Measuring and Visualizing Attention in Space with 3D Attention Volumes

Thies Pfeiffer
A.I. Group, Faculty of Technology, Bielefeld University
tpfeiffe@techfak.uni-bielefeld.de*

Abstract

Knowledge about the point of regard is a major key for the analysis of visual attention in areas such as psycholinguistics, psychology, neurobiology, computer science and human factors. Eye tracking is thus an established methodology in these areas, e.g., for investigating search processes, human communication behavior, product design or human-computer interaction. As eye tracking is a process which depends heavily on technology, the progress of gaze use in these scientific areas is tied closely to the advancements of eye-tracking technology. It is thus not surprising that in the last decades, research was primarily based on 2D stimuli and rather static scenarios, regarding both content and observer.

Only with the advancements in mobile and robust eye-tracking systems, the observer is freed to physically interact in a 3D target scenario. Measuring and analyzing the point of regards in 3D space, however, requires additional techniques for data acquisition and scientific visualization. We describe the process for measuring the 3D point of regard and provide our own implementation of this process, which extends recent approaches of combining eye tracking with motion capturing, including holistic estimations of the 3D point of regard. In addition, we present a refined version of 3D attention volumes for representing and visualizing attention in 3D space.

CR Categories: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented and virtual realities; H.5.2 [Information Interfaces and Presentation]: User Interfaces—Ergonomics; I.3.8 [Computer Graphics]: Applications—;

Keywords: 3d, gaze tracking, visual attention, visualization, motion tracking

1 Introduction

Humans live in and interact with a three-dimensional (3D) world. The eyes are the primary organ to perceive this world and the human brains have evolved to cope with all the particularities of our 3D surroundings. Knowledge about when and where humans target their visual attention is thus considered an important cornerstone for the investigation of the human mind. Hence, techniques to assess the direction of gaze have been developed and used in different fields. In linguistic research and cognitive sciences they are used, e.g., for investigations on language development and use [Tanenhaus et al. 1995]. In economics they provide insights for research on decision processes [Meiner and Decker 2010]. In sports they reveal differences in visual orientation strategies between experts and novices when it comes to quickly arbitrate game situations. In most studies, however, researchers have used a restricted two-dimensional (2D) world to test their hypotheses.

It may certainly be valid to test many hypotheses with either 2D or 3D stimuli, if the relevant effects can be expected to scale. This assumption, however, might not hold in general. Especially in areas of research addressing orientation, navigation, motor planning or spatial language, the additional dimension often requires collaterally increased efforts. A reduction of such scenarios to 2D is thus likely to render artifacts in the data which might be difficult to factor out without assessing the hypotheses on 3D stimuli, too.

Figure 1: *A head-mounted eye-tracking system by Arrington Research. It is augmented with an optical target for the DTrack2 tracking system by ART and with polarized filters for stereo presentations in a virtual environment. The modifications allow 3D gaze tracking in real and virtual environments.*

Mobile eye-tracking technology is one key technology to provide means to assess visual attention in arbitrary 3D environments, either in the laboratory, or in the fields (see Fig. 1). Especially the video-based approaches using scene-cameras to record a section of the field of view of the observer in parallel to monitoring his eye movements have gained a lot of attention in recent years. The analysis of the recorded data, however, is costly, as these devices basically provide video material with overlaid gaze-cursors and any classification of the fixations requires immense manual effort.

Recent developments, such as the mobile glasses developed by Tobii [Tobii Technology AB 2010] and SMI [SMI 2011], provide a sound technical basis for an analysis of visual attention during mobile interactions in a 3D environment. However, they still operate on a 2D or 2.5D [Marr 1982] abstraction and additional effort is required to extract real 3D data from those systems.

In the following, we will present our work on defining the general process and technical set-up to measure 3D point of regards of moving observers and provide a definition of 3D Attention Volumes that can be visualized using volume-based rendering to support the qualitative analysis of the distribution of visual attention of a group of observers in a complex 3D environment. This approach is thereby applicable both to computer-generated stimuli and real-world scenarios.

*Not for redistribution. The definitive version was published in Proceedings of the Symposium on Eye Tracking Research and Applications 2012, http://dx.doi.org/10.1145/2168556.2168560

2 Related Work

This work relies on the concept of the point of regard [Dodge 1907] and the relevancy of the sequence of fixations for investigating internal processes [Yarbus 1967] (later coined *scanpaths* [Norton and Stark 1971]).

2.1 Geometry-based estimation of the 3D point of regard

Bolt envisioned a gaze-based interaction system called *Gaze-Orchestrated Dynamic Windows* for the control of multiple windows as early as 1981 [Bolt 1981]. In his vision, visual attention was used for **online** control of the presentation of multimodal stimuli. In a very controlled setting, the eye gaze of an observer was to be measured and, while it is unclear to what extend this vision was realized, Bolt mentions R.A. Foulds work on a combination of eye tracking and head tracking to provide a point of regard in the environment. In Bolt's scenario, the target stimuli were 2D planes (the "World of Windows") and thus no full 3D point of regards were to be measured. However, the observer was able to move freely.

One of the first approaches to measure visual attention in space for experiments has been realized by Roetting, Goebel and Springer [Rötting et al. 1999]. They combined an eye-tracking system with an attached scene-camera and a 6DOF head-tracking in addition to the eye-tracking cameras. The point of regard was determined **offline** in a two-staged process. First the object contours were identified from at least two different perspectives on the image frames provided by the scene-camera. This was done to create a geometry model to approximate the object in space. In a second step, the fixations, which were mapped as 2D points on the image of the scene-camera by the eye-tracking system, were classified and for each frame the observed object (model of interest) was determined. With this approach the authors were able to take account for perspective changes on one axis semi-automatically. They report, that this method had been used for the evaluation of a new 3D radar display.

The described approach relies on the geometry of the target objects to classify the fixations and provides object-centered information about the distribution of attention. The classification of the fixations happens after a projection of the geometries on the image plane. An explicit computation of the line of sight is not made. For the real objects used in their scenario, these geometries have been created manually from scratch. If content is already being described by geometries, as it is the case for augmented or virtual reality, this process is simplified. First approaches - also object-centered - to measure visual attention in virtual reality were using Head-Mounted Displays (HMDs) [Tanriverdi and Jacob 2000; Duchowski et al. 2001; Duchowski et al. 2004]. Instead of the detour using a recorded image from the scene-camera, the systems could directly use the projection rendered for the particular eye from the framebuffer. Thus, the normal picking operations could be used to identify the object below a point of regard. Alternatively, the known position of the eye and the detected 2D fixation on the projection display in the HMD could be used to cast a ray into the 3D world to determine the intersection with the 3D object geometry [Duchowski et al. 2001] and thus identify the 3D point of regard and the model of interest.

Later, the work on estimating the 3D point of regard in 3D virtual worlds has been transferred from HMD-based projections to free interactions in front of large projection screens, such as CAVEs (e.g. [Pfeiffer 2008]). In these scenarios, the projection screen has no longer a fixed static position relative to the eye of the observer, as in the HMD-based settings. Such settings are thus closer to real-world scenarios, where a moving observer is attending to a 3D scenery. However, they still rely on an exact knowledge of the geometries of the relevant objects. This information is easily at hand in virtual reality scenarios, as it is the basis for the generation of the visual display of such systems, but it is not in real-world scenarios.

2.2 Holistic estimation of the 3D point of regard

Holistic methods to estimate the 3D point of regard do without a geometric model of the target objects. Instead, they integrate multiple information sources. One approach is the triangulation of the 3D point of regard based on at least two measured lines of gaze. These can be either provided by a binocular eye-tracking system [Essig et al. 2006; Pfeiffer et al. 2009], or by integrating over time [Kwon et al. 2006]. The holistic methods estimate the 3D point of regard based on information of the observer only. This way the described disadvantages of geometry-based methods are avoided.

In a desktop-based set-up using a static anaglyphic stereo projection of dots, Essig et al. [2006] showed, that the triangulation of the two visual axes was outperformed by their own approach based on machine-learning. They used a parameterized self-organizing map to learn the mapping from the 2D coordinates provided by the eye-tracking system to the 3D point of regard. Pfeiffer et al. [2009] later extended this approach to shutter-based projections and 3D object presentations, also in a static desktop-based set-up.

Real-life scenarios, such as walking through a supermarket, are currently in the domain of mobile eye-tracking systems using a scene-camera. These systems provide qualitative data which can be easily accessed, in terms of gaze cursors overlaid on the recorded scene video. However, they require costly manual annotations to collect quantitative data. Recent advances in mobile localization and holistic estimation of the 3D point of regard [Pirri et al. 2011], however, allow for a free interaction in such complex scenarios while still allowing for a statistical analysis with a reduced effort for manual annotation.

2.3 Visualizing the point of regard for 2D stimuli

For 2D stimuli, there exists a set of established visualization techniques to depict the recorded information on visual attention.

In *scanpath* visualizations, the location of a measured point of regard is depicted as circle. The diameter of this circle is often determined by the duration of the fixation on this particular point of regard. It could, however, also be used to represent the area of high acuity. The sequence of the point of regards is visualized by representing the saccades from one point of regard to the next by straight lines. This way, scanpaths provide one view of a particular visual exploration integrating over time. Scanpaths are, however, not suited to assess the aggregated visual attention over several participants. The resulting visualization would be too confusing (compare Fig. 3 and Fig. 5).

For an interpersonal visualization of point of regards, *attention maps* [Pomplun et al. 1996] were introduced, which are now often referred to as *heatmaps* (see Fig. 2). Attention maps do not use a discrete depiction of every point of regard. They are density surface maps that integrate the amount of overt visual attention targeted at every point of the stimuli, often per pixel of a digital 2D stimuli, over participants and over time. For the visualization, this attention map is then used to generate an overlay over the original stimulus material. While there is relative freedom in how the overlay is generated, it is typically a color-coded map, similar to the heat-images generated by infrared cameras and hence the name heatmap. The heatmap thus highlights the areas where many point of regards aggregate optically and shadows areas where less point

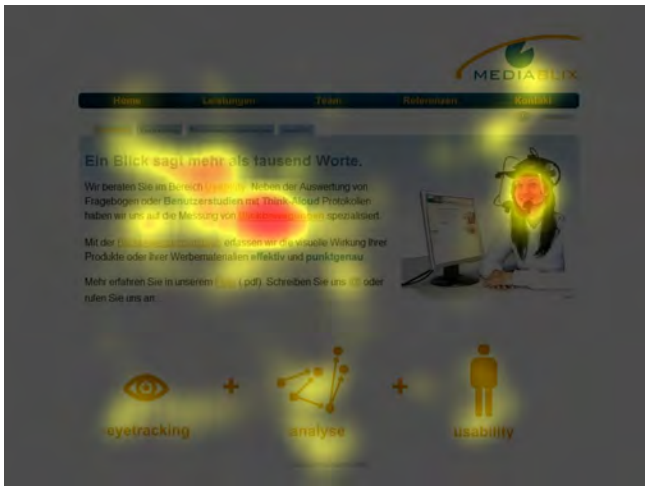


Figure 2: *The distribution of visual attention over 2D content visualized using a heatmap. The example shows the aggregated visual attention of ten participants on a webpage. Areas that receive a high level of attention are colored in red.*

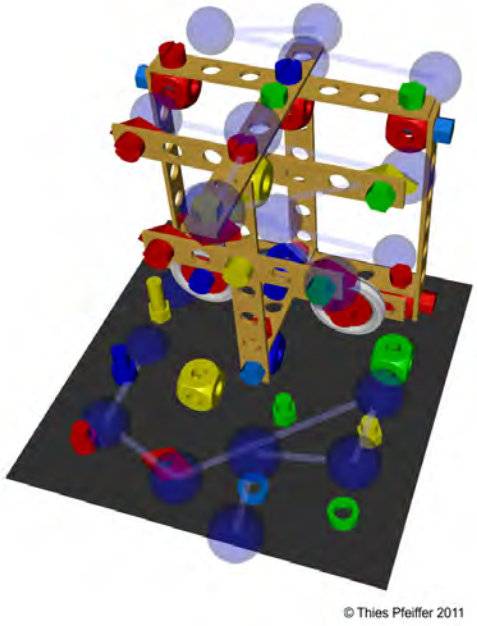


Figure 3: *3D scanpaths can be used to visualize the sequence of fixations made by a single observer over target objects. In this case the sequence of objects to fixate was given by instruction. The first fixation was on the lower left. Fixations are represented as spheres and connected to their successor by a cylinder.*

of regards were detected. This way the hot spots of visual attention pop out and a quick qualitative feedback is facilitated. It has to be stressed that there is no standard defining which parameters are computed and how they are exactly mapped to a visual representation when generating heatmaps, so care has to be taken when interpreting a presented heatmap [Bojko 2009].

So far, the mentioned work focused on the analysis of visual attention on 2D products or stimuli. However, what is successful for the analysis of 2D material could also be of help for 3D products, where we find multiple levels of depth, for example considering see-through head-up displays in modern cockpits, or multiple perspectives, for example considering the design of cars or the ergonomics of complex machineries.

2.4 Visualizing the point of regard for 3D stimuli

Once a 3D point of regard has been estimated, either by intersection with geometries or holistically, a visualization as 3D scanpath is straight forward (see Fig. 3). Formula and examples are given in Section 4.1.

More advanced methods today are, for example, object-centered visualizations. They colorize the target geometry with a single color representing the level of attention (model of interest). More fine-grained details provide the surface-centered visualizations. Similar to a 2D heatmap they create specific textures representing the distribution of visual attention over the object [Stellmach et al. 2010].

In the literature, all object-centered visualizations are used to visualize information about 3D point of regards calculated by geometry-based algorithms. They are, however, not generally bound to this, as the coloration could also be determined by projecting holistically estimated point of regards onto the surfaces of the objects.

Recently, there are several approaches to transfer the notion of attention maps from 2D content to 3D content. The so-called 3D *Attention Volumes* [Pfeiffer 2010; Pfeiffer 2011] model the 3D point of regards as Gaussian distribution in space. For the generation of the visualization, a volume-rendering technique is used. Pirri et al. presented later an alternative approach called 3D *Saliency Maps* [2011]. They model the 3D point of regards as single points in space, which are then scored and aggregated in bundles. The bundles are in turn used to identify objects of regard, thus 3D saliency maps are a mixture between a 3D heatmap and 3D regions of interest.

Both approaches, the 3D Attention Volumes and the 3D Saliency Maps, use data from a holistic estimation of 3D point of regards. Based on the descriptions and the choice of naming, it seems as if the 3D Attention Volume approach comes from a background of describing data from eye-tracking experiments, while the 3D Saliency Maps are originated in research on predictive models for eye movements.

3 Measuring the point of regard in 3D

In the section on related work, the work has been clustered around two primary methods to estimate the 3D point of regard: geometry-based and holistic estimations. We will discuss the two methods in this section and develop a general procedure and technical set-up for the construction of measuring systems for 3D point of regards of a moving observer. For this, we will extend and combine several of the approaches described in the related work.

Geometry-based estimation For the geometry-based estimation of the 3D point of regard, at least one eye has to be tracked, but not necessarily both. Preferably this is the default dominant eye. In combination with some sort of motion tracking, the visual axis of that eye can be reconstructed in 3D space. The eye-tracking system only needs to be calibrated to a single plane in space, as it is common with desktop-based systems.

To arrive at an estimation of the 3D point of regard, the distance of the focus of gaze has to be determined. Geometry-based approaches use an up-to-date geometry model for this. Their assumption is, that

the first object that is hit by the reconstructed visual axis is the target of the fixation and thus the intersection of this very object and the visual axis is taken as the 3D point of regard. This requires that the geometry model is a good approximation of the real world and if dynamic scenarios are considered, the model needs to be updated whenever geometry moves or changes.

While the geometry-based estimation suggests to deliver rather precise data, it turns out that its underlying assumptions are rather strong. First of all, the assumption that the first geometry to be hit by the visual axis is the one fixated at does not hold in general. Trivial examples where this assumption breaks are windows or other transparent objects, mirrors or see-through Head-Up-Displays where the observer might actually look beyond (or through) the first geometry. This is especially relevant, as eye tracking is often used in the analysis of cockpit design and assistive functions for drivers. Other examples are filigree objects smaller than the local accuracy of the eye-tracking system, such as text or fences. They might easily be missed and thus not be considered.

This last problem is increased even more when considering that our attention is not restricted to the single small point where our visual axis intersects a target. It is rather sufficient to bring the target within the area of high visual acuity. Thus, the assumption that this intersection point is the 3D point of regard is also too strong. Instead of the visual axis, which is basically a line or a vector, a cone should be used to compute the intersection. The opening angle of this cone would then match the angle of high visual acuity.

Also, if the target is fixated with one eye, but the sight of the other eye is occluded by another object, it is difficult to arbitrate which object really is the target of the fixation. In the case of monocular eye tracking, the eye under observation will win, but this is not necessarily the dominant eye in this situation.

Holistic estimation The holistic estimation requires the greater technical effort. It works best if both eyes are tracked with a binocular system. In addition, the calibration procedure may require additional effort, as for some methods [Essig et al. 2006] the reference points are not only distributed over a plane in space, but over a volume and thus the required number of reference points increases from 9/16 to 27/48 or even more. If the holistic estimation is done with a monocular system integrating over time, it has necessarily a higher latency than binocular approaches, will provide irregular sampling intervals and the required perspective shifts accumulate errors over several measurements.

The holistic estimation is independent of a geometric model and thus does not require a timely update of the model. This makes holistic estimations ideal candidates for measuring attention in real world experiments (see Fig. 4). If, however, the attention is to be analyzed on an object-level, e.g., similar to regions of interest, models required. These models can be of coarse resolution and algorithms such as nearest neighbor can be used to map 3D point of regards to close object geometries.

In principle, holistic estimations also provide a single point in space as estimation of the 3D point of regard. They are thus subject to similar arguments as the geometry-based estimations regarding the spread of the area of high acuity. However, most holistic approaches take this into account, as they require an additional mapping step to identify the model of interest based on the 3D point of regard, which is not necessary for the geometry-based estimations.

The holistic estimation anchors a 3D point of regard absolutely in the 3D world (world centered), while the geometry-based estimation provides in addition a position of the 3D point of regard on the target object the geometry is a part of (object centered). This makes

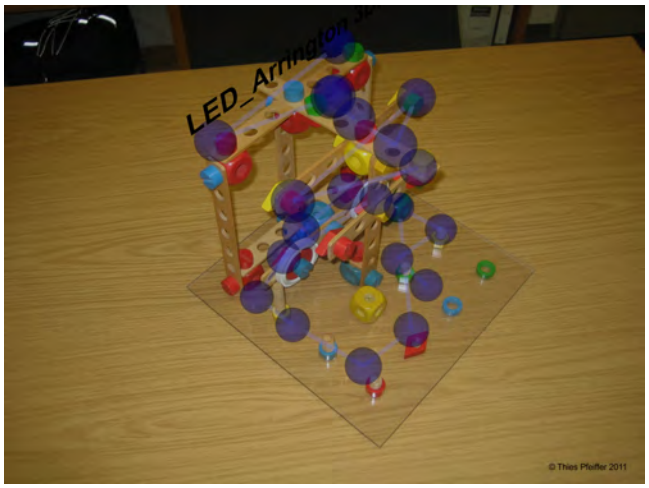


Figure 4: *Using the holistic estimation of the 3D point of regard, the attention over real-world objects can be derived without a geometric model of the environment. Here, a 3D scanpath on a real-world copy of the 3D construct is shown.*

the geometry-based estimation a better candidate if the scenario under consideration includes moving or changing objects. The holistic estimation will require an additional step to map the global 3D point of regard to a local 3D point of regard in object space. On the other hand, based on the volume of high acuity around a 3D point of regard, which could be called 3D *volume of attention*, visual attention can be computationally spread over several objects, which are close together. Instead of a 3D point of regard on a single object, as in the geometry-based estimation, estimations of 3D volumes of regard can be computed for several objects. The holistic estimation will then provide a more realistic model of attention.

3.1 General procedure and technical set-up

For the measurements of the 3D point of regards, the following units of equipment can be identified:

- eye tracker: monocular for a geometry-based estimation, binocular for geometry-based or holistic estimation.
- geometry model database: for geometry-based solutions, a database holding the detailed environment model is required. For smaller environments this could be a 3D scenegraph. This requirement could also include an active process for acquiring the environment model in real-time, such as SLAM (simultaneous localization and mapping) [Leonard and Durrant-Whyte 1991].
- body tracking/localization: for small spaces an outside-in tracking system, such as a VICON [Vicon Motion Systems 1984] or an ART [advanced realtime tracking GmbH 2009] system is required. Larger spaces should be approached with an inside-out tracking system, such as AR toolkit [Kato and Billinghurst 1999] or SLAM.
- data fusion unit: integrates the information provided by the eye tracker and the tracking/localization system, has a chain of matrix transformations to map the incoming coordinates to a coherent world model.
- solution for calibration: depends on the overall set-up, this could be a laser-pointer, one or more visual markers, a projection screen, or something else.

- 3D point of regard estimating unit: implementation of either a geometry-based or a holistic estimation algorithm which is fed by the data generated in the data fusion unit.

The first two requirements and the last have already been discussed. The third requirement has also been mentioned, but the distinction between outside-in and inside-out tracking has not been drawn.

For an **outside-in** tracking, active tracking sensors are attached to the environment. Their range of operation covers a certain tracking volume. The human is tracked purely optically or, in most cases, is marked with a certain tracking target. This target is lightweight and does not hinder the human's movements. Examples of this are the optical marker-based tracking systems sold by VICON or ART, just to name a few. In addition to being a lightweight solution, the accuracy of such a marker-based tracking system is very high (sub-millimeter range). However, the tracking volume covered by such a system is suited for laboratories or smaller shops, but not for large supermarkets or outdoors. In the ideal-case no markers are required and the user is tracked purely optically. However, while there is large progress in this area of research, the accuracy and frame-rate of such marker-less systems has not reached the performance of the marker-based systems.

Inside-out tracking works the other way around: the sensors are attached to the human. If it is a marker-based system, the markers are placed at distinct points in the environment. Such systems easily cover large spaces. The drawback is, however, the additional, sometimes obtrusive, payload on the human. Also, these systems have not reached the accuracy of outside-in tracking systems yet.

The eye-tracking system provides at least the orientation of one eye. The body tracking system provides the position and orientation of the human's head. Both information are fed into a **data fusion unit**, whose task is the construction of the direction of gaze in absolute world coordinates. This process is described in more detail in the following section.

Both the eye-tracking system and the data fusion unit may require **calibration** information to adjust the computations to the specific human user, which is described in Section 3.3.

3.2 Data Fusion Unit

The input of the data fusion unit consists of a matrix M_H (H for head) describing the position and orientation of the head of the human in an absolute world coordinate system provided by the tracking system, as well as one or two rotations describing the orientation of the eye(s) M_{LE}, M_{RE} (LE/RE for left eye/right eye) provided by the eye-tracking system.

The data fusion unit further requires information about the position of the eyes in relation to the origin of the head $M_{RP\rightarrow LE}$, $M_{RP\rightarrow RE}$ (RP for reference point). The origin of the head is a certain point of reference, which is either defined directly by the marker of the tracking system, or a distinct point, such as the point between the eyes. In the latter case, an additional transformation $M_{H\rightarrow RP}$ is needed to describe the relation between the marker attached to the head and this point of reference. The transformation from the reference point to the eye (and to the marker on the head) has to be determined only once, e.g., during a calibration step. The full chain of transformations is: $M_H M_{H\rightarrow RP} M_{RP\rightarrow LE} M_{LE}$ for the left and $M_H M_{H\rightarrow RP} M_{RP\rightarrow RE} M_{RE}$ for the right eye respectively.

3.3 Calibration

In a preparation step, the transformations $M_{H\rightarrow RP}$ and $M_{RP\rightarrow(L/R)E}$ have to be determined. The transformation $M_{H\rightarrow RP}$ can be fixed if the marker is attached to the eye-tracking gear. Otherwise, all transformations have to be measured. Depending on the body-tracking/localization system, an additional calibration could be required for the calculation of M_H. Marker-based outside-in tracking systems typically require such a calibration procedure only once when the sensors have been arranged in the environment or if vibrations could have altered the sensor orientations.

Optical eye-tracking systems typically also require a calibration procedure in which the system adjusts its parameters to the particularities of the eyes of the current human user. All vendors of eye-tracking systems known to the authors support a calibration procedure in their software which requires a sequence of fixations on reference points presented on a 2D plane. Automatic approaches thereby often use a computer screen for the presentation of the 2D plane. Other systems, such as mobile ones, support a semi-automatic calibration procedure in which the operator highlights a sequence of reference points in the field of view of the human user of the eye-tracking system, e.g., by using a laser pointer. After the calibration procedure, the system has learned a mapping from the detected features of the eye in the coordinate system of the camera targeted at the eye to a 2D coordinate system defined by the 2D plane used for calibration. During the calibration procedure, the human user normally has to remain motionless, which is sometimes even supported by using a chin rest. Alternatively, additional compensation mechanisms have to be implemented based on a tracking of the human's head movements.

Given the estimated 2D gaze position provided by the eye-tracking system relatively to the 2D plane used for calibration and the distance from the human's head to the 2D plane during calibration, the orientations $M_{(L/R)E}$ of the left and right eye can be computed, if it is not already provided by the eye-tracking system directly.

3.4 Own set-up and data acquisition

Our own set-up consists of the DTrack2 tracking system by ART, which is an outside-in marker-based tracking system. For the eye-tracking, an Arrington Research BS007 system with View Point software is used (see Fig. 1), as has been suggested by Pfeiffer [Pfeiffer 2008].

Based on the argumentation in Section 3, we decided to use a holistic estimation of the point of regard. For this we implemented the approach described by Essig et al. for binocular eye-tracking systems [Essig et al. 2006]. The data fusion of head and gaze was done as described in Section 3.2.

For the calibration, we developed an automatic approach for a free moving observer. This was done by using a projection system to present the reference points for the calibration procedure on a wall. The positions of the reference points were then adjusted dynamically to the perspective of the human user by taking into account the position and orientation of the head. This way, the reference points were moving with every head movement, but at the same time remained stable in relation to the center of the eye.

To test the set-up and record 3D point of regards, the eye movements of ten participants were recorded. The participants were asked to fixate a series of objects presented in a fixed sequence. The sequence was given verbally by the instructor. The target objects were arranged in a small construct depicted in Fig. 3, inside a cubic volume with an edge length of 30 cm. Individual objects were small, with sizes of about 2 to 4 cm. The construct was both mod-

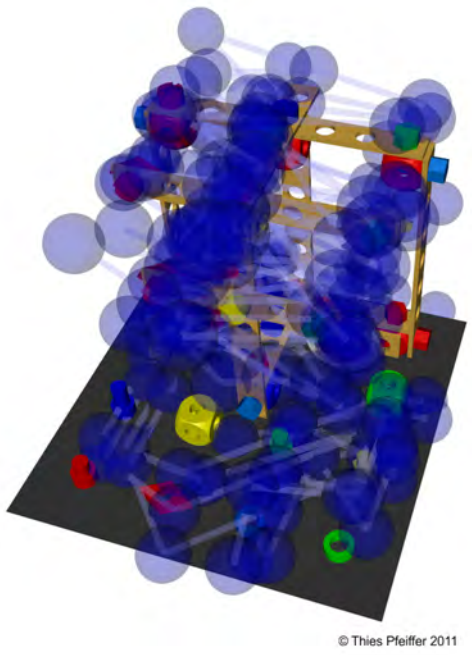


Figure 5: *3D scanpaths, however, are not suited to aggregate data over several persons. The figure shows the 3D scanpaths of ten observers.*

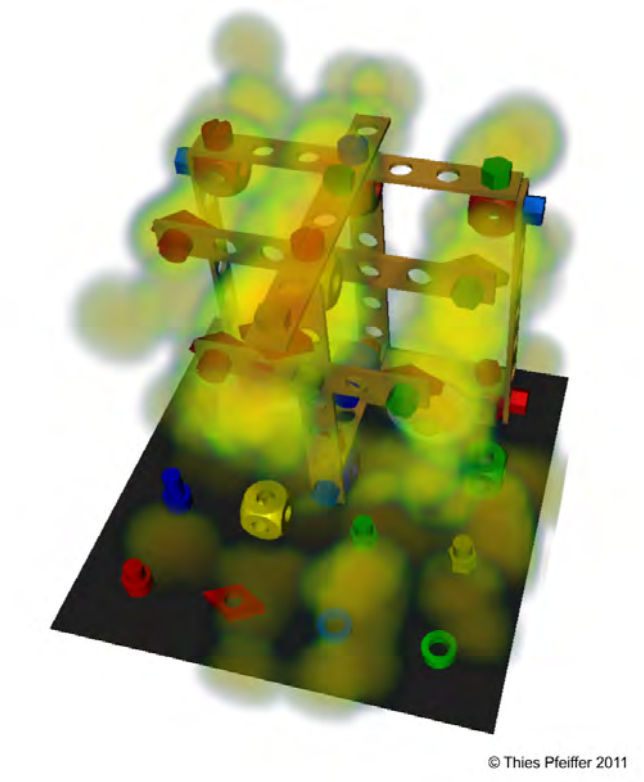


Figure 6: *3D Attention Volume of the same data-set as used for Fig. 5. The volumes that received a higher amount of attention pop out more clearly using the volume-based rendering with color coding. The effect is even more intriguing when exploring the visualization interactively.*

eled with virtual objects as depicted in Fig. 3 and with real objects as depicted in Fig. 4.

4 Visualizing 3D point of regards

The recorded 3D point of regards were visualized using two methods. First, a naive 3D scanpath visualization was developed, to assess the quality of the data. Second, an extended 3D attention volume model was developed.

4.1 3D scanpaths

The resulting 3D point of regards for an individual participant are depicted in Fig. 3 and Fig. 4 as 3D scanpath. In this 3D scanpath the 3D point of regard is extended to spheres approximating the volume of high acuity around the estimated 3D point of regard. The radius of the spheres is thereby determined by the visual angle of high acuity around the optical axis. This angle is about 2.5° for the area of the fovea. The function $POR_{sphere}(\vec{x})$, POR stands for point of regard, decides for every point in space, whether it is a member of the fixation or not, based on an implicit definition of a 3D sphere.

$$\begin{aligned} POR_{sphere}(\vec{x}) &: (\vec{x} - p_{\vec{POR}})^2 \leq r(\vec{x})^2 \\ r(\vec{x}) &: |\vec{x} - p_{\vec{eye}}| \tan\alpha \\ \text{with } POR_{sphere} &: \text{membership function for } \vec{x} \\ \vec{p}_{POR} &: \text{3D point of regard} \\ \vec{p}_{eye} &: \text{3D position of the observing eye} \\ \alpha &: \text{angle of high visual acuity} \end{aligned} \tag{1}$$

If data from several participants should be aggregated, 3D scanpaths are no longer a helpful visualization, as shown in Fig. 5.

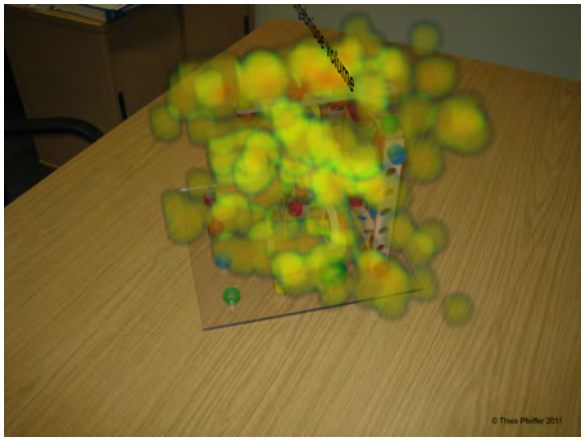
4.2 3D Attention Volumes

Before we provide our own definition of a 3D Attention Volume, we reconsider that the way the visualization of the 3D point of regard expressed in (1) is already a description of a volume. Points in space for which $POR_{sphere}(\vec{x})$ holds are members of this volume, all others are outside the volume. $POR_{sphere}(\vec{x})$ thus already defines a 3D *volume of regard*, instead of a single 3D point of regard.

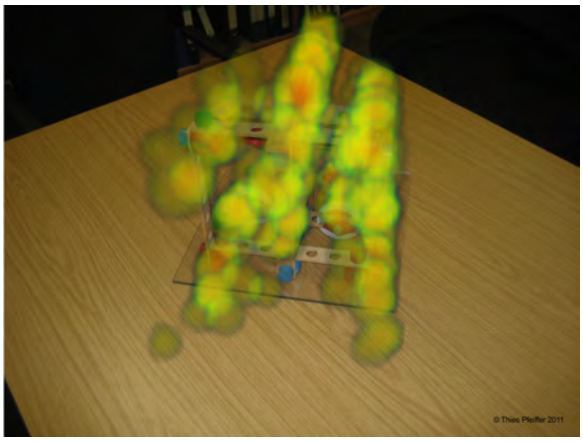
In our model of the 3D Attention Volume (3DAV), the discrete membership function defining the sphere is replaced by a continuous weighting function $3DAV(\vec{x})$. This weighting function realizes a Gaussian distribution around the measured 3D point of regard. The Gaussian distribution models the acuity around the visual axis and thus provides a fine-grained approximation of the distribution of visual attention in space. In addition, the distribution is slightly distorted in depth by taking the area of high visual acuity into account for every single point in space, thus the distribution gets broader the more distant it is from the observing eye.

$$\begin{aligned} 3DAV(\vec{x}) &: d(t) e^{-\frac{|\vec{x}-\vec{p}_{POR}|^2}{\sigma(\vec{p}_{eye},\vec{x})}} \\ \text{with } d(t) &: \text{amplification factor depending on the duration} \end{aligned} \tag{2}$$

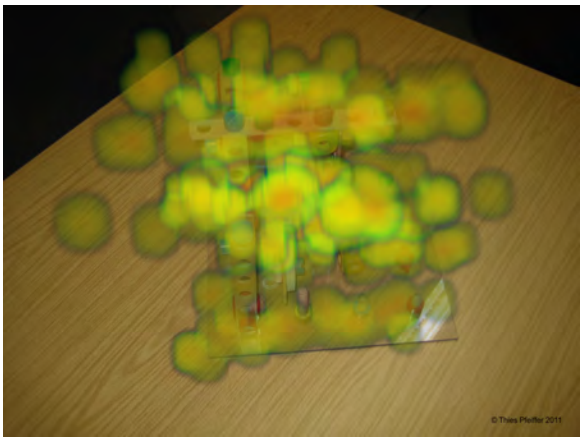
The function $3DAV(\vec{x})$ assigns a value to each point in space which represents the visual attention that has been spent on this



(a) *Picture of the scene from the left side of the construct*



(a) *Picture of the scene from the back side of the construct*



(b) *Picture of the scene from the right side of the construct*

Figure 7: *The figure shows a 3D Attention Volume of data recorded on real-world objects. The volume is rendered from the perspective of the camera and overlayed on the camera image in a post-processing step. The procedure is similar to that used in augmented reality, where 3D objects are rendered into the live view of a camera.*

Figure 8: *Additional perspectives can be rendered from any perspective, independently of the original recordings. With appropriate tools yet to be developed, the recorded 3D Attention Volumes could be reviewed interactively using augmented reality by overlaying the real scenery with the rendered volumes.*

point. The amplification factor $d(t)$ amplifies the distribution depending on the duration of the fixation. Longer durations will lead to higher amplitudes of the Gaussian function.

Aggregated visualizations of 3D Attention Volumes for multiple fixations and participants can be created by integrating over all the 3D Attention Volumes for the individual fixations and normalizing the values afterwards.

The 3D Attention Volumes can be visualized using volume rendering techniques (see Fig. 6 for the 3D Attention Volume for all 10 participants). Following the color-coding of 2D heatmaps, levels of high visual attentions are given a red shading and less warmth colors are used to shade lower levels of visual attention.

The volume-rendering approach also does not require knowledge about or the presence of object geometries. It can thus be used together with all kinds of methods to estimate the 3D point of regards. In particular, 3D Attention Volumes can even be used to depict visual attention on real 3D environments.

As the 3D Attention Volume models are independent of perspective, they can be rendered from different views and tracking shots can be created for offline viewing (see Fig. 7 and Fig. 8). Knowledge about the geometries, however, could be used to increase the visual quality of the rendering, for example, to correctly consider partial occlusions of 3D Attention Volumes by foreground objects.

5 Conclusion

The established methods for measuring visual attention are restricted to 2D stimuli. After a review of techniques to assess the 3D point of regard and a presentation of current visualizations of visual attention, we presented our own approach for measuring the 3D point of regard, which extends holistic estimations previously developed for desktop-based set-ups to 3D set-ups with a moving observer, either with virtual or real-world 3D objects as targets.

The volume-based modeling of visual attention can be seen as the logical extension of the established attention map models for 2D content to 3D. In fact, other 3D visualizations, such as the surface-based ones, turn out to be special cases of 3D Attention Volumes: they are the projections of the volumes on a surface.

Using volume-based rendering, 3D Attention Volumes can be visualized interactively as an overlay on the target objects. Together with the 3D scanpaths and the object- and surface-based visualizations there are now pendants for all of the established 2D visualizations of visual attention available to assess visual attention in 3D space.

Advances can be expected from a more fine-grained modeling of the 3D extension of the volume of visual high acuity. The presented model approximates this volume roughly using a Gaussian distribution, which is comparable to the approximations used for 2D heatmaps. The reality, however, is much more complex.

In practice, methodical aspects play an important role. The 3D attention tracking system should be easy and fast to setup and calibrate. In addition, it should support long interaction periods without interceptions by re-calibrations or drift corrections of the gear. In 3D scenarios, however, the user will naturally move around - in contrast to the 2D condition where the users remain seated and rather motionless in front of a computer screen. Thus while the chain

of tools and methods for assessing visual attention in 3D space is now complete, it would require several more iterations to make it as convenient to operate as the 2D tools of today.

The increased interests in interactions of people in natural settings and novel mobile uses of interaction technology demand for tools to record and visualize eye-gaze patterns in natural 3D environments. The review in this paper gives reason to raise the expectations that these demands will be satisfied in the near future.

Acknowledgements

This project has been partly funded by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center SFB 673, "Alignment in Communication".

References

ADVANCED REALTIME TRACKING GMBH, 2009. Homepage. Retrieved October 2011 from `http://www.ar-tracking.de`.

BOJKO, A. A. 2009. Informative or misleading? heatmaps deconstructed. In *Proceedings of the 13th International Conference on Human-Computer Interaction. Part I: New Trends*, Springer-Verlag, Berlin, Heidelberg, 30–39.

BOLT, R. 1981. Gaze-orchestrated dynamic windows. *Proceedings of the 8th annual conference on Computer graphics and interactive techniques*, 109–119.

DODGE, R. 1907. Studies from the psychological laboratory of wesleyanuniversity: An experimental study of visual fixation. *Psychological Monographs*.

DUCHOWSKI, A. T., MEDLIN, E., GRAMOPADHYE, A., MELLOY, B., AND NAIR, S. 2001. Binocular Eye Tracking in VR for Visual Inspection Training. In *Virtual Reality Software and Technology ACM: Symposium on Virtual reality software and technology*, ACM Press, S. A. S. I. G. on Computer-Human Interaction und SIGGRAPH: ACM Special Interest Group on Computer Graphics and I. Techniques, Eds., 1–8.

DUCHOWSKI, A. T., COURNIA, N., CUMMING, B., MCCALLUM, D., GRAMOPADHYE, A., GREENSTEIN, J., SADASIVAN, S., AND TYRRELL, R. A. 2004. Visual Deictic Reference in a Collaborative Virtual Environment. In *Eye Tracking Research & Applications Symposium 2004*, ACM Press, San Antonio, TX, 35–40.

ESSIG, K., POMPLUN, M., AND RITTER, H. 2006. A neural network for 3D gaze recording with binocular eye trackers. *The International Journal of Parallel, Emergent and Distributed Systems 21*, 2, 79–95.

KATO, H., AND BILLINGHURST, M. 1999. Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In *Augmented Reality, 1999.(IWAR'99) Proceedings. 2nd IEEE and ACM International Workshop on*, IEEE, 85–94.

KWON, Y.-M., JEON, K.-W., KI, J., SHAHAB, Q. M., JO, S., AND KIM, S.-K. 2006. 3D Gaze Estimation and Interaction to Stereo Display. *The International Journal of Virtual Reality 5*, 3, 41–45.

LEONARD, J., AND DURRANT-WHYTE, H. 1991. Simultaneous map building and localization for an autonomous mobile robot. In *Intelligent Robots and Systems' 91.'Intelligence for Mechanical Systems, Proceedings IROS'91. IEEE/RSJ International Workshop on*, Ieee, 1442–1447.

MARR, D. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. Freeman, San Francisco.

MEINER, M., AND DECKER, R. 2010. Eye-tracking information processing in choice-based conjoint analysis. *International Journal of Market Research*, 5, 591–611.

NORTON, D., AND STARK, L. 1971. Scanpaths in saccadic eye-movements during pattern perception. *Science*, 308–311.

PFEIFFER, T., LATOSCHIK, M. E., AND WACHSMUTH, I. 2009. Evaluation of Binocular Eye Trackers and Algorithms for 3D Gaze Interaction in Virtual Reality Environments. *Journal of Virtual Reality and Broadcasting 5*, 16 (jan).

PFEIFFER, T. 2008. Towards Gaze Interaction in Immersive Virtual Reality: Evaluation of a Monocular Eye Tracking Set-Up. In *Virtuelle und Erweiterte Realität - Fünfter Workshop der GI-Fachgruppe VR/AR*, Shaker Verlag GmbH, Aachen, M. Schumann and T. Kuhlen, Eds., 81–92.

PFEIFFER, T. 2010. Tracking and Visualizing Visual Attention in Real 3D Space. In *Proceedings of the KogWis 2010*, Universittsverlag Potsdam, Potsdam, J. Haack, H. Wiese, A. Abraham, and C. Chiarcos, Eds., 220–221.

PFEIFFER, T. 2011. *Understanding Multimodal Deixis with Gaze and Gesture in Conversational Interfaces*. Berichte aus der Informatik. Shaker Verlag, Aachen, Germany, December.

PIRRI, F., PIZZOLI., M., RIGATO, D., AND SHABANI, R. 2011. 3d saliency maps. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 9–14.

POMPLUN, M., RITTER, H., AND VELICHKOVSKY, B. 1996. Disambiguating complex visual information: Towards communication of personal views of a scene. *PERCEPTION-LONDON- 25*, 931–948.

RÖTTING, M., GÖBEL, M., AND SPRINGER, J. 1999. Automatic object identification and analysis of eye movement recordings. *MMI-Interaktiv 2*.

SMI, 2011. Smi eye tracking glasses homepage, September. Retrieved October 2011 from `http://eyetracking-glasses.com/`.

STELLMACH, S., NACKE, L., AND DACHSELT, R. 2010. 3d attentional maps: aggregated gaze visualizations in three-dimensional virtual environments. In *Proceedings of the International Conference on Advanced Visual Interfaces*, ACM, 345–348.

TANENHAUS, M. K., SPIVEY-KNOWLTON, M. J., EBERHARD, K. M., AND SEDIVY, J. C. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science 268*, 1632–1634.

TANRIVERDI, V., AND JACOB, R. J. K. 2000. Interacting with eye movements in virtual environments. In *Conference on Human Factors in Computing Systems, CHI 2000*, ACM Press, New York, 265–272.

TOBII TECHNOLOGY AB, 2010. Tobii glasses Homepage, June. Retrieved October 2011 from `http://www.tobiiglasses.com`.

VICON MOTION SYSTEMS, 1984. Homepage. Retrieved April 2010 from `http://www.vicon.com`.

YARBUS, A. L. 1967. *Eye Movements and Vision*. Plenum Press, New York.