Eurographics Workshop on Visual Computing for Biology and Medicine (2014)
I. Viola, K. Bühler, and T. Ropinski (Editors)

A Comparative User Study of a 2D and an Autostereoscopic 3D Display for a Tympanoplastic Surgery

A. Baer[1], A. Huebler[1], P. Saalfeld[1], D. Cunningham[2] and B. Preim[1]

[1]Department of Simulation and Graphics, Universiy of Magdeburg, Germany
[2]Institut of Graphical Systems, University of Cottbus, Germany

Abstract

This paper presents the design and execution of a comparative experimental between-participant study with 42 participants. We investigated depth perception comparing a 2D display with a glasses-free 3D autostereoscopic display in detail and conducted a follow-up study with the new 3D zSpace technology including a stylus as input device. This work comprises the design of a tympanoplastic training scenario used as the study's "real world task". Participants had to position a prosthesis implant to reconstruct the ossicular chain and thus a patient's hearing ability. The study revealed an overwhelming support of the 3D autostereoscopic display compared to a 2D display regarding depth judgment, task completion time and the number of required scene and prosthesis interactions.

Categories and Subject Descriptors (according to ACM CCS): B.4.2 [Input/Output and Data Communications]: Input/Output Devices—Image Display G.3 [Probability and Statistics]: Experimental Design—

1. Introduction

In order for surgeons to navigate through a patient's anatomy, they usually must rely on very accurate depth judgment and spatial orientation abilities. The correct localization of the surgical instruments as well as the identification of relevant anatomical and pathological structures usually demands high perceptual skills. These skills are essential for performing fine dissections, to avoid injuring risk structures, or for the correct position and alignment of implants. In particular, the microscopes or endoscopes used for minimal interventions provide a very restricted field of view, which contains limited (stereoscopic) depth cues and spatial information. Thus, especially surgeons-to-be need a lot of training and trial-and-error experience to gain adequate surgical skills and experiences, e.g., depth judgment and spatial assessment of structure relationships.

During the past decades, various 2D and 3D visualizations and illustration techniques were developed to display segmented structures derived from computer tomography (CT) or magnetic resonance imaging (MRI). These techniques were applied to patient-specific data and focus on the integration of additional depth cues and spatial information to ease and support anatomy exploration [JM07] and navigation, e.g., interactive cutaways [BGKG06], ghosting [GNBP11] or ambient occlusion [ZIK98]. However, most of them were developed and evaluated assuming that the visualizations would be presented using 2D displays.

Beyond the depth cues of visualization techniques, however, stereoscopic cues and motion parallax are the most significant sources of depth information [WFG92]. In- and output devices like stereoscopic monitors, haptic devices or 3D navigators and stylus input devices are developed to enable intuitive 3D visualizations, navigations and interactions with 3D scenes. However, 3D displays are still not accepted in surgery, even though their technology enables binocular vision without specific visualization techniques and thus, should support depth perception. Compared to newly introduced techniques, the viewer does not need any information or introduction to interpret and understand the visualization. When using a 3D display, a simple rendering of structures is sufficient to explore a patient's anatomy. In the past, 3D displays suffered from negative side effects of stereopsis. Viewers either complained about perceptual deficiencies such as the perception of double images or physical effects like nausea and headache [UH07,RB14]. The latest 3D imaging systems provide improved image quality and resolution, similar to 2D monitors. Therefore, 3D displays represent a potential alternative for surgery training systems compared to mani-

DOI: 10.2312/vcbm.20141190

fold developed visualization techniques. Moreover, stereoscopic depth cues provided by the used microscopes can be integrated in a training scenario and thus used to improve the surgical skills.

Although perceptual experiments are increasingly used to improve virtual and augmented environments, the vast majority of findings and techniques from psychophysical studies exclusively use really simple stimuli, e.g., letters, graphical objects such as circle and triangle or colors. These must be adapted to complex structures and scenarios e.g. in a surgery or human anatomy. This paper focuses on the design and execution of a comparative experimental study investigating the effectiveness of 2D versus 3D displays for an otologic training scenario. We compare a 2D display with a glasses-free 3D autostereoscopic display in detail and investigate in a follow-up pilot study a 3D zSpace system (`www.zSpace.com`) using a stylus as input device.

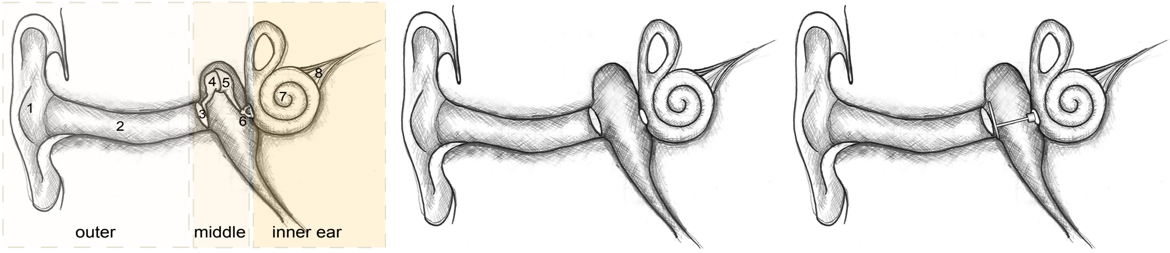
2. Previous Work

Guidelines and findings from psychophysical studies are increasingly being used in computer graphics to improve virtual and augmented environments. In particular, depth and spatial cues have been investigated with respect to how they can enhance the effectiveness of 3D and 2D visualizations [BCFW08]. Several perceptual experiments have examined the efficiency of depth cues such as occlusion, transparency [CWMC09], shadow, fog, shape from shading [BGCP11] and from texture [KHSI04], and depth of field [GSBH13] that contribute to the mental reconstruction of 3D objects from 2D images. Spatial perception studies of anatomic structures were presented by Ritter [RHD*06] and Baer [BGCP11]. Depth judgment tasks were used to analyze the visualization of vascular branches and aneurysm illustrations. Beyond the depth cues applied to 2D images, stereopsis and depth from motion parallax are the most significant mechanisms supporting observers with depth information.

Braun et al. [BLR11] examined the relative efficacy of stereopsis and motion parallax as well as their interaction for depth judgment accuracy. Although they found a relatively low correlation between the presented depth difference and the associated detection frequency, they were able to show that both cues improved accuracy a bit (with stereopsis helping a bit more). Manzey et al. [MRBH*09] presented a survey to 213 surgeons in order to examine the perceived consequences and human factors issues of image-guided navigation. They found issues in increased time pressure and mental demands of inexperienced users, but evaluated navigation and the support of stereopsis and motion parallax as helpful.They conclude that familiarization training with image-guided navigation is very important.

Training for laparoscopic surgery has been the most established application of simulator training and was analyzed in several experimental studies [WRB*09]. Training for otologic surgery is still a new research domain but essential to improve the cognitive surgical skills and therefore the surgery results. Feng et al. [FRH10] compared the impact of a conventional laparoscopic monitor system (2D), a high-definition monitor system (HD), and a stereoscopic display (3D) on an individual's performance in laparoscopic training. Participants had to move surgical instruments to a set of targets using all display systems. The users expressed a strong preference for HD systems. However, the actual quantitative analysis indicates that HD monitors offer no statistically significant advantage and may even worsen the performance compared with standard 2D or 3D laparoscopic monitors. The use of 3D displays in a laparoscopic task was also investigated by a number of researchers [WHK*12, WRK*14, KSK13]. Wagner et al. [WHK*12] compared a 3D display system, the DaVinci robotic system and a 2D display using an eye patch for monocular vision. They tested spatial relationships, grasping, positioning, dexterity, precision, hand-eye, and hand-hand coordination. Their results indicate: the more complex the task is, the more 3D vision accelerates task completion time compared with 2D vision. They state that the gain in task performance is independent of the surgical method. Wilhelm et al. [WRK*14] revealed that a glasses-based 3D display shows the best performances compared to a mirror-based theoretically ideal 3D display, a 2D and a 3D autostereoscopic display, in the order they are listed. Nevertheless, the glasses were subjectively rated as disturbing and autostereoscopic displays were favored. Furthermore, these studies showed that the results for inexperienced surgeons using 3D were in line with the results of experts using 2D.

Stelter et al. [SEWL*11] investigated navigation systems for endoscopic sinus surgery. Navigated operations lasted longer, but all sinuses were found, while in the group without navigation five were not found at all. Recently, Gerber et al. [GBG*14] investigated a surgical planning tool to robotically perform minimally invasive cochlear access. This planning tool supports the definition of landmarks and the safe direct cochlear access trajectory definition. However, trainee surgeons tend to overestimate the possibilities of the system and to underestimate the risks. An extensive training with the navigation system and different pathologies in advance might reduce this overestimation and support the surgical skills. Virtual training improves the accuracy, the time taken to perform a task, and minimizes errors compared to no training [GAPD09]. To design a virtual training scenario, an appropriate 3D training environment is required. Visualization techniques as well as 2D and 3D in- and output devices have to be investigated, to provide an optimal 3D environment with appropriate depth cues. Several studies document the advantages of 3D displays including shutter and polarized glasses or glass-free autostereoscopic systems [BvBKS96,JDG*04]. Nevertheless, their usage and effectiveness regarding depth perception for complex sce-



(a) anatomy (b) conductive hearing loss (c) total ossicular prostheses (TORP)

Figure 1: *(a) The three parts of the human ear. The (1) pinna and (2) auditory canal are the outer ear. (3) Eardrum, (4) mallus, (5) incus, and (6) stapes bone are the inner ear and the (7) cochlear and (8) auditory nerve belong to the inner ear. (b) Conductive hearing loss occurs when the ossicular chain (the movable parts (3)-(6)) is damaged, lost or when its mobility is impaired. (c) Prosthesis implants are used to bridge the gap between (3) and (6) and restore the hearing ability.*

narios and medical domains have to be evaluated, since they are still not part of the clinical workflow.

3. Medical Background

Three parts of the human ear are responsible for converting sound waves into electrical impulses: the outer, the middle, and the inner ear. Each of these has tiny, complex structures that detect vibrations, transmit mechanical energy, or convert mechanical energy into electrical nerve impulses (see Fig. 1 (a)). If one part or structure is malformed, damaged, lost or not fully functioning, the patient's ability to hear is either impaired or totally lost.

Deafness is the partial or total inability to hear caused by sensorineural or by conductive hearing loss. *Sensorineural loss* occurs when there is a damage to the inner ear (cochlear) or to the nerve pathways from the inner ear to the brain. *Conductive hearing loss* is caused by problems with the ear canal, eardrum, middle ear, or abnormalities in mobile portions of the ear. These movable parts transmit sound from the outside to the inner ear. Conductive hearing loss occurs when these movable parts are damaged, lost (see Fig. 1 (b)) or when their mobility is impaired.

3.1. Treatment of Deafness

Nowadays, hearing can be restored using cochlear implants for sensorineural and a tympanoplastic surgery for conductive hearing loss. A cochlear implant consists of an internal and external component. The internal component is surgically inserted under the skin behind the ear, and a narrow wire is threaded into the inner ear. The external component is connected to the internal one and sound waves are converted to electrical impulses bypassing the defective inner ear and providing patients with the ability to hear [MM13]. A tympanoplastic surgery re-establishes the ossicular chain and the non-functioning ossicles may be reshaped to fit properly or replaced with an prosthetic (artificial) implant. Gaps between intact stapes and either the incus, malleus handle or eardrum are bridged with a *partial ossicular prosthesis* (PORP) [GB09]. If there is no stapes superstructure and the prosthesis connects the stapes footplate to the other ossicles or eardrum, it is called a *total ossicular prosthesis* (TORP) (see Fig. 1 (c)). However, successful reconstruction and implanting presents significant challenges. Besides a working knowledge of available materials, prosthesis design, reconstruction techniques and their adaptability to a variety of problems, profound surgical skills are essential [CPI97]. Reconstruction results are variable, with some procedures giving complete normal hearing and some giving no improvement in hearing at all [GB09]. The results achieved from the implantation of a PORP or TORP are dependent on [SE84]:

1. the pathology present in the area to be implanted and the surgical skills employed in the implantation,
2. the implants' biomechanical properties, and
3. the biocompatibility of the material implanted.

Otologic surgeons-to-be need an extensive and long training and trial-and-error experience in hearing restoration, including the correct prosthesis length judgment and trimming during surgery as well as the correct prosthesis positioning.

3.2. Workflow Analysis

Tympanoplastic surgery is a minimally invasive intervention using a micro-surgical technique to enlarge the view of the ear structures. The surgeon tries to position a prosthesis implant through the ear canal using surgical instruments while looking through a microscope. In clinic routine, stereoscopic microscopes are used to support the spatial orientation and improve the navigation. In detail, the eardrum is elevated and lifted forward and the prosthetic implant is placed into the middle ear underneath the remaining eardrum to bridge the gap between the eardrum and the damaged or missing bone structures, sketched in Figure 2 (b). The major challenge is the length estimation of the prosthesis. It may require placing the prosthesis in the ear several times to es-

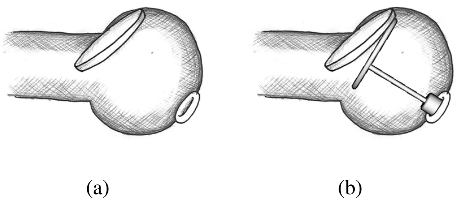


Figure 2: *(a) The conceptual setup of conductive hearing loss with no stapes bone, but the footplate. (b) TORP implants are used to bridge the gap between the eardrum and the footplate and to restore the hearing ability.*

timate the correct length. The correct length is such that the prosthesis just touches the undersurface of the eardrum without tenting it and bridging the existing gap, as seen in Figure 2 (a). The surgeon has to perceptually combine the images seen through the microscope with his actions. Additionally, a patient-specific depth judgment of the structure relationships is required, since vital anatomy landmarks may be obscured by disease or exposed to serious injury.

4. Experimental Setup

Initially, we analyzed the tympanoplastic surgery workflow by observations and interviews (see Section 3.2). We defined the essential tasks that require high-perceptual skills of an experienced surgeon and thus are essential for a training scenario: a TORP implant surgery with no incus, malleus, or stapes bones, but footplate existent (see Fig. 2 (a)). This surgery is chosen based on the simplicity and minimal number of existing structures. By focusing on one prosthesis type, we minimize bias factors, e.g., different prosthesis types, different structures and individual anatomy. Bias factors are every variation from one to another stimulus that may influence the results. Thus, participants were asked to select the correct TORP size and position it correctly. A correct position is achieved when the TORP touches the eardrum without penetrating it and bridging the gap between eardrum and stapes footplate, explained in Section 5.5. We do not expect that one display is better or worse than the other in every single investigated aspect, but we expect a difference. Because of that, we defined one- and two-tailed hypotheses.

One-tailed hypotheses. We postulate one-tailed hypotheses for object placement and orientation, since this correlates directly with depth perception.

We hypothesize that a 3D autostereoscopic display enables participants to place TORPs more accurately than a 2D display:

$\mathbf{H}_{accTransl}$: *as measured by smaller translation deviations.*

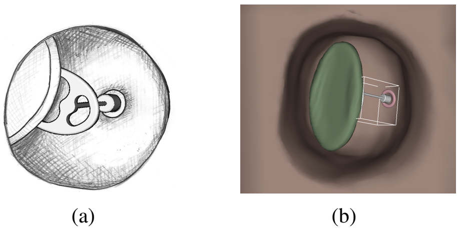


Figure 3: *(a) The surgeon's view through the microscope. The ear canal with the eardrum slightly folded to the side. (b) We imitate the surgeons' field of view and the participants explore the stimuli just as the surgeon. The same initial point of view, restricted field of view, and scope of action.*

$\mathbf{H}_{accRot}$: *as measured by smaller angular deviations.*

Two-tailed hypotheses. We claim that the interaction and task completion results are different.

We hypothesize that there is a difference between the 3D autostereoscopic display and the 2D display:

$\mathbf{H}_{taskTime}$: *in the mean task completion time.*

$\mathbf{H}_{actionScene}$: *in the number of scene interactions.*

$\mathbf{H}_{actionTORP}$: *in the number of TORP interactions.*

5. Controlled Perceptual Study

This work comprises a controlled perceptual user study comparing the depth perception support of a 2D and an autostereoscopic 3D display for a tympanoplastic training.

5.1. Participants

Our participants were recruited from various parts of the university including medical experts. Although it is recommended to recruit prospective users, in our case ENT (ear, nose and throat) surgeons, participants from the general population can also provide useful insights. Moreover, it is likely that the measured results can be applied to prospective users concerning the perceptual effectiveness, even though ENT surgeons may achieve better accuracy, because of their clinical experience. 42 participants, 21 using the 2D and 21 using the 3D autostereoscopic display participated. The 2D group comprised 13 female and eight male participants aged between 18 and 35 with a mean age of 24.5 years and the 3D autostereoscopic group comprised nine female and 12 male participants aged between 20 and 46 with a mean age of 27.4 years. In summary, we took care of two similar groups with respect to the age and gender ratio.

5.2. Stimuli

Based on the workflow presented in Section 3.2, we designed the stimuli setup similar to the surgeon's view

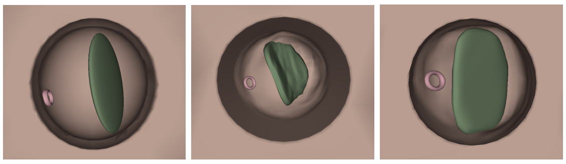

(a) simple (b) moderate (c) difficult

Figure 4: *(a) Simple, (b) moderately difficult, and (c) difficult stimuli were used during the study. Increasing occlusion caused by the eardrum and non-parallel orientations of eardrum and footplate lead to more TORP positioning effort.*

through the microscope, as sketched in Figure 3 (a). The same initial point of view, restricted field of view and scope of action were used. Each stimulus represents a middle ear scene viewed through the ear canal with the eardrum slightly folded to the side and a footplate at the back (see Fig. 3 (b)). No sophisticated illustration or illumination technique was applied so that we can focus on the effect of the display technology. Only color, ambient, and diffuse illumination was used to distinguish and identify the relevant structures.

We had four patient-specific petrous bone CT datasets provided from our collaboration partner at the university medical center. Additionally, we generated ten further datasets. All patients had a conductive hearing loss. Two of the four datasets contained 230× 230 × 208 voxels and a voxel length of 0.2 mm and two contained 230× 230 × 123 voxels and a voxel length of 0.353 mm. All structures were segmented using thresholding methods and were based on the segmentation results of the surface morphology. Five other scenes were generated manually based on the characteristics of real datasets, surgery inspections, video records and in cooperation with our medical experts. To unify and ease the stimuli scenes, a wall behind the footplate was integrated and thus closed the middle ear in the back. On the one hand, this wall restricted the field of view and unified the individual stimuli. On the other hand, the middle ear scene and study task was easier to understand. The virtual 3D prosthesis implant models were provided by KURZ GmbH Medizintechnik (www.kurzmed.de). They manufacture the TORP implants used by our medical experts for tympanoplastic surgery. We scaled those models along their longitudinal axis to get different TORP lengths.

5.3. Apparatus

All participants were tested alone by daylight and the stimuli were viewed from a distance of approximately 0.7 m (each stimulus subtended 17.8° of a visual angle on average). Nine different middle ear stimuli were presented to the 2D group and to the 3D autostereoscopic group using a 24″ monitor at a resolution of 1920 × 1200 pixel. The autostereoscopic display is a custom-built 3D display by Fraunhofer HHI. The

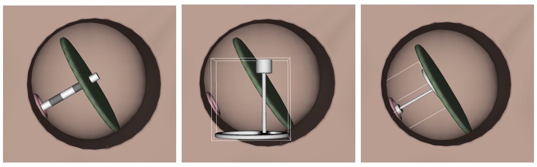

(a) start (b) TORP (c) end position

Figure 5: *(a) Initially, a depth judgment is required to chose the correct TORP length. (b) The TORP is displayed and has to be positioned between (c) the eardrum and the footplate.*

display is a 3D zeroCreative www.zerocreative.com display. A head-tracking unit and the corresponding tracking technology and software was developed and integrated by Fraunhofer HHI [BLR11].

5.4. Pilot Study

A prior within-participants pilot study with ten participants and 14 stimuli was conducted. This study revealed the visibility of disturbing crosstalk (ghosting) and binocular parallax of the images using the autostereoscopic display. To minimize these artifacts, we reduced the color saturation of the eardrum and the footplate, the contrast and eliminated sharp shadows by adapting the ambient and diffuse illumination term. Aside from that, the participants complained about the study duration, the number of stimuli and most frequently commented on the varying stimuli difficulty. Due to this, we chose nine stimuli (the four real datasets and five generated scenes) categorized into three simple, three moderately difficult and three difficult stimuli (see Fig. 4). By selecting stimuli of each category, we assure different degrees of difficulty similar to patient-specific anatomy. These categories are defined by the eardrum's orientation combined with the caused occlusion. Increasing occlusion of the middle ear cave and footplate leads to increasing positioning effort and demands better depth perception (see Fig. 4 (c)). Additionally, if the disk-like eardrum and footplate are oriented parallel (facing each other), it is easier to position the TORP. Figure 4 (a) illustrates a stimuli scene defined as a simple stimulus. The ear drum and footplate are almost parallel. Increasing deviation of this orientation aggravates the positioning.

5.5. Procedure

The within-participants pilot study has shown that a between-participants design is definitely the best for our purpose. This design prevents stimuli recognition and thus prior knowledge of depth and 3D orientation of the individual stimuli. If participants perform the study with the 2D display first, they get a mental model of the 3D stimuli scene, by rotating and interacting with the individual scenes and vice versa by the depth visualization of the 3D displays. Thus, we had the 2D and the 3D group.

In advance, all observers were instructed in written form. Two practice trials followed to familiarize each participant with the task and interaction technique. Additionally, we used two stimuli to minimize outlier, variance, and learning effects. These stimuli were not used during the experiments. For each middle ear scene, participants were asked to estimate the appropriate prosthesis length and then to position the TORP. During surgery, the cutting nib with known extension is used to estimate the distance between eardrum and footplate. We provide a yardstick that can be displayed, as shown in Figure 5 (a). If desired, the yardstick is illustrated as a small cylinder with differently colored areas of known extension (0.5 mm). Since we do not provide an implant trimming simulation, the participants had to choose between different TORP lengths. Besides the correct length, three other TORPs with different lengths (step size of 0.25 mm) were presented. The participants saw the possible TORP length as displayed values. After selecting the desired length, the chosen TORP was placed axis-aligned in front of the middle ear (Fig. 5 (b)). Participants had to navigate the TORP through the ear canal to the correct position, shown in Figure 3 (b). A correct position is achieved when the TORP touches the eardrum without penetrating it and bridging the gap between eardrum and stapes footplate. The TORP's orientation is defined as optimal when the TORP's "wheel" part fully touches the eardrum and the longitudinal axis is 90 degrees to the eardrum (compare Fig. 3 (b) and 5 (c)). This concrete orientation definition supports the task understanding and facilitates the object placement. Interaction with the TORP was achieved using a bounding box widget that is displayed around the TORP (Fig. 3 (b) and 5 (c)).

Each middle ear scene was presented until the participants pressed a "Ready" button to indicate that they were satisfied with their position and ready to move on to the next. The stimuli were shown randomly to each participant and at the end of the study the participants were asked for comments and personal opinions.

6. Results and Analysis

This is a between-participant study with nine stimuli divided into three per degree of difficulty. In detail, two independent samples of 21 participants each, the 2D and the 3D autostereoscopic display as the two between factors and the three degrees of difficulty as within factors lead to a 2 × 3 factorial analysis of variance (ANOVA) with $\alpha = 0.05$. A pair-wise analysis is performed using the T-test.

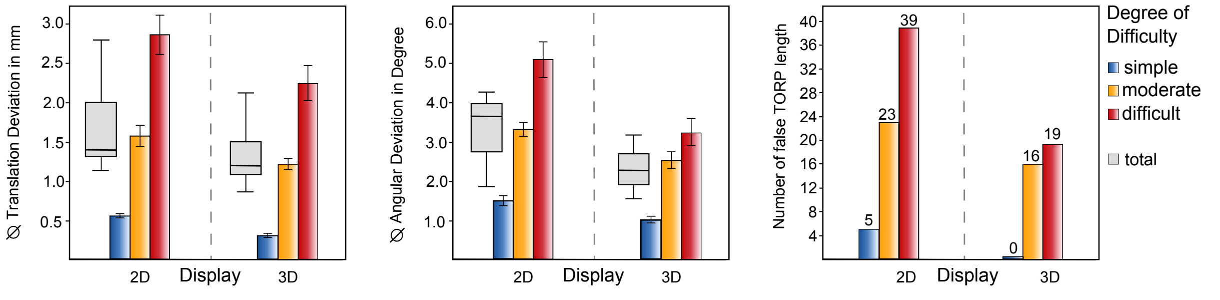
We quantitatively analyzed accuracy in terms of the object placement (translation deviation) and orientation (angular deviation), the interaction effort regarding scene and TORP, and the task completion time. Additionally, we evaluated the chosen implant length and can therefore qualitatively analyze the depth judgment and distance assessment of eardrum and footplate. The individual results will be outlined in the following sections.

6.1. Depth Perception

As explained in Section 5, participants had to perform a depth judgment and based on that chose the right TORP. To validate the length estimation and accuracy, a *gold standard* TORP and position is required for each individual middle ear stimulus. These *gold standard* TORPs were generated in advance by our medical experts. The position, the orientation, and the chosen TORP length were determined and analyzed for each stimulus.

Object placement. The accuracy of the object placement is defined as the average translation deviation of each TORP compared to the *gold standard* TORP. For each participant and stimulus, the 4×4 transformation matrix T_{sample} is recorded. This matrix includes the translation and rotation performed by the participants to place the TORP. To calculate the translation accuracy, we defined three control points along the TORPs longitudinal axis (both ends and a mid point). These points are multiplied with T_{sample} and thus transformed according to the participant's TORP transformation. The points' distance to the *gold standard* TORP is then calculated. We refer to this as $\Delta trans$. A little variation of $\Delta trans = \pm\, 0.2\ mm$ is negligible, since this corresponds to the footplate's diameter. As long as the TORP's foot (thin end) is positioned within the footplate, a correct hearing restoration is possible. The 3D group achieved an average of $\mu_{\Delta trans} = 1.29\ mm$ and thus a more accurate position estimation result than the 2D group with $\mu_{\Delta trans} = 1.69\ mm$. With $F(1,40) = 9.56,\ p \leq 0.05$ and $\eta^2 = 0.246$ an effect exists and $\mathbf{H}_{accTrans}$ is likely to be true. Significant differences exist for the simple ($p < 0.01$ and $t(40) = 5.06$) and the moderately difficult stimuli ($p \leq 0.05$ and $t(40) = 2.17$). Difficult stimuli are marginal not significant with $p = 0.07$ and $t(40) = 1.82$ (2D group with $\mu_{\Delta trans} = 2.905\ mm$ and 3D group with $\mu_{\Delta trans} = 2.29\ mm$). As shown in Figure 6 (a), the degree of difficulty is chosen properly. The average results confirm the three gradations. Especially, stimuli with high difficulty show higher translation deviations. The autostereoscopic display improves the depth perception and facilitates the TORP placement especially for simple (2D group with $\mu_{\Delta trans} = 0.57\ mm$ and 3D group with $\mu_{\Delta trans} = 0.32\ mm$)and moderately difficult (2D group with $\mu_{\Delta trans} = 1.601\ mm$ and 3D group with $\mu_{\Delta trans} = 1.254\ mm$) stimuli.

TORP orientation. To define the orientation accuracy, the TORPs is treated as a 3D unit vector $\vec{v}$. This vector $\vec{v}$ is transformed by T_{sample} and the angular difference Δrot to the *gold standard* orientation is calculated. Rotations around the longitudinal axis are not considered, since the majority of our participants is just medically knowledgeable. Moreover, this rotation has less impact when analyzing depth perception, it just defines the position of the TORPs' "wheel" part relatively to the ear canal. The 3D group achieved less angular deviations ($\mu_{\Delta rot} = 2.27°$) than the 2D group ($\mu_{\Delta rot} = 3.32°$), as shown in Figure 6 (b). Both results are



(a) (b) (c)

Figure 6: *The average results of the (a) position estimation, (b) the orientation, and (c) the number of chosen TORPs with a false length. The differences reveal the advantages of the 3D autostereoscopic display compared to the 2D display for the TORP implanting task.*

close to the *gold standard*. Angular deviations of $\Delta rot > 2^\circ$ are recognizable. The ANOVA confirmed that $\mathbf{H}_{accRot}$ with $p \leq 0.01$ and $\eta^2 = 0.390$ is likely to be true and $\mathbf{H}_0$ according to orientation accuracy is highly unlikely. Compared to the translation results, the average angular deviations of the 2D versus 3D group reveal a higher difference. Significant differences exist for the simple ($p < 0.01$ and $t(40) = 2.85$), the moderately difficult ($p = 0.01$ and $t(40) = 2.69$), and the difficult stimuli ($p \leq 0.01$ and $t(40) = 3.21$).

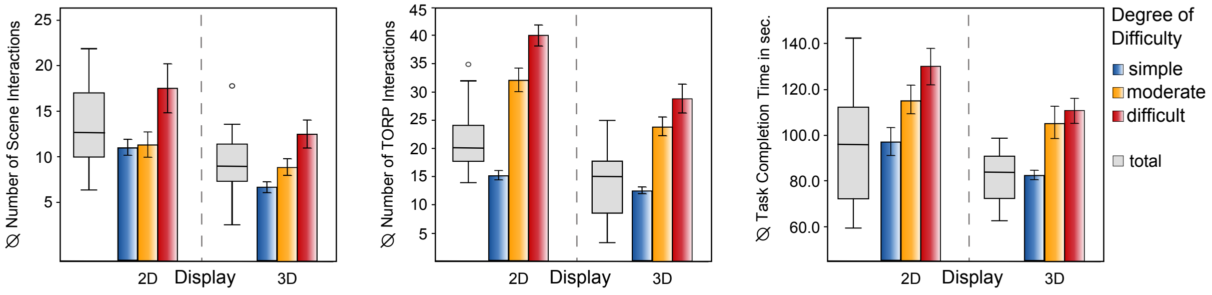
TORP length. The effectiveness of depth judgment of the structure relationship is analyzed by the chosen TORP length. Figure 6 (c) presents the number of all falsely chosen TORPs. We count the number of all TORPs that had a wrong length. 189 choices had to be made in each group. All TORPs chosen for simple stimuli by participants of the 3D group were correct. A TORP with a deviating length of 0.25 *mm* was chosen five times during the study with the 2D group. The moderately difficult stimuli lead to 23 out of 67 false choices in the 2D and 16 out of 35 in the 3D group. All falsely chosen TORPs by the 3D group deviate from the correct length by 0.25 *mm*. In the 2D group 52 % of all choices for a moderately difficult stimuli deviate by 0.25 *mm*, 21 % deviate by 0.5 *mm*, and 26 % deviate by 0.75 *mm*. While the wrong choices for the difficult stimuli of the 3D group (one of 19 false choices deviates by 0.5 *mm* and one by 0.75 *mm*) are similar to the moderately difficult stimuli, there are more falsely chosen TORPs in the 2D group (39), as shown in Figure 6 (c). This gives only a little insight into the depth perception of structure distances, but has to be evaluated in detail. Participants of the 3D group chose more correct sized TORPs' and positioned these TORPs more accurate by less translation and angular deviations. Even for the simple and moderately difficult stimuli are significant differences existent between the two displays.

6.2. Interaction

The interaction effort with the stimuli scene and the TORP was measured by recording the number of performed interactions. This was used to verify task completion times and to analyze whether interaction with the scene and thus motion was used to perceive depth cues. We did not measure the path length. Since this is considered to be a training scenario, the required time is interesting and therefore measured but not crucial. Participants were able to zoom in the scene until the camera reached the entry of the middle ear and a restricted scene rotation was possible, too. The restriction is defined by the middle ear bounding box. As long as the middle ear entry is almost fully oriented to the viewer, rotation is possible. For the TORP, rotation and translation are the two provided interaction techniques. As mentioned above, the TORPs bounding box served as the interaction widget and thus was displayed as thin box wire.

Scene interaction. We confirm $\mathrm{H}_{actionScene}$ with a significant difference between the 2D and the 3D group ($F(1,40) = 10.19, p \leq 0.05$ and $\eta^2 = 0.203$). Participants of the 2D group performed on average $\mu_{scene} = 13.26$ scene interactions, while participants of the 3D group required on average $\mu_{scene} = 9.34$ scene interactions until the TORP was placed, as seen in Figure 7 (a). A significant difference between 2D and 3D display is confirmed with $p \leq 0.01$ and $t(40) = 4.08$ for simple stimuli. Especially the difficult stimuli required more scene (2D: $\mu_{scene} = 17.48$ and 3D: $\mu_{scene} = 12.49$) and TORP interactions (2D: $\mu_{TORP} = 40.0$ and 3D: $\mu_{TORP} = 28.84$). While the simple and moderately difficult stimuli exhibit almost the same number of scene interaction using a 2D display (simple: $\mu_{scene} = 11.02$ and moderate: $\mu_{scene} = 11.29$), the average number of TORP interactions shows a clear distinction between simple, moderate, and difficult.

TORP interaction. Comparing the 2D with the 3D display, $\mathrm{H}_{actionTORP}$ ($F(1,40) = 17.55,\ p \leq 0.01$ and $\eta^2 = 0.305$



(a) (b) (c)

Figure 7: *The average results of the (a) task completion time, (b) the number of scene interactions, and (c) the number of TORP interactions. Participants of the 3D group required less scene and TORP interactions and thus less task completion time to position the TORP.*

) is confirmed with an average number of TORP interactions $\mu_{TORP} = 29.12$ of the 2D group and $\mu_{TORP} = 21.75$ of the 3D group. A significant difference exists for the simple ($p \leq 0.05$ and $t(40) = 2.47$), the moderately difficult ($p \leq 0.05$ and $t(40) = 2.98$), and the difficult stimuli ($p = 0.01$ and $t(40) = 3.45$). Comparing the results in Figure 7 (b), participants that performed the study using the 3D autostereoscopic display required less TORP interactions to position the implant. Nevertheless, as shown in Section 6.1, they are more accurate than the participants of group 2D. A detailed analysis revealed that especially rotating the TORP required more interaction than the translation. Additionally, a pair-wise comparison using the T-Test revealed a significant difference even for each degree of difficulty. Participants of the 3D group required significantly less scene interactions as well as TORP interactions compared to the 2D group when positioning the TORP. Hence, the support of the 3D autostereoscopic display in terms of less number of required interactions is statistically confirmed.

6.3. Task Completion Time

The task completion time result with $F(1,40) = 4.77$, $p \leq 0.05$ and $\eta^2 = 0.107$ shows a marginal effect. We confirm that $\mathbf{H}_{taskTime}$ is likely to be true. As visualized in Figure 7 (c), the 3D group required an average time of $\mu_{time} = 99.72$ *sec*, while the 2D group required on average $\mu_{time} = 114.12$ *sec* to position the TORP. Moreover, a significant difference of the required time for the simple stimuli exist between the 2D and the 3D group ($p \leq 0.05$ and $t(40) = 2.19$). The difference of the difficult stimuli with $\mu_{time} = 129.98$ *sec* for the 2D and $\mu_{time} = 110.82$ *sec* for the 3D group is with $p = 0.055$ and $t(40) = 1.97$ not significant as well as the difference between the moderately difficult stimuli $p = 0.335$ and $t(40) = 0.98$.

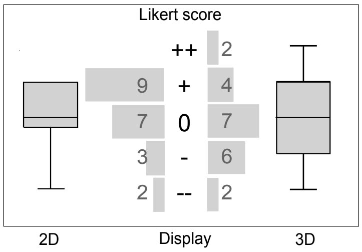


Figure 8: *The qualitative results of the two displays. While the 2D group tend to rate the display with* $+$ *or* 0*, the 3D group results are normally distributed around* 0.

6.4. Qualitative Results

Participants of the within-participant pilot study were asked to compare both displays and state their display preference. We used a five point Likert scale, with each pole representing one display compared to the other. Four of the ten participants rated the 2D display as *good* compared to the 3D display. Four did not tend to one display and rated *neutral* and three rated the 3D display with *good* since they liked the 3D depth visualization, even though there were some artifacts (see Section 5). No participant rated one display using *very good*.

Participants of the between-participants study had to rate one display. The personal preference regarding the display and the provided support for this task had to be rated. Therefore, we provided a five point bipolar Likert scale $(-,--,0,+,++)$. As illustrated in Figure 8, the results of the 3D group are normally distributed around 0. That means the participants tend to a neutral opinion for the TORP implanting task with a 3D display. 19% rated with $+$, 33.3% with 0 and 28.6% with $-$. While two participants strongly liked the stereo visualization and rated with $++$, no participant of the 2D group rated the 2D display with $++$. Overall, the participants of the 2D group rated this display either with 0 or with +. In summary, the 3D group tends to a neutral opinion and thus to no support of the 3D autostereoscopic display, while the 2D group reveals a tendency to + and therefore to a little support of the 2D display.

7. Discussion

All postulated hypotheses are statistically confirmed. We reject $\mathbf{H}_0$, since there are differences between the two displays and depth perception is more supported using a 3D autostereoscopic display compared to a 2D display. The quantitative analysis reveals the advantages of the stereoscopic visualizations. Participants of the 3D group chose more correct TORPs, positioned them more accurately in terms of smaller translation and angular deviations, and required less task completion time. Thus, the autostereoscopic display provides a more correct depth estimation compared to the 2D display. However, the TORP's position and thus translation deviation was biased by the occlusions of TORP and footplate. Participants used this information to find the correct position. Occlusion as a bias factor was identified by the participants' comments and observation and should be addressed and eliminated in further studies.

Comparing the number of interactions and time results, the 2D group tried to improve the depth perception by scene and TORP interactions. Furthermore, the results show that participants of the 3D group required almost as much time as the 2D group without the same interaction effort. Observations during the study and a few participants' comments showed that using the 3D autostereoscopic display is not as comfortable compared as the 2D display. Participants had to refocus during the study to receive a proper 3D stereo image. Moreover, there are still a few disturbing crosstalk effects left that hamper the 3D viewing and slow down the TORP placement. These disadvantages have an impact on the qualitative analysis results. Disturbing artifacts and a demanding 3D visualization caused by the display technology lead to a more negative rating of the 3D autostereoscopic display. The quantitative results of the participants' performances, however, indicate the advantages of the 3D display.

zSpace. Additionally, we performed a follow-up study with a zSpace system. This is a 3D virtually holographic imaging display with a passive circular 120 Hz stereo 3D polarization technology. This glass-based systems includes an optical tracking giving an angle-dependent stereoscopic view of 1920 × 1080 pixels full HD with time-interleaved stereo frames. Additionally, a six DOF stylus input device is provided for intuitive 3D manipulation and navigation. Seven participants (two female and five male, ages 27-37 with a mean age of 30.5 years) performed the task with a zSpace system. Every single participant was enthusiastic positioning the TORP using the stylus as input device. They all rated the display and the stylus with ++. The 3D stereo effect is very realistic and the view angle-dependent visualization technology enables the scene exploration by moving the head. It is possible to look on the side of the eardrum and thus to verify the TORP position without rotating the scene. In detail, the participants required on average $\mu_{scene} = 0.138$ scene interactions and $\mu_{TORP} = 4.57$ TORP interactions and completed the task on average in $\mu_{time} = 22.34$ seconds (simple: $\mu_{time} = 13.75$ *sec.*, moderate: $\mu_{time} = 21.78$ *sec.*, difficult: $\mu_{time} = 30.37$ *sec.*). The stylus enables a really fast object placement with a few stylus interactions, since six DOFs are provided. Moreover, the system combined with the stylus enables a more accurate object placement with $\mu_{\Delta trans} = 0.79$ *mm* and $\mu_{\Delta rot} = 1.78°$.

The results illustrate the correlation between increasing difficulty and perceptual effort. Thus, the importance of different stimuli with a varying degree of difficulty is revealed. Significant differences with $p \leq 0.01$ between the degree of difficulty for object placement, orientation, TORP interaction and task completion time were achieved. Differences with $p \leq 0.05$ were achieved for scene interaction. Thus, our categories are confirmed and simple, moderate, and difficult were chosen carefully.

8. Conclusion

Our between-participant study with 42 participants revealed a clear preference of the 3D autostereoscopic display compared to a 2D display for a tympanoplastic training scenario. Moreover, the follow-up study revealed the overwhelming depth perception improvement with the zSpace system using a stylus input device. The view angle-dependent stereoscopic view enables a scene exploration without a manually scene interaction. This leads to a faster task completion performance. The stylus enables a more comfortable and precise object placement. These results combined with the enormous personal preference reveals the high potential of the zSpace system. Participants were impressed by the realistic 3D visualization and depth perception.

We designed 3D stimuli to imitate a mirco-surgical TORP positioning task. This experimental study statistically confirmed the support of the 3D autostereoscopic display compared to a 2D display and outlined the potential of the new 3D zSpace technology. Depth judgment for TORP length estimation and positioning as well as the number of required interaction and the task completion time revealed the advantages of 3D stereo visualizations. Moreover, we showed that even if the personal preferences tend to a 2D display, the quantitative results reveal a better and more accurate performance with the autostereoscopic 3D display.

Besides the advantages regarding accuracy, the 3D display predominated the 2D display in terms of the TORP's length estimation, the required number of interactions and the task completion time. Nevertheless, the advantages are present as long as the visualizations including color, saturation and contrast are adapted, and visible artifacts are minimized. The 3D autostereoscopic display improves depth perception, but is not suitable for every visualization. Passive 3D display system with the 3D image being compounded by the used glasses reveal disturbing crosstalk effects. Especially highly-contrast and saturated colors enhance the perceived ghosting images. Thus, there is still room for improvement of 3D technology to overcome the differences to 2D and active 3D stereo systems and to support the depth perception for individual visualizations.

Acknowledgements The authors would like to thank Dr. U. Vorwerk for medical datasets and knowledge and M. Luz (Technical University Berlin) for her support and most helpful explanations of statistical analysis. The work was partly funded by the German Ministry of Education and Research (BMBF) within the *STIMULATE* project (grant no. 03FO16101A).

References

[BCFW08] Bartz D., Cunningham D. W., Fischer J., Wallraven C.: The role of perception for computer graphics. In *Eurographics State-of-the-Art Report 4* (2008). 2

[BGCP11] Baer A., Gasteiger R., Cunningham D. W., Preim B.: Perceptual evaluation of ghosted view techniques for the exploration of vascular structures and embedded flow. *Comp. Graph. Forum 30*, 3 (2011), 811–820. 2

[BGKG06] Bruckner S., Grimm S., Kanitsar A., Gröller E.: Illustrative context-preserving exploration of volume data. *IEEE Trans. Vis. Graph. 12*, 6 (2006), 1559–1569. 1

[BLR11] Braun M., Leiner U., Ruschin D.: Evaluating motion parallax and stereopsis as depth cues for autostereoscopic displays. In *Proc. SPIE 7863, Stereoscopic Displays and Applications XXII,* (2011), vol. 7863. 2, 5

[BvBKS96] Buess G., van Bergen P., Kuner W., Schurr M.: Comparative study of various 2d and 3d vision system in minimally invasive surgery. *Chirurg 67*, 10 (1996), 1041–1046. 2

[CPI97] Carrasco V. N., Pillsbury III H. C.: *Revision otologic surgery*. Thieme Medical Publishers, Inc., 1997. 3

[CWMC09] Chan M.-Y., Wu Y., Mak M.-H., Chen W.: Perception-based transparency optimization for direct volume rendering. *IEEE Trans. Vis. Graph. 15* (2009), 1283–1290. 2

[FRH10] Feng C., Rozenblit J., Hamilton A.: A computerized assessment to compare the impact of standard, stereoscopic, and high-definition laparoscopic monitor displays on surgical technique. *Surgical Endoscopy 24*, 11 (2010), 2743–8. 2

[GAPD09] Gurusamy K., Aggarwal R., Palanivelu L., Davidson B.: Systematic review of randomized controlled trials on the effectiveness of virtual reality training for laparoscopic surgery. *The British Journal of Surgery 95*, 9 (2009), 1088–97. 2

[GB09] Graham J., Baguley D.: *Ballantyne's deafness*, 7th edition illustrated ed. Wiley, 2009. 3

[GBG*14] Gerber N., Bell B., Gavaghan K., et al.: Surgical planning tool for robotically assisted hearing aid implantation. *Int. J. Computer Assisted Radiology and Surgery 9*, 1 (2014), 11–20. 2

[GNBP11] Gasteiger R., Neugebauer M., Beuing O., Preim B.: The flowlens: A focus-and-context visualization approach for exploration of blood flow in cerebral aneurysms. *IEEE Trans. Vis. Graph. 17*, 12 (2011), 2183–2192. 1

[GSBH13] Grosset P., Schott M., Bonneau G.-P., Hansen C.: Evaluation of depth of field for depth perception in DVR. In *IEEE Pacific Visualization* (2013), pp. 81–88. 2

[JDG*04] Jourdan I. C., Dutson E., Garcia A., et al.: Stereoscopic vision provides a significant advantage for precision robotic laparoscopy. *British Journal of Surgery 91*, 7 (2004), 879–885. 2

[JM07] John N. W., McCloy R.: Navigating and visualizing three-dimensional data sets. *The British Journal of Radiology 77*, 2 (2007), 108 – 113. 1

[KHSI04] Kim S., Hagh-Shenas H., Interrante V.: Conveying shape with texture: Experimental investigations of texture's effects on shape categorization judgments. *IEEE Trans. Vis. Graph. 10*, 4 (2004), 471–483. 2

[KSK13] Kunert W., Storz P., Kirschniak A.: For 3d laparoscopy: a step toward advanced surgical navigation: how to get maximum benefit from 3d vision. *Surg. Endosc. 27*, 2 (2013), 696–9. 2

[MM13] Mudry A., Mills M.: The early history of the cochlear implant: a retrospective. *JAMA Otolaryngology– Head & Neck Surgery 139*, 5 (2013), 446–453. 3

[MRBH*09] Manzey D., Röttger S., Bahner-Heyne J. E., et al.: Image-guided navigation: the surgeon's perspective on performance consequences and human factors issues. *The international journal of medical robotics + computer assisted surgery : MRCAS 5* (2009), 297–308. 2

[RB14] Read J. C., Bohr I.: User experience while viewing stereoscopic 3d television. *Ergonomics 57*, 8 (2014), 1140 – 1153. 1

[RHD*06] Ritter F., Hansen C., Dicken V., et al.: Real-time illustration of vascular structures. *IEEE TVCGIEEE Trans. Vis. Graph. 12*, 5 (2006), 877–884. 2

[SE84] Shea J. R., Emmett J. R.: Polyethylene TORPs and PORPs in otologic surgery. In *Proc. of the First International Symposium Biomaterials in Otology* (1984), pp. 137–152. 3

[SEWL*11] Stelter K., Ertl-Wagner B., Luz M., et al.: Evaluation of an image-guided navigation system in the training of functional endoscopic sinus surgeons. a prospective, randomised clinical study. *Rhin 49*, 4 (2011), 429–37. 2

[UH07] Ukai K., Howarth P. A.: Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations. *Displays 29*, 2 (2007), 106 – 116. 1

[WFG92] Wagner L. R., Ferwerda J., Greenberg D.: Perceiving spatial relationships in computer-generated images. *Iee Comp. Graphics & Applic.* (1992), 44–58. 1

[WHK*12] Wagner O., Hagen M., Kurmann A., et al.: Three-dimensional vision enhances task performance independently of the surgical method. *Surg. Endosc. 26*, 10 (2012), 2961–8. 2

[WRB*09] Wiet G., Rastatter J. C., Bapna S., et al.: Training otologic surgical skills through simulation-moving toward validation: A pilot study and lessons learned. *Journal of graduate medical education 1*, 1 (2009), 61–66. 2

[WRK*14] Wilhelm D., Reiser S., Kohn N., et al.: Comparative evaluation of hd 2d/3d laparoscopic monitors and benchmarking to a theoretically ideal 3d pseudodisplay: even well-experienced laparoscopists perform better with 3d. *Surgical Endoscopy 28*, 8 (2014), 2387–2397. 2

[ZIK98] Zhukov S., Iones A., Kroning A.: An ambient light illumination model. In *Proc. of Eurographics Rendering Workshop* (1998), pp. 45–56. 1