Augmented Reality Evaluation: A Concept Utilizing Virtual Reality

Philipp Tiefenbacher, Nicolas H. Lehment, and Gerhard Rigoll

Institute for Human-Machine Communication,
Technische Universität München, Germany
{Philipp.Tiefenbacher,Lehment}@tum.de
http://www.mmk.ei.tum.de

Abstract. In recent years the field of augmented reality (AR) has seen great advances in interaction, tracking and rendering. New input devices and mobile hardware have enabled entirely new interaction concepts for AR content. The high complexity of AR applications results in lacking usability evaluation practices on part of the developer. In this paper, we present a thorough classification of factors influencing user experience, split into the broad categories of rendering, tracking and interaction. Based on these factors, we propose an architecture for evaluating AR experiences prior to deployment in an adapted virtual reality (VR) environment. Thus we enable rapid prototyping and evaluation of AR applications especially suited for applications in challenging industrial AR projects.

1 Introduction

The AR technology has advanced rapidly over the last years and the number of real world applications increases. Industrial AR applications have always been targeted, but until today only partly succeeded. The development of AR applications for industrial settings is challenging. The development process of industrial plants is complex and in general years pass until the plant is reality. AR applications, on the other side, need to be adapted to the exact use-cases to yield real value. Therefore, we present a mixed reality (MR) environment, which enables early development of AR applications through the visualization of the CAD data of the industrial line in the virtual reality (VR). Besides the benefits of an earlier development, the isolated and controlled environment of the VR allows advanced user evaluation for AR applications.

A survey by Dunser *et al.* showed that only 8 % of all considered publications in AR conducted user evaluations. Swan *et al.* conclude in [17] that there is still space to identify proper user interfaces and user interaction requirements to known usage domains, which is reflected by the rising importance of evaluation in AR publications [6,21]. We therefore see a high potential in frameworks, which enable thorough evaluation of AR systems under a wide range of conditions. In this paper we propose a general concept for evaluating AR applications, which takes advantage of the controlled conditions afforded by full VR systems.

R. Shumaker and S. Lackey (Eds.): VAMR 2014, Part I, LNCS 8525, pp. 226–236, 2014.

Hereby, the system is able to simulate Head-Mounted Displays (HMD) as well as to integrate mobile device like smart phones. We feel that the success of AR depends on the acceptance by the users and aim to improve said acceptance by providing a reliable and reproducible test bench for future AR applications. Section 2 discusses current work in the field of mixed reality evaluation concepts. Then, Section 3 presents the architecture of the evaluation bench with the corresponding parameters. Also the CAVE is described shortly. The following Section 4 presents a variety of different properties, which affect the AR experience and should be evaluated. Lastly, Section 5 summarizes the work and gives an outlook.

2 Related Work

Khan *et al.* describe in [11] a CAVE setting for evaluating the intrusiveness of virtual ads on a smart phone. Here, the CAVE solely supports the evaluation of the user experience. Furthermore, only the head of the subjects are tracked and based on the position of the user, virtual content is displayed on the smart phone. The mobile device itself is not tracked.

First approaches investigating the influence of tracking failures in AR scenes like jitter are presented in [15,19]. Vincent *et al.* test in [19] the influence of jitter based on artificial normal distributed noise. Ragan *et al.* provide in [15] a proof of concept for simulating an AR system in a virtual environment, the experiment results show a significant influence of jitter to the task completion time.

Besides the evaluation of the tracking accuracy also approaches for evaluating the impact of the latency in AR scenes exist [13]. VR systems have also been used for simulating outdoor AR systems. Gabbard *et al.* present in [7] an AR scene in a CAVE, the users wore an optical see-through display, which showed virtual objects registered within the VR environment. Here, different designed virtual texts overlay heavily textured outdoor scenes. In the controlled environment of the CAVE, the users chose the best recognizable text designs. Another work about the human perception of AR scenes was done by Knecht *et al.* in [12]. Knecht focused on the influence of rendering global illumination for augmented objects to allow for photo-realistic augmented reality scenes. The different types of rendering did not change the completion time for a positioning task of a virtual cube.

Furthermore, Lee *et al.* compare in [14] different levels of visual quality for searching tasks in a mixed reality (MR) environment. The results show that the completion time of most searching tasks are independent of the visual quality of the rendering part of the MR. Also the completion times of the same task in MR and AR are not significantly different, which further motivates our approach.

3 Evaluation Architecture

Our main contribution is an overview of experience criteria for AR applications, which are depicted in the evaluation concept called Augmented Reality Evaluation (ARE). Figure 1 shows this concept. ARE works as a link between the physical hardware to any AR application.

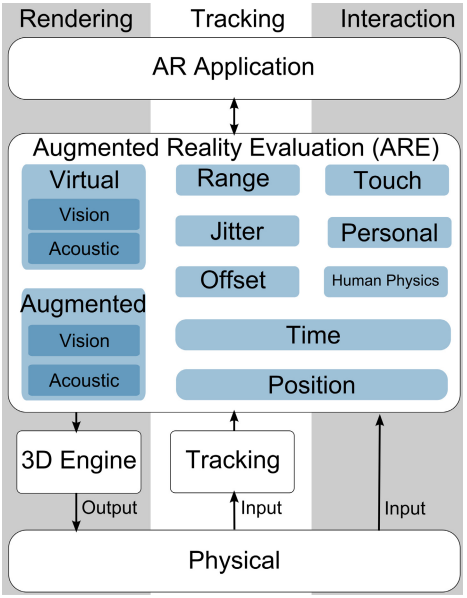

Fig. 1. The test bench separated into rendering, tracking and interaction. Each part is subdivided into possible evaluation criteria.

It is implemented as a CAVE virtual reality (VR) system, which in turn simulates the entire AR experience and partly real hardware. This constrained environment enables complete control of AR related parameters. We separate these parameters into the three different main parts of any AR system [3]: rendering, tracking and interaction. Azuma defined these parts in [3] as *scene generator*, *tracking and sensing* and *display device*. We extend the definition of the displaying part to the interaction part, as the display is generally more than just a visualization. Commonly the touch display is also used as interaction device for the AR scene. In the case of a HMD, the interaction part also may include gesture or speech recognition. The computational unit is neglected as it is included in every mentioned main part.

The physical layer includes all the necessary hardware for the MR environment like the canvas, projectors, tracking targets and PCs. The rendering encompasses both the stimulation of the real world and the augmented reality content, which is just visible in the field of view of a Head-Mounted Display or on the surface of a tablet PC, as demonstrated in Figure 3. Section 4.1 describes the quality metrics of the rendering part in more detail. A professional tracking system (DTrack), which works on the basis of infrared cameras and reflecting tracking targets, delivers precise tracking data with a rate of 60 Hz to our ARE concept. The tracking results of the DTrack system are altered according to the five depicted properties in Figure 1. Then ARE forwards this modified tracking data to an arbitrary AR application. The subject's view is tracked using glasses with

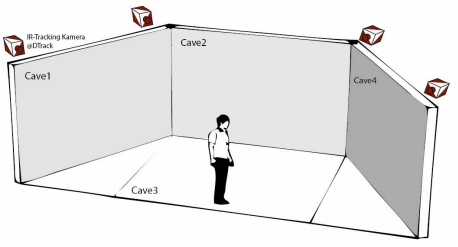

Fig. 2. Scheme of the four sided CAVE and the setup of the tracking cameras. Each of the four walls is projected through two *projectiondesign* projectors. The projectors are equipped with *Infitec* filters, which enable a 3D experience.

tracking targets mounted on it. The position of the head defines the rendered scene in the CAVE. When simulating a virtual HMD display, an additional camera with an arbitrarily defined field of view renders the contents of the AR scene. Besides the simulated HMD also physical mobile devices can be included. For this, tracking targets are also equipped at the mobile device. In this case, the tracking system registers both the head of the subject and the mobile device. The mobile device has to render the virtual scene in the CAVE, as well as the additional AR content. Thus, the mobile device has to fulfil high performance requirements. The smaller field of view of the mobile device, however, limits the virtual space which has to be rendered. Lastly, the interaction with the AR scene is also split into five different properties. The simulated HMD has currently no input options, consequently the only way for the subject to directly interact with the scene would be through gesture or/and speech recognition using a Wizard-of-Oz approach. Nevertheless, AR tasks like searching for 3D content can be performed and the user experience evaluated.

Fig. 3. A user with a tracked mobile device stands inside a virtual industrial line, which is also visible on the mobile device. The user interacts with additional AR content only accessible on the touch device. In this scenario the head of the subject, as well as the mobile device are tracked.

4 Evaluation Properties

4.1 Properties for Rendering

The side effects in the virtual environment are very limited, enabling to simulate real world conditions with respect to actual user perception. This can be achieved by rendering the virtual scene in different ways, imitating possible limitations and noise of the real environment. For instance, the influence of changing outdoor illumination as evaluated in [7] can easily be tested within a controlled environment. For a start, we propose four different environmental properties, which might influence the visual AR experience and are easily reproducible in our concept:

- available light
- background texture
- disruptive visual environment (dust, powder)
- acoustic noise

The first three items are part of the visual rendering property, the last belongs to the acoustic property of previous Figure 1. The available light includes the brightness of the virtual scene as well as the number and the spots of light sources. In a plant or an other indoor environment, it is very likely that more than one light source brightens the setting.

The background texture can be changed from very simple, flat ones to more rich ones, which is similar to the idea in [7]. It should be noted that, our main goal is not an easier reproducibility of outdoor scenes in the VR simulation. The influence of disruptive visual environment shares the same features as the background textures, however this time the distractions are of a more dynamic nature and overlays the background textures. Finally, the acoustic noise in the VR rendering part provides an additional disturbance factor to the task. Hereby, the noise can surround the subject or be perceived in a more directional fashion, where the source of the noise comes from a distinct direction. The Head-related transfer function for the audio output, however, is not implemented yet.

The type of noise should fit to the presented scene in the CAVE. In our case an industrial plant is visualized and typical sounds of heavy machinery are the best choice.

The camera of the portable device, which displays the augmented reality content is purely virtual. Hence, we can control the following camera parameters of the AR rendering part:

- rendering latency
- rendering quality
- focal length
- field of view
- aperture

A holding parameter defines the update rate of the AR content, which only affects the user experience in the case of animated 3D items. The quality of the

3D items has to be detailed offline. The basis of the 3D items is CAD data, thus through reducing the faces and vertices, the rendering quality of the 3D items can be decreased from very realistic to sketchy. A volumetric lens renders not the whole 3D data as we are culling specific parts of the 3D items [18] based on the view frustum. A view frustum is a geometric shape similar to a pyramid. The position of the pyramid's peak specifies the position of the lense. The shape of the volume defined through the view frustum is rigid, so moving the viewer's head has no affect on the volume of the frustum. The focal length is determined by setting the near and far plane for the view frustum, whereas the field of view can easily be changed through setting the left, right, top and bottom of the view frustum. The brightness of the light in the virtual scene defines the aperture of the virtual camera.

4.2 Properties for Tracking

Tracking is a crucial part in any AR scenario as bad content registration impairs user experience and task fulfilment. The precise tracking available in a virtual environment allows for measuring the influence of important performance metrics of a tracking system. In our test bench approach, we can alter four different tracking metrics and evaluate their impact on user experience. These metrics are:

- update rate
- precision
- jitter (spatial and temporal)
- necessary range of tracking (see Figure 4)

As stated above, the tracking system delivers new tracking data with a frequency of 60 Hz, which also defines the upper threshold of the update rate. Based on this maximum rate, slower update rates can be simulated by emitting the same tracking data for multiple new inputs. At this step, an additional temporal random noise can be added to the update rate. The random jitters of both the spatial and temporal jitter are Gaussian distributed. The temporal jitter is just added to the altered update rate.

The precision of the tracking data can be separated into the rotational precision and the translational precision. Therefore, two additional sources of noise may influence the tracking outcome respectively. First, the translational position is modified. For that, a fixed translational offset vector $\mathbf{T_o}$ is subtracted from the original position of the object $\mathbf{X}$

$$\mathbf{X_o} = \mathbf{X} - \mathbf{T_o}. \tag{1}$$

Hereby, $\mathbf{T_o}$ holds the offset for each axis. Then, the spatial jitter is added to each axis.

Huynh *et al.* analysed different metrics for 3D rotation in [5]. We choose an easy to calculate metric, which overcomes the problem of ambiguous representation and is also bi-variant, when calculating a metric for the distance between

two rotations. Therefore, the proposed metrics in [5] cannot be used for determining a new rotation matrix based on such a distance metric.

Hence, the three Euler angles are separately defined based on the Equation

$$\alpha_o = \alpha - r_o. \tag{2}$$

The offset r_o to the current angle ranges between $-\pi \leq r_o \leq \pi$. Finally, the random rotational jitter is added to each angle. For evaluation purposes a rotational distance metric m_r between the final rotation matrix $\mathbf{R}_o$ and the original rotation $\mathbf{R}_x$ is computed with Equation (3).

$$m_r = ||\mathbf{I} - \mathbf{R}_o \mathbf{R}_x^T||_F \tag{3}$$

Here, $\mathbf{I}$ denotes the identity matrix. The *Frobeniusnorm* is defined as follows

$$||\mathbf{R}||_F = \sqrt{\sum_i \sum_j R_{i,j}^2}. \tag{4}$$

The distance of the altered rotation to the original rotation changes with every update of the tracking data. The reason therefore lay in the fact that for each update a fix offset r_o is added but also random noise (jitter). So the metric m_r records the changes in the rotational distance. Lastly, the tracking system follows the targets not in the whole area of the CAVE. An upper and lower threshold for each axis determine the valid area for delivering valid tracking data.

Marker based tracking is still widely used as it leads to quite accurate tracking and also to little performance requirements. A drawback of marker based tracking is the limited range of available tracking, as in each frame a fiducial marker must be recorded and detected. Furthermore, when working at an industrial plant, markers can only attached at certain spots on the machine. Either way the tracking is limited to specific regions. Thus, we also restrict the range of the tracking to examine the influence to certain AR applications in the mixed reality environment. In the best case, we are able to recommend a necessary tracking area to fulfil a certain AR task, like maintaining a part of the machine in a satisfactory manner.

Figure 4 shows the favoured positions of 18 different subjects in bird's-eye view of a study. Here, the subjects had to interact with three AR items in the CAVE.

4.3 Properties for Interaction

Now that we can control experimental conditions, the interaction metrics can be exercised. Widely used questionnaires are the System Usability Scale [4] and the NASA TLX [9] for more challenging tasks. Our system, however, can gather a host of additional data, such as:

- amount of touch points
- area and range of touch points (see Figure 5)

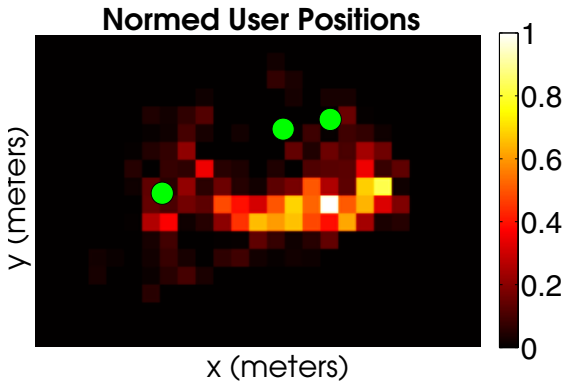

Fig. 4. The heat map of positions displays the favoured positions of the subjects during an AR task. Here, the green dots describe the location of AR models. Brighter rectangles indicate the favoured positions of the participants during the whole test.

- human physics
- time (task, interaction)

Human physiology need to be taken to account as in the one hand the interaction happens in real places, which sometimes might be hard to reach. On the other hand the human biometrics like the size of the hand may influence the usability of an AR application according to different sized touch devices.

In first studies, we concluded that in some cases it is beneficial to separate the completion time into the whole task time and the interaction time. The interaction time is just the time, in which the subjects are really touching and interacting with the scene. In the case of evaluating different interaction concepts for an AR scene, the whole completion time is additionally influenced by repositioning tasks of the subjects. The difficulty of a certain task also influences the completion time, as the participants have to find the right way to solve the task. This affects the study outcome, when performed as within-subject design as it is hard to counterbalance the training effects. The consideration of just the interaction time reduces such training effects. Figure 5 depicts an heat map of exemplary touch points on the touch-screen of the tablet. Here, a lot of touch points are indicated in the middle of the screen, thus the human hands need to be big enough to perform touch events in these area easily.

5 Detailed Properties vs. Evaluation Categories

The ARE architecture features the parts initially proposed by Azuma, however, there exists also a classification of different types of evaluation concepts of AR systems. In the following the rendering, tracking and interaction part are linked to these new types of classification. Dunser *et al.* classified in [6] four different types of evaluation concepts for AR systems. Three of them were initially introduced by Swan *et al.* in [17].

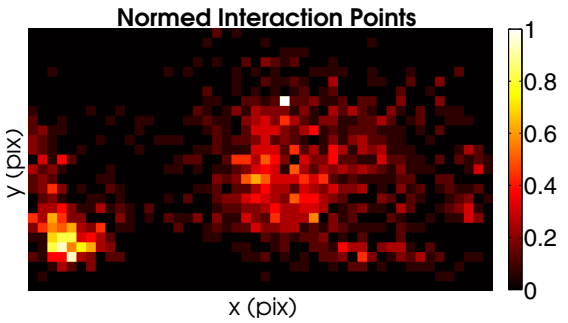

Fig. 5. The heat map of touch points classifies the positions on the screen in regard to the total number of clicks. Brighter regions received more touch clicks than darker regions.

Experiments determining how the subjects perceive and realize the AR context belong to the *Perception* type. The second type of evaluation is classified as *Performance*. These experiments mainly examine the users' task performance with the goal to improve the task execution through the help of AR. The *Collaboration* experiments detail the interactions in an AR scene between multiple users. The collaborative AR can be split into face-to-face [1,10] and to remote collaboration [8,20]. The new category introduced in [6] is called *Interface or system usability studies* abbreviated as *Usability*. This type does not need to involve the measurement of the task performance, instead the user experience is identified. Figure 6 combines the evaluation criteria with the three components

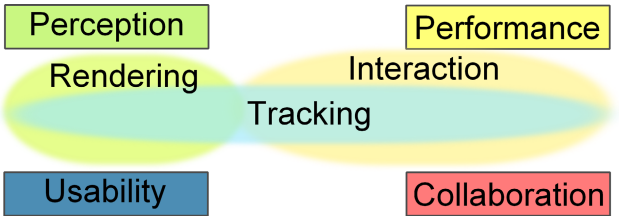

Fig. 6. Relation of the three parts every AR systems consists of to the evaluation criteria in [6]

every AR system consists of. The rendering part of an AR scene mainly belongs to the *Perception* category. Since the proposed properties in the rendering part affect primarily the cognition of the scene.

The interaction with the AR scene and the quality of the tracking are measured as the *Performance* of the task. The AR task mostly depends on the featured interaction idea, when a certain quality in tracking is guaranteed. Photo-realistic rendering, calculation of lightning effects [12] and the visualization of occlusion [2]

improve the user experience (*Usability*). Therefore, the rendering part plays an important role. Shah *et al.* show in [16], however, that neglecting occlusions in AR scenes lead to incorrect display, which might be noticed by the user. So, the users' perception might be wrong in some cases, which lead to wrong task operations and might also influence the performance.

Lastly, when working with multiple user's in an AR scene, the interaction parts within the group are of main interest (*Collaboration*).

6 Conclusion and Outlook

Inspired by the definition of the main AR parts by Azuma, we envisaged individual evaluation metrics for each part on the basis of a mixed reality environment. The detailed description of evaluation criteria for AR applications gives other researchers a guideline for their AR evaluations. We use a CAVE setting for the evaluation architecture as AR applications can be evaluated more rapidly and easily in VR scenes than in real scenes.

The use of a mixed environment, however, also implies some shortcomings. The area for the AR scene is limited to a desktop scenario due to the size of the CAVE. Testing AR applications in wide area scenarios is not possible.

Furthermore, the use of projectors for the presentation of the virtual scene restricts the maximum intensity of ambient light. A brighter room reduces the quality of the virtual experience as the projectors have a limited luminance. So, the experiments can only be conducted in an almost dark room.

Beside this, the evaluation architecture can be advanced to include photo realistic rendering or occlusion aware rendering of the AR objects, which is currently not integrated. The forth evaluation category, *Collaboration*, is also not included, yet. Hence, face-to-face as well as remote scenarios for collaborative AR have still to be incorporated in the CAVE environment.

Acknowledgement. The research leading to these inventions has received funding from the European Union Seventh Framework Programm (FP7/2007-2013) under grant agreement n° 284573.

References

1. Morrison, A., Mulloni, A., Lemmelä, S., Oulasvirta, A., Jacucci, G., Peltonen, P., Schmalstieg, D., Regenbrecht, H.: Collaborative use of mobile augmented reality with paper maps. Computers & Graphics 35(4), 789–799 (2011)
2. Allen, M., Hoermann, S., Piumsomboon, T., Regenbrecht, H.: Visual occlusion in an augmented reality post-stroke therapy scenario. In: Proc. CHINZ. ACM (2013)
3. Azuma, R.: A survey of augmented reality. Presence 6(4), 355–385 (1997)
4. Brooke, J.: Sus-a quick and dirty usability scale. Usability Evaluation in Industry 189, 194 (1996)
5. Du Huynh, Q.: Metrics for 3d rotations: Comparison and analysis. Mathematical Imaging and Vision 35(2), 155–164 (2009)

6. Dünser, A., Grasset, R., Billinghurst, M.: A Survey of Evaluation Techniques Used in Augmented Reality Studies. Technical report (2008)
7. Gabbard, J.L., Swan, J.E., Hix, D., Lucas, J., Gupta, D.: An empirical user-based study of text drawing styles and outdoor background textures for augmented reality. In: Proc. VR, pp. 11–18, 317. IEEE (2005)
8. Gauglitz, S., Lee, C., Turk, M., Höllerer, T.: Integrating the physical environment into mobile remote collaboration. In: Proc. MobileHCI, pp. 241–250. ACM (2012)
9. Hart, S.G., Staveland, L.E.: Development of nasa-tlx (task load index): Results of empirical and theoretical research. Human Mental Workload 1(3), 139–183 (1988)
10. Kaufmann, H., Schmalstieg, D.: Mathematics and geometry education with collaborative augmented reality. Computers & Graphics 27(3), 339–345 (2003)
11. Khan, V., Nuijten, K., Deslé, N.: Pervasive application evaluation within virtual environments. In: Proc. PECCS, pp. 261–264. SciTePress (2011)
12. Knecht, M., Dünser, A., Traxler, C., Wimmer, M., Grasset, R.: A framework for perceptual studies in photorealistic augmented reality (2011)
13. Lee, C., Bonebrake, S., Hollerer, T., Bowman, D.A.: The role of latency in the validity of ar simulation. In: Proc. VR, pp. 11–18. IEEE (2010)
14. Lee, C., Rincon, G.A., Meyer, G., Hollerer, T., Bowman, D.A.: The effects of visual realism on search tasks in mixed reality simulation. Visualization and Computer Graphics 19(4), 547–556 (2013)
15. Ragan, E., Wilkes, C., Bowman, D.A., Hollerer, T.: Simulation of augmented reality systems in purely virtual environments. In: Proc. VR, pp. 287–288. IEEE (2009)
16. Shah, M.M., Arshad, H., Sulaiman, R.: Occlusion in augmented reality. In: Proc. ICIDT, vol. 2, pp. 372–378. IEEE (2012)
17. Swan, J.E., Gabbard, J.L.: Survey of user-based experimentation in augmented reality. In: Proc. VR, pp. 1–9. IEEE (2005)
18. Viega, J., Conway, M.J., Williams, G., Pausch, R.: 3d magic lenses. In: Proc. UIST, pp. 51–58. ACM (1996)
19. Vincent, T., Nigay, L., Kurata, T.: Handheld augmented reality: Effect of registration jitter on cursor-based pointing techniques. In: Proc. IHM, pp. 1–6. ACM (2013)
20. Shen, Y., Ong, S.K., Nee, A.Y.C.: Augmented reality for collaborative product design and development. Design Studies 31(2), 118–145 (2010)
21. Bai, Z., Blackwell, A.F.: Analytic review of usability evaluation in ISMAR. Interacting with Computers 24(6), 450–460 (2012)