Augmented Reality User Interface Evaluation

Performance Measurement of Hololens, Moverio and Mouse Input

https://doi.org/10.3991/ijim.v13i03.10226

Alaric Hamacher(✉), Jahanzeb Hafeez, Roland Csizmazia
Kwangwoon University, Seoul, Republic of Korea
stereo3d@kw.ac.kr

Taeg Keun Whangbo
Gachon University, Seongnam, Republic of Korea

Abstract—Recent innovation in the field of Augmented Reality (AR) and Virtual Reality (VR) has brought new devices on the market. The price for consumer products dropped significantly. Many industries see a big future in AR business and applications. The present research focuses on the user input performance of these AR-devices. This paper proposes an evaluation procedure using a server based input interface with a built-in assessment control. The evaluation is performed by test persons exposed to two AR devices: Microsoft Hololens and Epson Move- rio BT-200. A conventional mouse input is used as a benchmark. The assessment reveals a trend of strength and weaknesses of each device and can orient developers to create more optimized AR experiences and improve the user experience.

Keywords—Augmented Reality, Input, Performance

1 Introduction

New generations of AR device emerge at every major trade show. The market seems to advance to maturity and as the base of users increases many industries forecast a successful future in AR services, businesses and applications[1].

While the field of application is constantly increasing, a general problem in AR remains: human interaction. Research has been done to find appropriate, generic and new input methods for AR. These methods include gestures, voice input, trackers, markers or other haptic devices.

Although many methods have been proposed in this sector, not all interfaces proposed by AR devices are as easy and efficient as a mouse. The following can be observed: gestures are imprecise; input is slow; cloud processed voice input causes privacy problems; input methods require training similar to acquiring game skills[2]; input methods require custom content.

Manufacturers generally rely on their own Software Development Kit (SDK)s to provide interaction. Sensors and trackers are often so specific that content needs do be

developed with a target device in mind. Most of the time the experience will be related to a certain manufacturer of technology. Specific aspects of the performance or operation of AR devices have been evaluated for example for the Google Glass project in first experiences for lectures described by Ebner et al. [3].

The observation of these general problems and the underlying thoughts have motivated the present research. The purpose of this paper is to present an approach to a homogenized evaluation method in order to create a scientific assessment of the performance of existing input methods.

The most popular AR devices, Microsoft Hololens and Epson Moverio, have been selected for this research.

2 Previous Research

Most input devices are not specially designed for AR interaction. However some dedicated special devices exist for interaction with AR systems[4]. This section introduces some examples of so-called generic AR interfaces.

2.1 Tangible interfaces with tiles

This method was developed by Poupyrev et al. in early 2000 for AR, when the technology was in its infancy. Although performance of devices was very limited, many applications could already be foreseen [4]. This AR interface relied on a set of tiles. Acting like graphical boards, they could be overlaid by AR with symbols and custom designs. Since AR platforms imply awareness of the environment in form of video scanning, the idea behind this method is to use the mentioned tiles as optical markers. The computer system attached to the AR environment should then perform two tasks:

First, track the markers in the real works and map predefined object on them to give them a design and signification inside the computer application. The second task is to follow the interaction with theses semi-virtual objects. The tiles can be used to trigger actions. Such as for example copy, paste or delete. These tiles could not only serve to interact with one single application, but perform the same task in different applications[5].

2.2 Two handed interface for AR

Szalavári and Gervautz developed a specific AR interface called Two Handed Interface for AR[6]. Similar to the previous example, this interface is tangible. This means it can be interacted with by touch as if it was a real world object. Similar to the tile interface it is also of versatile appearance: the AR system is tracking the object and overlaying it the texture and interface elements that are desired for the interaction. This interface consists of a track pad with a pen pointer. Both are held in front of the view of the AR device, so they can easily be tracked and registered by the computer system. Multiple functions can be assigned to this pair of track pad and pen. Figure 1

shows a simulation in form of a stylus and a tablet. The shape of the interface can be changed as desired to give the user the impression to interact with simple buttons or if necessary with sliders allowing inputting more precise values.

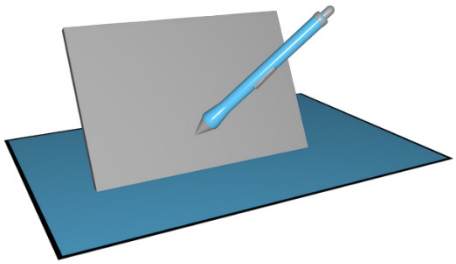

Fig. 1. Reconstruction of two hand interface for AR, Szalavári and Gervautz

The two handed interface for AR is an approach of tracking and mapping. It resembles in VR to the input sticks and trackers provided for the HTC Vive.

Another application using AR devices for tracking markers and hands are described by Menezes in his U-Academy learning modules showing how markers and hands can be used for advanced interaction in AR[7].

2.3 Palm Type

The human hand offers different possibilities for AR applications. Recent research shows how it can be used as an input or to visualize human hand anatomy as described by Boonbrahm et al.[8]. Wang et al.[9] developed a keyboard projected on the users hand to create a virtual input device as shown in figure 2. A similar approach had already been conceived by Dezfuli et al. for a palm based television remote control[10].

Palm Type was originally developed as an enhancement for the Google Glass project, which is similar to the Epson Moverio BT-200. Analog to the previously described methods, Palm Type provides a tangible user interface. In this case it the user's palm with little segments that remind a typewriter key board and using the body as a virtual input surface [11]. However, with some training the users can learn to map this mental keyboard to the lines and knuckles of the palm.

As opposed to the previously shown methods, the authors also perform a series of assessments to evaluate the performance of the new input method. The results are presented on the one hand as a numerical performance value, showing how many words per minute a user will be able to input on such an Palm Type keyboard for VR. On the other hand, the test persons are asked to rate the experience after the assessment. The evaluated is measured on a scale from zero to ten as shown in the results displayed in figure 3.

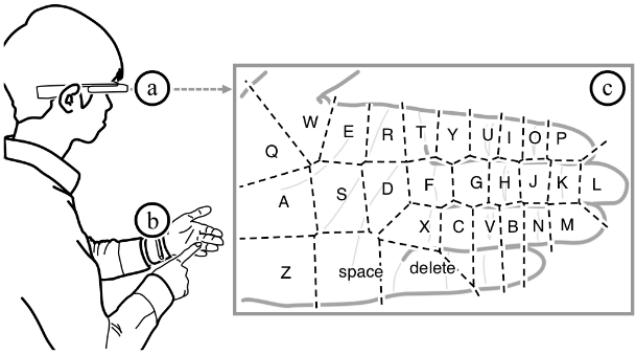

Fig. 2. Palm Type schema

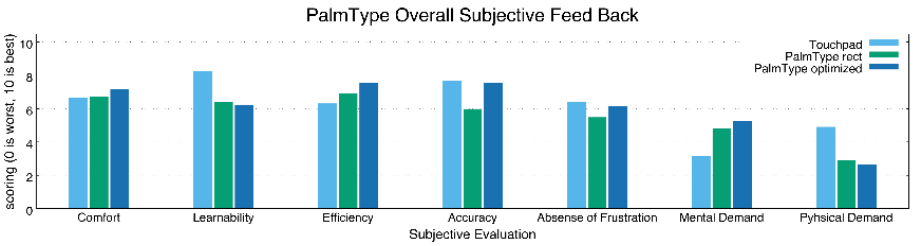

Fig. 3. PalmType subjective evaluation

The numerical results are published using the numerical benchmark Words Per Minute (WPM) ranging between 9.19 and 10.1. If word input is counted for AR devices, it is important to remember that writing performance is usually evaluated in Characters Per Minute (CPM)[12]. This is usually a requirement for typists. Real world values are 200 to 400 characters per minute entered on a keyboard, which corresponds approximately to one hundred WPM.

The work on Palm Type contains two important features that have inspired this research: First, tested devices are compared to an everyday device to set an independent benchmark. In this case, it will be a mouse attached to a laptop. Second, the performance is not only measured numerically, but it is followed by a subjective evaluation to reflect the overall satisfaction.

3 Method

Bach and Scapin state in their research [13] that a single assessment method for measuring Mixed Reality Systems (MRS) does not yet exist. According to the authors, this is due to the following factors: The field of AR is large and specialized. It is not easy to find experts who are competent for all systems. Many limitations lie in the technology itself, not easily to be measurable and traceable. The overall aim of the present assessment is to require as little instruction as possible and to give as much

introduction as necessary. The test persons should not be biased by the operators or the technology.

Therefore, the following measures were taken:

- Random order of experiments
- Instructions integrated in the assessment
- Test persons can run the assessment alone
- Test persons are chosen outside the lab environment

These precautions aim to eliminate most limitations in order to obtain significant evaluations. The following sections describe the different tests and methods that have been developed to perform the assessment.

3.1 Comparing and Benchmarking

The purpose of this research is an evaluation of the objective and subjective performance for AR input devices. For this a series of assessments using three different input methods as shown in figure 4 will be conducted:

(a)·mouse¤ (b)·touchpad¤ (c)·gestures

Fig. 4. Overview inputs methods for performance evaluation

3.2 Performance measurement

The following section describes the assessment user interface and the underlying server technology driving the assessment and collecting the results.

3.3 Assessment interface

The left part of figure 5 shows the interface visible to the user on the different AR devices. The top line (A) contains a small space for the instruction. In case of tasks with timeout, the background of the instruction space can optionally display a progress bar indicating the remaining time. Element (B) is a vertical slider. This element allows selecting values between 0 and 10. The lower part of the interface displays two large buttons. One labeled start (C) to begin the assessment. The other button (D) reads ok and can be triggered when the user has accomplished a task.

The assessment manager sees a different interface: It contains a text-field with the XML-assessment (E). The controls (F) for loading, starting the assessment and selecting the assessment devices. The lower side displays a real time log viewer (G) and a text-field displaying the XML-results (H).

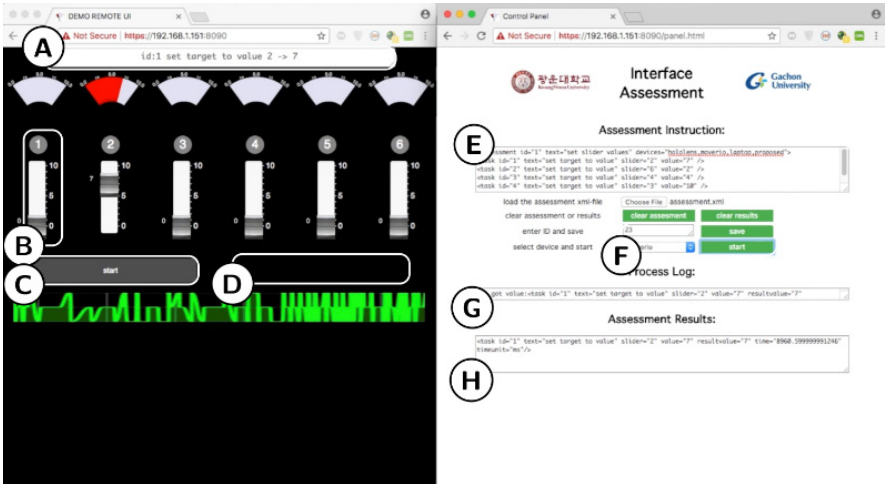

Fig. 5. Assessment interface: for user (left) and for manager (right)

3.4 Assessment server

In order to provide a reusable test environment for a large number of AR devices, the previously described user interfaces are generated by a server equipped with a wireless access point. The AR devices can connect to the access point and display the assessment interface using a web view element or web browser. In the present research server and access point were implemented using a Raspberry PI. This approach has two advantages: First, it assures a unified user experience. Second, it simplifies the assessment creation by not using any platform specific development environments.

3.5 Survey for subjective evaluation

The second part of the assessment consists of a subjective evaluation. Each test person is asked to answer questions and to rank their experience after having completed the tasks on the assessment server.

While many methods exist for such evaluations such as the Likert Scale [14] with a range form one to five or one to seven, these systems aim at identifying the affirmation of a certain hypothesis if form of statements such as: “I strongly agree” or “I strongly disagree”. While this method has many advantages to identify opinions and to reflect test person’s attitude, It is generally difficult to create a mean or to sum up a certain statement, which might even be contradictory [15].

For this reason, the test persons are asked to evaluate certain factors on a linear decimal scale ranging from zero to ten for these categories:

- Overall comfort (easiness to wear)
- Learning (effort necessary to learn operation)
- Efficiency (evaluation after own testing)
- Precision (evaluation after own testing)
- Frustration (description of level during testing)
- Mental demand (description of level during testing)
- Physical demand (description of level during testing)

The lower number describes a negative or uncomfortable experience. The higher number describes a positive or comfortable experience. The test persons are given unlimited time after the experiment to fill out a questionnaire for the survey.

Table 1. Assessment period

Phase	Start	End	Persons
Pretest	0	0	7
First Session	0	0	17
Second Session	0	0	10

4 Results

This section presents the results of survey and assessment. The first part describes composition and structure of the samples. The second part exposes results and subjective evaluation.

4.1 Overview and demography

The head of the questionnaire for the subjective evaluation includes general demographic information and an identification number to relate the subjective evaluation with the results measured by the assessment server. Most of the test persons are students in Seoul, South Korea. Other participants are partly teaching stuff, researchers and students whose major is media, converged software or information contents. Table 1 shows time ranges and sample amount of each assessment.

The assessment period took place between end of May and beginning of June in 2018. A series of seven pretests were conducted on 2018 May 18. These pretests had the purpose to identify ambiguities in the questionnaire and to optimize the survey. Tests on the user interface helped to identify problems and to improve the assessment process. The principal test session took place may 24 and 25 in the VR Medial Lab of the Kwangwoon University. The duration of each assessment was approximately 30 minutes for each device. Assessment assistants verified the proper functioning of the equipment. The test person received a brief introduction how to operate each device: HoloLens, Moverio BT-200 and a laptop with an ordinary office mouse served as configuration for benchmarking.

The first series of tests run for 18 persons without any timeout for the participants. They had all the necessary time to perform all the required tasks until they judged it completed. A second series of tests was performed on additional group of ten test persons on June 7 and 8. These test persons received the same questionnaire, but had to perform the tasks with a specific timeout for each device.

The order in which the participants assessed each device was chosen randomly in order to exclude this factor's influence on the participant's performance. Participants have not been selected by any criterion but accepted as a random group. Their participation was voluntary. The total number of participants is 27. Six of them were females, 21 were males. Table 2 shows the gender composition.

Table 2. Assessment participant gender

Gender	Amount
Female	6
Male	21
Total	27

4.2 Responses without timeout

The first set of results, as displayed in table 3, shows the mean interaction for each sample and device in absolute time in milliseconds. A resemblance in the pattern can be observed for all the samples: HoloLens has on average the longest response time, the mouse is in most cases the fastest input device.

The original aim to account for errors in the input methods, seemed biased by the fact that the test persons generally take as much time as needed in order to complete a task without any error. Table 4 reveals the mean and median response time for each device.

Table 3. Mean Response Time per Sample and Device in ms

Nr	Hololens	Moverio	Mouse
1	15375	8040	4690
2	10285	9858	4179
3	8544	8486	4100
4	17826	9121	3079
5	10050	12088	3894
6	32658	9424	4061
7	17670	7811	4532
8	25788	9711	2931
9	12368	15953	4319
10	15634	9866	3769
11	17925	24214	5096
12	17925	13252	3700
13	16364	8472	5010
14	7900	12180	3024
15	37937	9300	2621
16	24061	9545	3482
17	10756	21802	5300

4.3 Responses with timeout

The second set of samples was obtained by using the mean response time of each device as a timeout for the assessment. The second set of samples shows a much smaller deviation compared to the first one. Table 5 shows the mean response time and deviation of the first session without timeout and the second session with timeout.

The deviation of the response time remains much closer to the mean on all devices. Some people are still faster than the average limited by the timeout. Even if it is much smaller, the variance between test persons with fast and slow reaction times is the largest on HoloLens, and the smallest when using the mouse.

Table 4. Mean and Median Response Time per Device Data in ms

	Hololens	**Moverio**	**Mouse**
Response Time (median)	16364	9711	4061
Response Time (average)	17592	11713	3987

Table 5. Mean Response Time per Device and Standard Deviation in ms Data

Devices	**Mean without Timeout**	**SDV**	**Mean with Timeout**	**SDV**
Hololens	17592	8368	11867	2595
Moverio	11713	4754	10361	1473
Mouse	3987	793	3851	212

4.4 Subjective evaluation results

While most of the assessment performance could be measured in interaction and response time by implementation of an assessment server inside the user interface, the subjective evaluation occurred with no time constraint and requested the test person to rank their experiences.

Figure 6 shows the result of the subjective evaluation among all test subjects after the assessment. The result show some obvious effects: First, there seems to be a general order in all the categories, attributing the best properties to mouse as input interface followed by Moverio and eventually HoloLens. Second: While some device input methods are evaluated below the average, no input method is really evaluated with zero points. In all cases, the mouse as interactive input seems to represent the ideal case ranging in almost all cases at the top with eight or nine in average. Table 6 shows the subjective evaluation overview data.

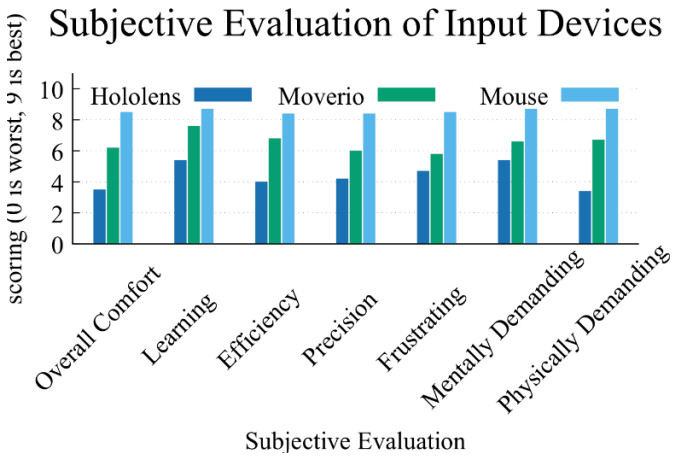

Fig. 6. Subjective evaluation results, all devices cumulative

Regarding the overall comfort, me mouse ranks at the top of the evaluation. Hololens is evaluated 3.5 as the most uncomfortable among the tested devices.

The results regarding learnability show HoloLens ranks last one on this category, meaning that most test persons considered it the hardest to learn how to correctly interact with the device.

Efficiency is evaluated below average on HoloLens with a score of 4.0. While the mouse is evaluated the most efficient input device with a score of 8.4, Moverio ranks above the average.

Regarding precision, all AR devices are evaluated less precise than the mouse. Among the AR devices, HoloLens ranks the lowest.

Table 6. Mean subjective rating overview, samples N = 27

question	Hololens	Moverio	Mouse
Overall Comfort	3.5	6.2	8.5
Learning	5.4	7.6	8.7
Efficiency	4	6.8	8.4
Precision	4.2	6	8.4
Frustrating	4.7	5.8	8.5
Mentally Demanding	5.4	6.6	8.7
Physically Demanding	3.4	6.7	8.7

The level of frustration is comparable to the previous scores. The test persons evaluate HoloLens with a score of 4.7, which is ranks in a neutral region of frustration. All other devices seem less frustrating to use.

The scores regarding mental and physical demand required for the operation of the device show again that the mouse ranks the highest. With a score of 3.4 HoloLens has the lowest ranking in physical demand, which means that the test persons judged it to require the most efforts in order to operate the device properly.

4.5 Future research

The present research shows a trend in the performance of the tested AR input devices. Future assessments should increase the amount of samples in order to gain a higher significance regarding operation and subjective results. The subjective evaluation was conducted after the practical assessment. Future research should include an additional questionnaire to measure the expectations of the user before the assessment. This would allow to draw additional conclusions toward expected and real performance.

5 Conclusion

The present research has proposed a method for objective and subjective performance evaluation using an assessment server in combination with user surveys. The present research indicates that performance and satisfaction of the contemporary AR devices far from being satisfactory. Although the technology makes big steps forward, assessment metrics indicate that there is a need for further improved human input devices. New input methods, such as gestures or touch devices emerge in AR, however most of them are ranked far behind traditional input methods such as the mouse. This research shows that performance of human input interfaces in AR still has large room for improvement of overall performance, satisfaction and user comfort.

6 Acknowledgement

The present research has been conducted by the Research Grant of Kwangwoon University in 2018.

7 References

[1] Francesca Bonetti, Gary Warnaby, and Lee Quinn. Augmented reality and virtual reality in physical and online retailing: A review, synthesis and research agenda. In Augmented Reality and Virtual Reality, pages 119–132. Springer, 2018. https://doi.org/10.1007/978-3-319-64027-3_9

[2] Valerie J Shute and Fengfeng Ke. Games, learning, and assessment. In Assessment in game-based learning, pages 43–58. Springer, 2012. https://doi.org/10.1007/978-1-4614-3546-4_4

[3] Markus Ebner, Herbert Mühlburger, and Martin Ebner. Google glass in face-to-face lectures-prototype and first experiences. International Journal of Interactive Mobile Technologies (iJIM), 10(1):27–34, 2016. https://doi.org/10.3991/ijim.v10i1.4834

[4] Ivan Poupyrev, Desney S Tan, Mark Billinghurst, Hirokazu Kato, Holger Regenbrecht, and Nobuji Tetsutani. Developing a generic augmented-reality interface. Computer, 35(3):44–50, 2002. https://doi.org/10.1109/2.989929

[5] Ivan Poupyrev, Desney S Tan, Mark Billinghurst, Hirokazu Kato, Hol- ger Regenbrecht, and Nobuji Tetsutani. Tiles: A mixed reality author- ing interface. In Interact, volume 1, pages 334–341, 2001.
[6] Zsolt Szalavári and Michael Gervautz. The personal interaction panel– a two-handed interface for augmented reality. In Computer graphics forum, volume 16. Wiley Online Library, 1997.
[7] Paulo Menezes. An augmented reality u-academy module: From basic principles to connected subjects. International Journal of Interactive Mobile Technologies (iJIM), 11(5):105–117, 2017. https://doi.org/10.3991/ijim.v11i5.7074
[8] Poonpong Boonbrahm, Charlee Kaewrat, Presert Pengkaew, Salin Boonbrahm, and Vincent Meni. Study of the hand anatomy using real hand and augmented reality. International Journal of Interactive Mobile Technologies (iJIM), 12(7):181–190, 2018. https://doi.org/10.3991/ijim.v12i7.9645
[9] Cheng-Yao Wang, Wei-Chen Chu, Po-Tsung Chiu, Min-Chieh Hsiu, Yih-Harn Chiang, and Mike Y. Chen. Palmtype. Proceedings of the 17th International Conference on Human- Computer Interaction with Mobile Devices and Services - MobileHCI '15, 2015. https://doi.org/10.1145/2785830.2785886
[10] Niloofar Dezfuli, Mohammadreza Khalilbeigi, Jochen Huber, Murat Özkorkmaz, and Max Mühlhäuser. Palmrc: leveraging the palm sur- face as an imaginary eyes-free television remote control. Behaviour & Information Technology, 33(8):829–843, 2014. https://doi.org/10.1080/0144929X.2013.810781
[11] Joanna Bergstrom-Lehtovirta, Sebastian Boring, and Kasper Hornbæk. Placing and recalling virtual items on the skin. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pages 1497–1507. ACM, 2017. https://doi.org/10.1145/3025453.3026030
[12] Scott I MacKenzie. Kspc (keystrokes per character) as a characteristic of text entry techniques. In International Conference on Mobile Human- Computer Interaction, pages 195–210. Springer, 2002.
[13] C´edric Bach and Dominique L Scapin. Obstacles and perspectives for evaluating mixed reality systems usability. In Acte du Workshop MIXER, IUI-CADUI, volume 4. Citeseer, 2004.
[14] Elaine Allen and Christopher Seaman. Likert scales and data analyses. Quality progress, 40(7):64, 2007.
[15] Ron Garland. The mid-point on a rating scale: Is it desirable. Marketing bulletin, 2(1):66–70, 1991.

8 Acknowledgement

The present research has been conducted by Research Grant of Kwangwoon University in 2018.

9 Authors

Alaric Hamacher is professor for 3D Contents and Virtual Reality in the Graduate School of Smart Convergence, Kwangwoon University. Alaric Hamacher graduated in directing and producing from the Academy for Television and Cinema Munich and

holds a MA in Film Sciences from Paris VII. He directed stereo 3D on many 3D commercials and corporate movies. His present research focusses on Augmented and Virtual Reality, 360VR.

Roland Attila Csizmazia is professor for statistical programming and office automation at Kwangwoon University. Currently, he is working on his dissertation at Korea University in Industrial Management Engineering.

Jahanzeb Hafeez currently works at the Graduate School of Smart Convergence, Kwangwoon University. Jahanzeb does research in Close Range photogrammetry, Structure-from-motion, Engineering and Medicine, Computer Graphics and Algorithms.

Taeg Keun Whangbo received the M.S. degree from City University of New York in 1988 and the Ph.D degree both in Computer Science from Stevens Institute of Technology in 1995. Currently, he is a professor in the Department of Computer Science, Gachon University. He was also the researcher in Samsung Electronics from 2005 to 2007. His research areas include Computer Graphics, Deep Learning and AR/VR.

Article submitted 2019-01-29. Resubmitted 2019-02-24. Final acceptance 2019-02-24. Final version published as submitted by the authors.