5$ are employed to train the model, we achieve the best predictive performance, which means that all the features are partially correlated with the QoI score. Otherwise, when the motion features from the captured video are simply used to train the model, the predictive performance decreases. The result shows that stronger real and virtual remapping differences stimulate a higher degree of degraded QoI.

TABLE II
PERFORMANCE OF 100 TRIALS OF RANDOMLY CHOSEN TRAIN AND TEST SETS

Features	Train		Test	
	SRCC	PLCC	SRCC	PLCC
(f_1, f_2)	0.4032	0.4178	0.4142	0.3996
$+f_3$	0.5138	0.5389	0.5083	0.5128
$+f_4$	0.5828	0.6032	0.5739	0.5914
$+f_5$	0.6588	0.6727	0.6439	0.6631

VI. CONCLUSIONS

In this paper, we have proposed a framework to evaluate the interaction quality of hand gestures in the AR HMD environment. To this end, we first conducted a formative study to figure out the difficulties when the user interacts with virtual objects. The constructed dataset covers various key attributes discussed from the formative study. Using this AR content, we performed a subjective assessment by adopting a 5-points Likert scale. To further validate our dataset, we show how this dataset can quantify the level of QoI. Therefore, we first design simple features using the recorded data, and then, these features regressed with the corresponding subjective scores for the supervision of QoI prediction model. We expect that our proposed framework can more easily predict user QoI for newly created hand gestures in AR environments, which will lead to the implementation of more diverse functions in line with the rapid development of the AR industry.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT)(No.2020-0-00537, Development of 5G based low latency device – edge cloud interaction technology)

REFERENCES

[1] W. Xu, H.-N. Liang, Y. Chen, X. Li, and K. Yu, "Exploring visual techniques for boundary awareness during interaction in augmented reality head-mounted displays," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 204–211, IEEE, 2020.

[2] N. Katzakis, L. Chen, R. J. Teather, O. Ariza, and F. Steinicke, "Evaluation of 3d pointing accuracy in the fovea and periphery in immersive head-mounted display environments," *IEEE transactions on visualization and computer graphics*, 2019.

[3] N. Ogawa, T. Narumi, and M. Hirose, "Effect of avatar appearance on detection thresholds for remapped hand movements," *IEEE transactions on visualization and computer graphics*, 2020.

[4] H. G. Debarba, J.-N. Khoury, S. Perrin, B. Herbelin, and R. Boulic, "Perception of redirected pointing precision in immersive virtual reality," in *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 341–346, IEEE, 2018.

[5] M. Azmandian, M. Hancock, H. Benko, E. Ofek, and A. D. Wilson, "Haptic retargeting: Dynamic repurposing of passive haptics for enhanced virtual reality experiences," in *Proceedings of the 2016 chi conference on human factors in computing systems*, pp. 1968–1979, 2016.

[6] L. Kohli, M. C. Whitton, and F. P. Brooks, "Redirected touching: Training and adaptation in warped virtual spaces," in *2013 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 79–86, IEEE, 2013.

[7] D. A. G. Jauregui *et al.*, "Toward" pseudo-haptic avatars": Modifying the visual animation of self-avatar can simulate the perception of weight lifting," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 4, pp. 654–661, 2014.

[8] Y. Ban, T. Narumi, T. Tanikawa, and M. Hirose, "Displaying shapes with various types of surfaces using visuo-haptic interaction," in *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, pp. 191–196, 2014.

[9] J. Kim *et al.*, "A deep motion sickness predictor induced by visual stimuli in virtual reality," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[10] J. Kim, W. Kim, H. Oh, S. Lee, and S. Lee, "A deep cybersickness predictor based on brain signal analysis for virtual reality contents," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10580–10589, 2019.

[11] W. Kim *et al.*, "Modern trends on quality of experience assessment and future work," *APSIPA Transactions on Signal and Information Processing*, vol. 8, 2019.

[12] W. Kim, S. Lee, and A. C. Bovik, "Vr sickness versus vr presence: A statistical prediction model," *IEEE Transactions on Image Processing*, vol. 30, pp. 559–571, 2020.

[13] W. Kim, A.-D. Nguyen, S. Lee, and A. C. Bovik, "Dynamic receptive field generation for full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4219–4231, 2020.

[14] W. Kim, J. Kim, and S. Lee, "Quality of experience using deep convolutional neural networks and future trends," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1556–1559, IEEE, 2019.

[15] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1676–1684, 2017.

[16] J. Kim, W. Kim, S. Ahn, J. Kim, and S. Lee, "Virtual reality sickness predictor: Analysis of visual-vestibular conflict and vr contents," in *2018 tenth international conference on quality of multimedia experience (QoMEX)*, pp. 1–6, IEEE, 2018.

[17] R. I.-R. BT, "Methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union*, 2002.

[18] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *The international journal of aviation psychology*, vol. 3, no. 3, pp. 203–220, 1993.

[19] J.-Y. Lee *et al.*, "Estimating the simulator sickness in immersive virtual reality with optical flow analysis," in *SIGGRAPH Asia 2017 Posters*, pp. 1–2, 2017.

[20] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.