Influence of disparity on fixation and saccades in free viewing of natural scenes

Lina Jansen
Institute of Cognitive Science, University of Osnabrück, Germany

Selim Onat
Institute of Cognitive Science, University of Osnabrück, Germany

Peter König
Institute of Cognitive Science, University of Osnabrück, Germany

Humans select relevant locations in a scene by means of stimulus-driven bottom-up and context-dependent top-down mechanisms. These have mainly been investigated by recording eye movements under 2D natural or 3D artificial stimulation conditions. Here we try to close that obvious gap and presented 2D and 3D versions of natural, pink, and white noise images to human subjects. Importantly, ground truth distance was obtained for all image pairs by laser scanning. Recording eye movements, we investigated the influence of disparity information and of higher order scene correlations on basic saccade properties and on saliency of bottom-up information. Our results show that the removal of higher order correlations changed saccadic rate, length and main sequence, and the subjects’ explorative behavior. Introduction of disparity information countered these effects and alleviated differences between image categories. Disparity information had no effect on the saliency of monocular image features like luminance and texture contrast; however, without higher order correlations these features were uncorrelated to fixation locations. An analysis of binocular image features revealed that participants fixated closer locations earlier than more distant locations. Importantly, this also held for 2D natural images. Taken together, we conclude that depth information changes basic eye movement properties and provides a salient image feature.

Keywords: overt attention, binocular vision, natural stimuli, eye movements


Introduction

Under natural conditions, sensory organs are overwhelmed by a continuous inflow of signals. The sheer amount of available information has promoted the idea that it is behaviorally relevant signals that are selected for detailed processing, nicknamed “selection for action” (Rafal, Danziger, Grossi, Machado, & Ward, 2002; Rafal, Ward, & Danziger, 2006). An important contribution to such attentional processes is the directing of sensory organs toward the most relevant signal source. Overt attention in the form of exploratory eye movements brings selected regions of the visual scene onto the fovea for processing with high spatial detail. Hence, studies of eye movements provide a valuable opportunity to investigate the control of attention.

Two main factors contribute to the control of overt attention. On the one hand, top-down signals mediate information about the present task as already demonstrated by Yarbus (1967). On the other hand, low-level stimulus properties influence the deployment of attention in a bottom-up manner (Wolfe & Horowitz, 2004). Recent research proposes that these two types of signals involve separate cortical areas (Corbetta & Shulman, 2002). This raises the question of which factors contribute to top-down and bottom-up signals and how these signals interact.

A variety of models of overt attention have been proposed (Deco & Rolls, 2004; Hamker, 2004; Itti & Koch, 2000; Torralba, 2003). They range from demonstrations of principled mechanisms inspired by neuroscience to detailed engineered solutions for computer vision. Many of these concentrate on bottom-up mechanisms and are based on the concept of the so-called saliency map (Koch & Ullman, 1985). The basic saliency model of Itti and Koch (2000) is based on various image feature maps, like luminance, orientation, and color maps, which are extracted from the stimulus. On each feature map, center-surround differences are computed and the resulting maps are combined into a topographically organized saliency map. This map represents the conspicuity of each location in the image and is used together with an inhibition of return mechanism to predict where humans would fixate in an image.

In such saliency-based models, the selection of relevant features is partly inspired by the receptive field properties...
of neurons in early sensory areas and partly due to available psychophysical results. Several studies have reported a correlation between image feature values and the selection of fixated locations. For example, Reinagel and Zador (1999) reported that luminance contrast is higher at fixated than non-fixated regions, a finding which has been replicated in many other studies (Einhäuser & König, 2003; Parkhurst, Law, & Niebur, 2002; Tatler, Baddeley, & Gilchrist, 2005). Similarly, an increase of other low-level features, like texture contrast, has been observed (Parkhurst & Niebur, 2004). These correlations explain a substantial fraction of the variance in the distribution of selected fixation points. They are, however, far from perfect (Tatler, 2007; Tatler et al., 2005) and their relevance for the selection of fixation locations has been questioned. Three reasons against the relevance of this correlation are provided: First, some studies show that these correlations can be absent or even reversed when participants perform specific tasks such as searching for a target in an outdoor scene (Einhäuser, Rutishauser, & Koch, 2008) or walking to a target in a real-world environment (Turano, Geruschat, & Baker, 2003). If image features have a general bottom-up impact on the selection of fixation points, it is expected that the influence of the saliency on the oculomotor mechanisms must be present irrespective of the task. Second, models based on saliency miss important aspects of eye movements such as repetitive scanpaths on reviewed images (Foulsham & Underwood, 2008; Noton & Stark, 1971), saccade sequences, region-to-region saccades (Henderson, Brockmole, Castelhano, & Mack, 2007), and large variability on the fixation durations (Hollingworth & Henderson, 1998). Thus, saliency can only play a partial role in the selection of fixation locations. Third, it has not yet been decided whether these correlations represent a causal influence or arise due to correlation with salient higher order regular structures like objects. In this context, Baddeley and Tatler (2006) reported that using a model that incorporates edge, contrast, and luminance maps on two different scales they found that not luminance contrast but high spatial frequency edge information was the most salient feature. As this feature is more directly related to higher order structures, it questions the causal influence of low-level features. Supporting the doubt about causality, Henderson et al. (2007) reported that the higher amount of fixations on salient objects can be explained by the finding that salient regions are more informative. In summary, it is an open debate whether the correlation between image features and fixation selection is still present in real-world setups and, if a correlation exists, whether the correlation is causal or based on correlations with higher order regularities like objects.

Many image features have already been analyzed for their relevance in attentional processes but to give a complete list of relevant visual features is a difficult task (Wolfe & Horowitz, 2004). Moving from the argument that attention is selection for action and objects are important for action, cues communicating depth information are natural candidates for inclusion in this list. Human actions often involve manipulating objects that lie at a certain distance that is distinct from the background. Hence, depth cues potentially provide important information about behaviorally relevant objects in the visual field and should be regarded as relevant features. In contrast to well investigated image features like luminance and texture contrast, depth as a feature is more directly bound to objects, as irregularities in depth are mainly associated with changes of distinct objects.

In natural scenes, various depth cues are present (for a review, see Howard & Rogers, 1995). Monocular cues like linear perspective, texture gradient, interposition, and shading are combined with binocular cues like accommodation and stereo disparity. The latter cue describes the shift between the locations on the left and right retina on which a point in a 3D scene is projected. This shift is inversely proportional to the distance of the point from the viewer and thus is mainly used for depth perception at short and medium distances. Monocular cues alone are often sufficient to distinguish the potential ordering of objects in a scene and to estimate the depth of these objects. The depth information that is deduced from monocular cues will be called “deduced” and can be considered analogous to the association of motion to potentially moving objects in static scenes. From the list of binocular cues that establish depth perception, disparity is very important, as pure disparity information can mediate form vision (Julesz, 1960). Further reinforcing the importance of disparity information, neuroscientific studies have shown that the early visual cortex of cats and primates contain cells that are sensitive to disparity (Barlow, Blakemore, & Pettigrew, 1967; Cumming & Parker, 1999).

Based on these findings, Itti and Koch (2001) suggested that stereo disparity could also be a relevant attentional cue. Indeed, depth cues were included in a number of previous experimental studies. In most of these studies, random-dot stereograms or autostereograms were used to establish a depth percept. Using this technique, studies showed that humans shift their attention across depth planes when the perceptual load of the task is high (Atchley, Kramer, Andersen, & Theeuwes, 1997). Further, attentional shifts from far to near objects were found to be faster than shifts from near to far objects, but this viewer-centered asymmetric gradient also depended on the perceptual load (Arnott & Shedden, 2000). A recent study by Parks and Corballis (2006) generalized these findings to pictorial depth stimuli. However, in all of these studies only artificial stimuli were presented. Thus, little is known about the contribution of depth and especially disparity in natural scenes to overt attention.

In the present study, we analyzed eye movements of human subjects viewing natural 3D stimuli. A laser scanner and stereo camera setup were used to acquire natural stereoscopic images together with their ground truth disparity. From these images, 1/f noise (pink noise)
and white noise stimuli were generated that contained the same disparity information as their natural counterparts. The noise stimuli allowed an analysis of the influence of the binocular features independent from their correlation with higher order statistics. Because of the accuracy of the laser scanner, the image statistics of the noise images were not confounded by the quality of depth reconstruction. The 2D and 3D versions of the natural, white noise, and pink noise images were presented for 20 s on a 3D display while eye movements of human subjects were recorded in a free viewing task. This present setup allowed a detailed analysis of the influence of disparity information on the control of overt attention. Further, it allowed a direct comparison of the role of previously analyzed image features in conditions with (natural images) and without (noise images) object-related correlations.

Using the recorded data, we investigated on the one hand differences in the ability to perceive depth in natural and noise images and differences in basic eye movement statistics (such as saccade length or fixation duration) in 2D and 3D images. On the other hand, the saliency of the image features mean luminance, luminance contrast, texture contrast, mean disparity, and disparity contrast was investigated. Thus, we answer 4 questions:

1. Are participants able to recognize 3D structure equally well in all 3D conditions?
2. Are eye movement properties different when viewing 3D images compared to 2D images?
3. Does the saliency of 2D image features change with the introduction of disparity?
4. Are binocular image features salient?

Materials and methods

Stimulus acquisition

We acquired stereoscopic images using two identical digital cameras (5.0-megapixel CCD, Sony DSC-V1, Tokyo, Japan) mounted on a stereo rig with an inter-lens distance of 7.5 cm. All images show forest scenes containing no man-made objects. The depth information of these photographed scenes was measured using a 3D laser scanner (Sick AG, Waldkirch, Germany). By computing the geometric relation between the scanner and the cameras, the relation between the 3D laser scan and 2D image coordinates was determined. Disparity maps were then generated by projecting the measured depth information onto the images and finding point correspondences between the stereo images (as described in Martin et al., 2007). The resulting maps were restricted to a disparity range between −80 and 80 pixels.

From the stimulus pool, 28 stereo images were selected, undistorted, cropped to 1280 × 1024 pixels, and converted to grayscale. The left and right images of each stereo image pair were aligned with respect to each other by maximizing their cross-correlation. This displacement improves the depth percept by reducing the overall disparity and normalizing the disparity values such that locations perceived on the screen have zero disparity, locations in front of the screen negative, and locations behind the screen positive disparity.

Stimulus generation

Each generated image pair was used together with its disparity information to generate one set of six stimuli: 2D and 3D versions of natural, pink noise, and white noise images. In the 3D natural condition (3DN), a stimulus consisted of the left and right stereo image pair, while in the 2D natural condition (2DN) the pair consisted of two copies of the left image. In addition to the natural images, we generated pink and white noise images to study the influence of disparity without any correlated influence of other image features. The 3D noise images were generated from the measured disparity maps and, thus, contained natural disparity information.

Each left pink noise image was generated by randomizing the phase spectrum of the corresponding natural image while keeping the amplitude spectrum constant and respecting the symmetry of the phase spectrum. The corresponding right image was constructed by copying the left image and shifting the pixels according to the disparity in the natural image. Unfilled holes were filled by applying a slightly modified version (Bernhard, 2006) of an iterative texture inpainting algorithm (Hirani & Totsuka, 1995, 1996). Due to this process, the resulting pink noise image did not contain any visible artifacts.

Each left white noise image was generated by randomly selecting 256 grayscale values. The corresponding right image was a copy of the left image but with pixels shifted according to the disparity in the corresponding natural image. Holes were filled with random white noise pixels. Pairs of left and right noise images were used for the 3D pink noise (3DP) and white noise conditions (3DW) and pairs of two copies of the left image made up the 2D noise conditions (2DP, 2DW).

The 28 generated stimulus sets were split up into three training, one position calibration, and 24 main experiment sets. The training stimuli were necessary to allow the participant to become familiar with the 3D display and the stimulus types. The natural 3D image of the position calibration set was used as a reference image for the participants to check their 3D percept.

The stimuli of all sets were modified to have the same mean luminance and RMS contrast. Figure 1 shows a 2D natural, pink noise, and white noise image of one stimulus set together with the mean disparity map for the corresponding 3D images.
Stimulus presentation

Stimuli were presented on an 18.1 inch 3D display (SeeReal Technologies, C-s 3D display, Dresden, Germany), which allows viewing of 3D images without special eyeglasses. In this technique, a beam splitter is situated in front of a TFT panel and results in all odd columns of the image being projected to the left eye and all even ones to the right eye. Thus, a 3D perception emerges by presenting a vertically interlaced stereo image pair. A 2D perception is established by vertically interlacing two identical images.

Participants

Fourteen subjects with normal or corrected-to-normal vision participated in the experiment. They either received course credit or were paid for their participation. Two subjects attended only one of the two experimental sessions.

Experimental design

The experiment was split into two sessions with two presentation blocks each. Both sessions followed the same procedure. They comprised a training period followed by two presentation blocks that were separated by a short break. In the training period, subjects saw one stimulus of each condition and were instructed to describe what they saw and to change their position until they achieved an optimal 3D perception. After the training session, the eye tracker was calibrated and the position calibration image was shown with the instruction to report if there were any artifacts and whether 3D perception was possible. If the perception was disturbed or artifacts were present, subjects were instructed to find

Figure 1. Example of 2D stimulus set consisting of natural (top, left), pink noise (bottom, left), and white noise (bottom, right) images and mean disparity map (top, right) of the corresponding 3D stimuli. For display purposes, we made red–green anaglyphs of the 3D stimuli of two stimulus sets available; please refer to the supplementary material.
a new optimal position and the eye tracker was recalibrated if necessary. Afterward, the presentation block started with the instruction to study the images carefully over the whole presentation time and to press a button during an image presentation immediately after at least two depth layers are perceived in the image. Before each stimulus presentation, a fixation cross in the middle of the screen with zero disparity had to be fixated to check the calibration of the eye tracker.

Each stimulus was presented for 20 s and eye movements were continuously recorded. The presentation order was randomized with four constraints: The stimulus pool was divided into six sets. (1) Each set entailed one image of each stimulus set. (2) The conditions were equally distributed in each set. Consecutive stimuli could not belong to the same (3) condition or (4) stimulus set.

**Eye tracking**

Eye movements were recorded using the “Eyelink II” head-mounted system (SR Research, Osgoode, Ontario, Canada). Eye position data were collected binocularly, but analyzed only for the left eye as the input to this eye was identical over 2D and 3D versions of stimuli. During the experiment, the subject’s head was stabilized using a chin rest, which stood at a distance of 60 cm from the screen. Thus, a stimulus spanned 25.9 × 34.1 degrees of visual field. Each stimulus presentation block started with a nine-point grid calibration procedure. The experiment began after a stable global mean error smaller than 0.3° was reached for each eye.

**Data analysis**

**Recognition of 3D structure**

In order to explicitly know whether subjects perceived the depth of an image, they were required to press a button as soon as they recognized at least two depth layers in the image. Performance, measured in terms of recognition rate and time, was compared across 3D categories with a Friedman test. One of the 14 subjects misunderstood the task and pressed the button in all images. We excluded his data from this analysis.

As the button press signals the start of the 3D percept and the eye tracker was recalibrated if necessary. Afterward, the presentation block started with the instruction to study the images carefully over the whole presentation time and to press a button during an image presentation immediately after at least two depth layers are perceived in the image. Before each stimulus presentation, a fixation cross in the middle of the screen with zero disparity had to be fixated to check the calibration of the eye tracker.

Each stimulus was presented for 20 s and eye movements were continuously recorded. The presentation order was randomized with four constraints: The stimulus pool was divided into six sets. (1) Each set entailed one image of each stimulus set. (2) The conditions were equally distributed in each set. Consecutive stimuli could not belong to the same (3) condition or (4) stimulus set.

**Eye tracking data analysis**

To avoid distortions at the borders of the monitor, we cropped data with a distance of less than 3.75 degrees to the borders of the monitor. In addition, we discarded the first fixation of each stimulus presentation, as each presentation started with a fixation presentation in the center and biased the fixation to this region.

**Eye movement property analysis**

To test whether the introduction of disparity has an effect on basic eye movement properties, we analyzed the rate of fixation, the length of saccades (Euclidean, horizontal and vertical length), and the duration of fixations. These properties are highly correlated, but we started by analyzing them separately and then continued by discussing the overall effect. For the relationship between saccade length and velocity, called main sequence (Bahill, Clark, & Stark, 1975), a detailed analysis is provided. As the distributions of eye movement properties were non-normal and contained many outliers, we applied nonparametric statistical methods. The distribution of saccade length was lognormal and the distribution of fixation rate negatively skewed. For these distributions, we computed the median as summary statistics. Fixation duration distributions were slightly positively skewed and had a small tendency to bimodality. As the second peak was very small and only slightly influenced the median, we use the median as a summary statistic. As a control, we made an additional analysis for these eye movement properties, using the mode as summary statistics (as suggested for fixation duration analysis by Velichkovsky, Dornhoefer, Pannasch, & Unema, 2000). So, for all properties, conditions were compared by first computing the median/mode eye movement property for each image and then performing a sign test between the property distributions containing the median/mode property values of all images belonging to a chosen condition. The difference between the medians/ modes of these distributions was used as an effect size.

For the analysis of the main sequence, we estimated for each subject and condition the relation between the length of a saccade and its mean velocity using the following equation from Becker (1991):

\[
v = \frac{l \cdot v_m}{l_{50} + l},
\]
where \( v \) is mean velocity, \( l \) is saccade length, \( l_{50} \) is the asymptotic upper limit of mean velocity, and \( v_m \) is a constant that corresponds to the saccade length at which the mean velocity reaches half its maximum value. The parameters \( v_m \) and \( l_{50} \) were estimated using least squares fitting. To test for differences between conditions, the distributions containing the estimated \( v_m \) and \( l_{50} \), respectively, were compared using sign tests. As the ratio between mean velocity and peak velocity, a measure often used for the analysis of saccade dynamics, was shown to be linear (Becker, 1989), our analysis can be compared with the analysis of the dependence between peak velocity and saccade length.

As most studies tend to use presentation durations shorter than 20 s, we computed all statistics not only for the whole recorded data set but also for the first 5 s only. Furthermore, we analyzed the time course of the median value to investigate the temporal evolution of these basic eye movement parameters. As the results for the first seconds were mostly identical with the results for the whole stimulus presentation, we mainly report the results computed over the whole stimulus presentation time and additionally include any time-dependent differences within the first 5 s or trends over the whole 20-s time period. If a trend was found, we furthermore state the more detailed analysis and the average change from the first to the fifth and the fifteenth second, respectively. To analyze the difference in eye movement properties between 2D and 3D presentations for each image category, 21 sign tests (7 features × 3 categories) were computed. No multiple comparison correction was performed, but \( p \)-values are reported. For the discussion, we considered only \( p \)-values smaller than 0.01.

**Analysis of exploration pattern**

To investigate whether subjects’ visual exploration of an image changes with different stimulus types, we generated a 2D fixation distribution for each image. Each entry in the distribution represents the probability of fixation during a chosen time interval at the corresponding position in the image. These probabilities were estimated from the fixation data of all subjects. The fixation distributions were then convolved with a two-dimensional Gaussian with full-width at half-maximum (FWHM) of one degree to account for the size of the fovea and the finite precision of the eye tracker. A measure of the explorative behavior of subjects is the inter-quartile range of the horizontal and vertical marginal distributions of the 2D fixation distributions. This statistical dispersion measure is robust against outliers, more stable than the standard deviation, and shows the spread of the 75% of the data that are closer to the median. The inter-quartile ranges of the fixation distributions originating from different conditions were compared against each other using a sign test.

**Image feature maps**

To investigate the role of image features in the allocation of attention, we computed three monocular and two binocular image features: mean luminance, luminance contrast and texture contrast; and mean disparity and disparity contrast.

We defined the mean luminance of a given location as the weighted average of luminance values with a two-dimensional Gaussian with one degree FWHM used as the weighting function. In line with previous studies, the weighted standard deviation of luminance in a Gaussian patch with FWHM of one degree is defined as luminance contrast. Texture contrast is a generalization of luminance contrast. Thus, we define it as the standard deviation of luminance contrast values in a Gaussian patch with FWHM of five degrees. The binocular image feature maps were computed from the disparity maps that were measured using the laser scanner. Mean disparity and disparity contrast were computed on these maps in the same way as for mean luminance and luminance contrast. Mean disparity is inversely proportional to the mean distance and disparity contrast is a measure of depth discontinuity.

**Control fixations**

In order to determine the role of image features in the attentional process, we compared image features at fixated locations with image features at control locations. In line with previous studies (Einhäuser & König, 2003; Tatler et al., 2005), control locations for one trial were composed of all fixations made by the subject on all images of that trial’s condition, except for fixations made on the actual image itself. Importantly, for analysis of specific time intervals, control fixations are only composed of fixations taken from the same time interval. This design ensures that general biases that are unrelated to image features, like the central fixation bias reported by Tatler (2007), are accounted for by the control structure.

**Saliency measure**

By comparing the distributions of image features at fixated locations (actual distribution) with the distribution of features at control locations (control distribution), the saliency of an image feature was determined. A higher median feature value at fixated regions than at non-fixated ones indicates that subjects tend to look more often at regions with higher feature values than expected by random scanning of the scene. This comparison does not allow the attribution of causal effects but allows us to find correlations between the allocation of attention and image features.

As image features were not normally distributed, we used nonparametric statistics. Sign tests between actual
and control distributions, which both contain one median feature value for each trial in the chosen condition, detected saliency effects of image features. The area under the receiver-operating characteristic curve (AUC) between actual and control distributions was used as an effect size. This signal detection measure was first introduced in the context of saliency analysis by Tatler et al. (2005). We estimated the AUC non-parametrically by its relation to the Mann–Whitney U statistics (Faraggi & Reiser, 2002). In this case, the AUC between actual and control distributions represents the probability that in a randomly selected pair the actual value is higher than the control value. For each trial, we computed the AUC between actual and control distributions. Because the AUC values were normally distributed, the mean AUC over the trials in one condition represents the average effect of the feature in this condition. To allow comparisons across features and across conditions, 99 percent bootstrap confidence intervals were estimated for each mean AUC. In some cases, the confidence intervals of the mean AUCs can signal a significant saliency of an image feature, although the sign test does not confirm this saliency. This effect arises because the sign test only considers for each trial whether the actual or the control median feature value is higher, while the mean AUC considers the size of the effect for each trial. In the following analysis, we interpreted the significance of the sign test and used the mean AUC only as an effect size.

We reported for each feature the effect in each condition for the whole presentation time and for the data of the first 5 s. Further, the presentation time was divided into 5-s intervals and effect sizes were computed separately for each interval. Due to the current debate about the time course of saliency in the first 5 s, we also show for this time interval the saliency effect of each image feature on a second-by-second basis. For each trial, we computed for each second the difference between the AUC value at the presentation time and then compute the mean over each second the difference between the AUC value at the second-by-second basis. For each trial, we computed for the whole presentation time and for the data of the first 5 s. Further, the presentation time was divided into 5-s intervals and effect sizes were computed separately for each interval. Due to the current debate about the time course of saliency in the first 5 s, we also show for this time interval the saliency effect of each image feature on a second-by-second basis. For each trial, we computed for each second the difference between the AUC value at the specific second and the AUC computed over the whole presentation time and then compute the mean over subjects and images. Unfortunately, the number of subjects is not sufficient to provide inferential statistics for the second-by-second analysis. But as an indication of an effect, we test with a t-test whether the distributions of AUC differences in the first second are different from a distribution with zero mean. As an overall statistical test a repeated measure 2 × 3 × 4 ANOVA with the factors Stereoscopy (2D, 3D), Category (natural, pink noise, white noise), and Time (4 × 5 s time intervals) was applied to the mean AUC distributions of each image feature. Significant factors and interactions were further analyzed by inspecting their AUCs and confidence intervals. We only considered p-values smaller than 0.01 but did not account for multiple comparisons. Interactions between features were not investigated in this study.

### Results

#### Are participants able to recognize 3D structure equally well in all 3D conditions?

We tested whether participants could recognize 3D structure equally well in natural, pink noise, and white noise 3D images. Averaged over all conditions, each subject classified more than 95% of the stimuli correctly. There was a non-significant trend (p = 0.21) to a better mean recognition rate in natural (99 ± 2%, Mean ±SD) and white noise images (99 ± 7%) compared to pink noise images (94 ± 4%). The worst subjects in each 3D condition still classified 92% of the 3D natural images, 75% of the 3D pink noise images, and 85% of the white noise images correctly. The best subjects of each condition classified all 3D stimuli correct. Recognition times differed across categories (p = 0.05), but the pairwise post-hoc tests were not significant (p = 0.10). Thus, we found a trend for decreasing mean recognition times from pink noise (1.58 ± 1.39 s, Mean ±SD) to white noise (1.32 ± 0.83 s) to natural images (1.23 ± 0.84 s). Looking at the individual subjects’ performance in detail, the worst subjects in each 3D condition needed 3.12 s in 3D natural, 4.90 s in 3D pink noise, and 3.55 s in 3D white noise stimuli. The best subjects showed recognition times of 0.72 s in 3D natural, 0.68 s in 3D pink noise, and 0.66 s in 3D white noise stimuli. Concluding, humans were able to recognize natural binocular disparity information, either in isolation (3DW) or in their original context (3DN) with comparable performance. For pink noise images, the accuracy and speed of recognition were slightly reduced.

#### Are eye movement properties different when viewing 3D images compared to 2D images?

Most eye movement research is carried out using natural images lacking 3D information. Therefore, it is of interest to know whether differences in basic eye movement parameters exist between 2D and 3D images. Furthermore, eye movements could depend on the higher order image correlations in natural images. Thus, the eye movement parameters on natural images were compared to pink noise images that only contain second order information and to white noise images that do not contain any correlations. We define eye movement properties as the combination of the following parameters: rate of fixation, saccade length, saccade dynamics, fixation duration, and spatial distribution of fixations. The median eye movement parameters were compared across stimulus conditions (2D and 3D) and categories (N, W, P).
Rate of fixation

It was previously reported that participants make one to three fixations per second (Yarbus, 1967). However, how the rate of fixation depends on the image structure and content remains largely unknown. When comparing the rate of fixation in 2D and 3D images, we found for each category an increase with the introduction of disparity (N: p = 0.007, P, W: p < 0.001). The increase weakened from white noise (+0.55 fixations/s) to pink noise (+0.35) to natural images (+0.15). Comparing stimulus categories, the fixation rate made on 2D images decreased from natural (2.6 fixations/s) to pink (2.1) to white noise images (1.95). In 3D images, pink and white noise images were comparable (2.45 vs. 2.5). Irrespective of the presence of depth, the decrease from natural to pink noise (2D/3D: p < 0.001) and from natural to white noise (2D/3D: p < 0.001) images was significant. Considering only the data of the first 5 s, the results of the 2D/3D comparison and the category comparison were unchanged except a significant trend to more fixations in pink noise compared to white noise images (2D,3D: p < 0.001). These results show that the rate of fixation increased with the introduction of disparity and with the introduction of higher order scene information. The depth-induced increase was strongest in noise stimuli but was also present in natural images.

Saccade length

We measured saccade length as the distance between the locations of two successive fixations in degrees. The overall length was defined as the Euclidean distance, with the horizontal component as the horizontal distance and the vertical component as the vertical distance between the fixated locations.

The introduction of disparity information shortened the saccade length (Figure 2). The effect was significant for natural (p < 0.001) and white noise (p = 0.01) images where we found a decrease of 0.52° and 0.34°, respectively. On pink noise images, the decrease was on average larger than in white noise images (0.39°) but was not significant (p = 0.02). These reductions were mainly caused by a decrease in horizontal saccade length that was present in all categories (mean decrease: 0.48°, N: p < 0.001, P: p < 0.001, W: p = 0.007). The small changes in vertical saccade length were not significant (N/P: p = 0.06, W: p = 0.15). Comparing across categories, saccades performed during the white noise condition were shorter than saccades on pink noise (2D: p < 0.001, 3D: p < 0.001) and on natural images (2D: p < 0.001, 3D: p < 0.001). This difference arose due to a reduction in both vertical (mean decrease: −0.31°) and horizontal distances (mean decrease: −0.34°). Saccade lengths on natural and pink noise images were comparable (2D: p = 1.00, 3D: p = 0.84). Analyzing the time course of saccade length, we found a slight tendency to an overall decrease in saccade length with an average of −0.93° in the first 15 s, which arose in all conditions except 3D white noise and was mainly characterized by a strong decrease in the first 5 seconds (−0.63°). This decrease resulted from a shortening of both horizontal and vertical components. In conclusion, the additional stereoscopic information led participants to make shorter saccades. These changes in saccade length were mainly caused by a decrease in the horizontal direction. Regardless of the presence or absence of disparity, saccades were shorter on white noise images than on natural and pink noise images. The comparable saccade length in natural and pink noise images showed that the mere introduction of second-order correlations was sufficient to reach a length equivalent to that found for natural images. Irrespective of the stimulus condition, the saccade length decreased steadily over time, including a strong decrease in the first 5 s, for all conditions except 3D white noise.

Saccade dynamics

Studies have shown that saccade dynamics are affected by age and state of alertness as well as by task variables (Epelboim et al., 1997). Here we make a large effort toward natural viewing conditions with qualitative effects on perception, hence it can be hypothesized that these dynamics may be modified in the present paradigm. Therefore, we estimate the relation between saccade length and mean velocity for each condition separately and compare the estimated parameters $v_m$ (asymptotic upper limit of mean velocity) and $l_{s0}$ (saccade length at which the mean velocity reaches half its maximum value) across conditions.

As can be seen in Figure 3 for natural images, there is a strong dependence of mean velocity on saccade length. The filled circles show mean velocity for a specific saccade length, averaged across all trials within one condition. The curves shown were fitted to the data using
Equation 1 (Becker, 1991) and the parameters $v_m$ and $l_{50}$ were estimated. In all conditions (data not shown), the fit overestimates velocity for medium saccades and underestimates the parameter for long saccades, nevertheless it explained 83% of the variance of the data.

Comparing the estimated parameters $v_m$ and $l_{50}$ in 2D and 3D conditions, in all image categories we found a slight trend for smaller $l_{50}$ in 3D conditions, which was however not significant. In natural images, mean velocity reached half its maximum value at a saccade length of $7.3^\circ$ in 2D and $6.89^\circ$ in 3D images. The asymptotic upper limit of mean velocity was comparable in 2D and 3D images ($2D/3D: 0.3^{/\text{ms}}$). In pink noise images, half-maximum velocity was reached at a saccade length of $7.89^\circ$ in 2D images and $7.27^\circ$ in 3D images. The asymptotic upper limit was estimated at $0.31^{/\text{ms}}$ in 2D and $0.30^{/\text{ms}}$ in 3D white noise images.

Next we evaluated the temporal dependency of $v_m$ and $l_{50}$ and found a trend for a decrease of the estimates over the first 15 s by an average of $-0.06^{/\text{ms}}$ and $-2.47^\circ$ in 2D conditions and $-0.06^{/\text{ms}}$ and $-2.47^\circ$ in 3D conditions $v_m$ and $l_{50}$, respectively. This trend was present in each condition, but only significant for natural images ($v_m/l_{50}; p = 0.002$).

In summary, the changes in $l_{50}$ indicate that saccades at intermediate length are slightly faster in 3D conditions than in 2D conditions. The maximal asymptotic velocity is unchanged. Thus, certain factors of saccade dynamics are affected by the presence of depth information. No indication was given for a dependence on the category of the image. Further, they are affected by the time interval of stimulus presentation.

Figure 3. Mean velocity as a function of saccade length for 2D (left) and 3D (right) natural images. Filled circles show the means taken at 1 degree saccade length intervals. Error bars show $\pm 1$ SE. Solid lines show curves fitted to the data.

**Fixation duration**

The presence of disparity information produces an additional task for the coordination of eye movements. For example, the correspondence problem between different positions on each retina must be solved in order to perceive a coherent depth structure. Furthermore, the perception of depth necessitates vergence eye movements in order to foveate the target location correctly with both eyes. These tasks could perhaps be thought to prolong the duration of fixations. However, the analysis of the fixation rate, a measure that is not reciprocal to but highly correlated with the fixation duration, already suggested that the duration may not increase. And indeed, as shown in Figure 4, the introduction of disparity did not prolong but shortened the median fixation duration in noise images ($P: p < 0.001$, $W: p < 0.001$) and had no effect on natural images ($p = 1.00$). The decrease in duration was larger for white noise (50.37 ms) compared to pink noise images (28.54 ms). Using the mode as a summary statistics resulted in the same trend but with less significant results. Comparing categories, for 2D presentations we found significant increases of +60.33 ms in median fixation duration from natural to pink noise images (2D: $p < 0.001$) and of +31.42 ms from pink noise to white noise images (2D: $p < 0.001$). In the case of 3D presentations, the difference between the noise conditions vanished ($p = 0.21$), but the difference between natural and noise conditions remained ($P/W: p < 0.001$). Computing the mode made the fixation durations in pink noise images and natural images comparable, irrespective of the presence of depth. But there was still a trend to longer fixations in pink noise compared to natural images. Looking at the time course of the fixation duration over 20 s, there was in all conditions a permanent small increase over the 20-s presentation time. Comparing the first and fifteenth

Figure 4. Boxplot of the distribution of median fixation duration as a function of stimulus type.
second, the fixation duration increased on average about 25.42 ms.

Concluding, the additional tasks during 3D presentations did not prolong the duration of fixations. Further, fixations on noise images lasted longer than those on natural images. Irrespective of the condition, fixation duration increased over time.

**Image exploration**

We compared the inter-quartile ranges of the fixation distributions on natural, pink, and white noise images to investigate whether the relative reduction in stimulus content along these categories leads to a reduction in the degree of exploration of participants’ visual behavior. Further, we analyzed whether the additional depth information in the 3D conditions influenced the extent of exploration.

In 2D, dispersion of fixations was widest in natural images (horizontal: 9.49°, vertical: 5.28°) followed by comparable dispersion in pink (horizontal: 7.33°, vertical: 4.69°) and white noise images (horizontal: 6.16°, vertical: 4.44°). The difference between natural and pink noise (horizontal: $p < 0.001$, vertical: $p = 0.007$) and natural and white noise images (horizontal: $p = 0.007$, vertical: $p < 0.001$) was significant. Thus, the spatial extent of visual exploration decreased when higher order correlations were no longer present in the image. Additionally removing the second-order correlations does not significantly reduce the extent of explorative behavior.

Adding natural disparity information to the stimuli, the dispersion of fixations increased. But this trend was only significant in the case of pink (horizontal: $p < 0.001$, vertical: $p = 0.007$) and white noise images (horizontal: $p < 0.001$, vertical: $p = 0.007$). This increased dispersion mainly resulted from a larger horizontal dispersion in noise images (P: 1.70°, W: 2.52°) and larger dispersion in both directions in natural images (horizontal: +0.52°, vertical: +0.30°). As the increase was stronger for noise than for natural images, the difference in the spatial extent of exploration between natural and noise images disappeared. Thus, all 3D categories had comparable fixation distribution width. In conclusion, these results show that additional information, due to additional higher order correlations as well as due to additional binocular disparity information, increased the spatial extent of exploration of the participants. Further, the presence of depth information made subjects’ fixation dispersion comparable across categories.

Overall, we conclude that basic eye movement properties were different when participants studied images with binocular information compared to 2D images. The effect was largest on noise images but also present in natural images. Participants made more but shorter fixations on 3D images. The saccades were shorter and faster. A wider region of the image was explored. Comparing categories, the additional information in natural images compared to noise images led to more but shorter fixations and stronger visual exploration. The dynamics of saccades were unaffected.

**Does the saliency of 2D image features change with the introduction of disparity?**

The saliency of image features during overt attention under natural conditions has been intensely investigated in previous experiments (Einhäuser & König, 2003; Parkhurst et al., 2002; Tatler et al., 2005). As those studies employed only 2D images, it is not clear how to generalize these results to more natural stimulation conditions containing binocular depth cues. Furthermore, the temporal dynamics of saliency have not yet been investigated for periods longer than 5 s. Thus, we here investigated the influence of the monocular image features mean luminance, luminance contrast, and texture contrast. Besides an analysis of the entire 20-s presentation time, we also separately analyzed the first 5 s to allow a comparison to previous studies with shorter presentation durations and show the saliency time course over the first 5 s on a second-by-second basis (Figure 5). Further, the saliency of the features was compared across conditions.

**Mean luminance**

We analyzed whether the distribution of mean luminance at fixated regions differed from the distribution at control regions. Over the whole stimulus presentation, there was no significant effect of mean luminance for any stimulus type. Analyzing only the first 5 s, the results were
comparable in all conditions except 3D natural stimuli. There, actual mean luminance was slightly higher than control mean luminance (sign test: \( p = 0.01 \), AUC = 0.5038). In a second-by-second analysis of this first 5-s time interval, no clear time trend over the 5 s could be found. Further, for each condition no difference was present between the overall saliency of mean luminance and the saliency in the first second. To investigate saliency differences across conditions and across the entire presentation time, a repeated measure \( 2 \times 3 \times 4 \) ANOVA with the factors Stereoscopy, Category, and Time was applied. We observed no effect or interaction. In conclusion, mean luminance was, except in 3D natural images, uncorrelated with the selection of fixation points. The effect on 3D natural images was small and only present in the first 5 s of the stimulus presentation.

**Luminance contrast**

Many studies using 2D images have shown that luminance contrast is elevated at fixated locations compared to non-fixated locations. To generalize these results to 3D presentations and longer presentation times, we investigate the overall relation between the selection of fixation locations and luminance contrast and the time course of this relationship. For 2D natural images, we replicated the effect of higher luminance contrast at fixated locations and generalized this result to a presentation time of 20 s \(( p < 0.001)\) and to 3D presentation \(( p < 0.001)\). The effect size was small (AUC—2D: 0.53, 3D: 0.53, Figure 6). In 2D and 3D noise conditions, luminance contrast had no effect on attention. The results were identical when analyzing the data of the first 5 s separately. Analyzing this time interval in detail, no time-related saliency effect was found. Figure 5 shows for 2D and 3D natural images the mean difference between the effect at a certain second and the overall effect of luminance contrast. For 2D and 3D natural images, the effect is higher in the first second compared to the overall effect (2DN: 0.006 ± 0.001 (mean ±SEM), 3DN: 0.007 ± 0.001), but this very small trend is not significant. The three-way ANOVA supported the previous results, showing that Stereoscopy and Time had no effect on the saliency. The significant effect of category \(( p = 0.001)\) is explained by the absence of a luminance contrast effect in noise images. Thus, we found a significant correlation between luminance contrast and fixation selection that was constantly present in 2D and 3D natural images over a time period of 20 s. This effect was not present in noise images.

**Texture contrast**

Previous studies have shown that texture contrast has an effect on attention. Therefore, texture contrast maps were computed and the distributions of texture contrast at fixated locations were compared to the distributions at control locations. In the majority of natural image trials, actual texture contrast was higher than control texture contrast (2D: \( p < 0.001, 3D: p < 0.001 \)). As in the case of luminance contrast, the effect size in natural images was small (AUC—2D: 0.54, 3D: 0.53, Figure 5) and no effect was found in noise images. When analyzing only the first 5 s, the results were comparable. But as shown in Figure 5, for natural images we found a very small decrease in the correlation between fixation point selection and texture contrast over the first 5 s. No trend was found for noise images. Despite this small trend, the distribution of the differences between the effect size at second one and the overall effect size is in no condition different from zero. In the three-way ANOVA, only the factor Category was significant \(( p = 0.002)\). This may be due to the difference between natural and noise images. Concluding, the results showed that texture contrast was a salient feature in natural but not in noise images. The effect in natural images was constantly present during the 20-s presentation, with a very slight trend to a decrease in the first seconds. The size of this effect was small but comparable to the size of the luminance contrast effect (Figure 6).

Summarizing the monocular feature analysis, we showed that independent of stimulus type mean luminance did not have any influence on the allocation of attention. In contrast, luminance contrast and texture contrast had a small but significant effect in 2D and 3D natural images, which was not enhanced in the first seconds and almost constant over time. The size of these effects was independent of the presence of depth information. In 2D...
and 3D noise images, no saliency effect was found. Therefore, although basic eye movement properties changed with the introduction of disparity, the saliency of the monocular visual features was independent of the presence or absence of depth information.

**Are binocular image features salient?**

The saliency of monocular image features has been the subject of much research. The effect of binocular features on overt attention, however, has largely been ignored. Thus, we analyzed the effect of mean disparity, a measure for distance, and disparity contrast, a measure for depth discontinuity, on the selection of fixation locations in 3D images. As studies have shown that depth information can be inferred from other image cues (O'Shea, Govan, & Sekuler, 1997), monocular depth cues could have an effect on 2D images. To control for such an effect, we analyzed subjects’ 2D fixation data as if they were done on the 3D images, so that we effectively measured the saliency of the binocular features that were not present in the 2D image. Thus, we investigated whether there is a correlation between the selection of fixation points in 2D images and the disparity information that can be deduced from monocular depth cues (deduced disparity).

**Mean binocular disparity**

Looking at the data from the entire trial duration, deduced mean disparity did not have any effect in the 2D conditions. But in the 3D white noise condition, actual mean disparity was lower than control mean disparity in 67% of the trials ($p < 0.001$). This trend was not present in natural and pink noise images. The effect size of 0.45 in white noise images is comparable to the effect sizes of luminance and texture contrast in natural images.

These results changed when using only the data of the first 5 s (Figure 7). In this time interval, mean disparity not only had an effect in the 3D white noise condition ($p < 0.001$), but also in 3D natural ($p < 0.001$) and 3D pink noise images ($p < 0.001$). The effect was largest in white noise images (AUC = 0.38) followed by comparable effect sizes in natural and pink noise images (AUC = 0.45). Note that AUC values below 0.50 indicate that disparity at fixated locations is systematically smaller than at control locations. In 2D conditions, no significant effect was found. However, there was a slight trend toward a negative mean disparity effect in natural images. Overall, in 3D images participants looked significantly more often at closer locations during the first 5 s and a similar trend was observed in 2D natural condition.

A three-way ANOVA showed an interaction between Stereoscopy and Category ($p < 0.001$) and Stereoscopy and Time ($p < 0.001$). Further, the factor Time was significant.
Therefore, the introduction of depth did not have the same influence on each category and the time effect depends on the presence or absence of depth.

Looking at the time course of the mean disparity effect across the presentation duration, we showed that in all 3D conditions, participants looked first at closer locations before they gradually shifted their attention to locations further away (Figure 7). The same bias was present in 2D natural images, although no disparity information was given. A more fine-grained analysis of the first 5 s revealed that in all 3D conditions, the mean disparity effect decreased steadily over time (for natural images see Figure 5). The time trend for 2D natural images resulted from a more long-term effect and therefore cannot directly be seen in the first 5 s.

Testing whether the effect in the first second is larger than the overall effect, we found that the mean disparity effect is in all 3D conditions larger in the first second than on average ($p < 0.001$). The large effect in the first second could indicate a general behavioral orienting bias to look at a close depth plane followed by a small trend to look at nearby locations instead of a feature effect per se. As mean disparity values were not equally distributed across the image but were smaller at the bottom than at the top of the image, another explanation of the effect would be the existence of a bias to look from the central fixation cross first at the bottom of an image and then gradually at higher levels. For 3D natural images, the initial mean disparity effect could indeed arise from a trend to look downward, as in the first 5 s 52.53% of the fixations are below the center of the image (binomial test: $p = 0.001$). But the disparity effect is still present in the second time interval and there we even find a trend to fixate locations above the center (downward: 47.22%).

Further, such a downward trend could not be found for 3D noise images, where the disparity effect is largest. In these images, in the first 5 s subjects looked significantly more upward than downward ($p < 0.001$, 3DP: 40.45%, 3DW: 38.71%). In 2D natural images, the long-term effect in the first 10 s can also not be explained by a trend to look downward, as there was a slight inverse trend to look upward ($p = 0.001$, 45.19%). Thus, the effects of mean disparity cannot be fully explained by a general bias to look first at the bottom of the image. Whether the effect arises due to a behavioral orienting response to closer depth planes or is an effect of the saliency of disparity per se cannot be decided.

These results demonstrate that the presence of mean disparity changed the allocation of attention at the beginning of stimulus presentation. In all categories, participants first looked at closer locations, and then gradually shifted their attention to more distant locations. As the same trend was present in 2D natural images, this influence on attention can be completely explained neither by the presence of disparity (but it may arise partly due to other features in the image), nor by a general image-independent strategy to make more fixations below the center of an image.

### Disparity contrast

Disparity contrast is a measure of depth discontinuity. Thus, independent of the actual distance of a location it measures the discontinuity between the distance of a given location and its neighboring locations. It is possible that depth discontinuity has an effect on the allocation of attention, but little is known about this feature.

Analyzing the entire presentation time, deduced disparity contrast had no effect in any 2D condition. For the 3D conditions, we only found an effect of disparity contrast in noise images ($P: p = 0.003$, $W: p < 0.001$), but the effect size was small (AUC—$P: 0.52$, $W: 0.53$). The results were comparable when only analyzing the first 5 s, and a detailed analysis of these seconds did not reveal any time-related changes. In accordance with these results, no factor or interaction of the ANOVA was significant.

Concluding, in 3D noise images participants consistently looked more at depth discontinuities than at planar surfaces. Further, in all categories, there was a trend in 2D presentations to look more often at planar surfaces than in the 3D presentations (Figure 8).

Summarizing the binocular feature analysis, we showed that the introduction of depth information resulted in the new salient feature mean disparity. Regardless of the category, participants first looked at closer locations before gradually shifting their gaze to locations further away. This trend was also present, but alleviated, while presenting 2D natural images. Further, the participants...
attended in 2D images more to depth planes than in 3D images. This effect was larger in noise images than in natural images.

**Discussion**

In this study, we used 2D and 3D natural and noise images to investigate the effect of additional binocular information on basic eye movement parameters as well as on the correlations between image features and the allocation of attention. Importantly, ground-truth depth information was obtained for all stereo image pairs by additional laser scanning. We found a change in basic eye movement properties when disparity information was present. The additional information led to an increased number of fixations, shorter and faster saccades, and increased spatial extent of exploration. These effects were largest in noise images but also present in natural images. Despite these changes in basic parameters, the saliency of mean luminance, luminance contrast, and texture contrast was comparable across 2D and 3D stimuli. In natural images, luminance contrast and texture contrast were constantly elevated at fixated locations over a presentation time of 20 s. The size of the effect is, in accordance with previous studies, very small (Tatler et al., 2005). In noise images, the effect was absent. Mean disparity had a time-dependent effect in 3D stimuli. In all categories, participants looked first at closer objects and then shifted their gaze to locations further away. This effect was found even in 2D natural images, in which disparity is not present but can be deduced from other image cues. Disparity contrast was only elevated at fixated regions in 3D noise images but not in 3D natural images. Concluding, we showed that the results of studies using 2D stimulus presentations cannot automatically be generalized to 3D vision as even basic eye movement properties change with the introduction of disparity. Further, the additional disparity information can be used as an image feature that correlates with the allocation of attention to the same extent as previously investigated monocular features like luminance and texture contrast.

In the current experiment, the stimuli were displayed on a 3D display that provides stable 3D perception without using any additional tools like shutter or red–green glasses. As this depth perception necessitates vergence movements, it solves the convergence problem. However, vergence movements are coupled with accommodative movements (Wann, Rushton, & Mon-Williams, 1995) and this automatic coupling must be overcome to achieve a 3D percept. There is some evidence for a possible distortion of the 3D impression by such a decoupling (Watt, Akeley, Ernst, & Banks, 2005), but stereo systems that allow natural accommodation are still in the development stage (for an example system, see McQuaide, Seibel, Kelly, Schowengerdt, & Furness, 2003). Thus, the only alternative would be to track eye movements in a real-world environment (Einhauser et al., 2007; Hayhoe & Ballard, 2005; Rothkopf, Ballard, & Hayhoe, 2007; Schumann et al., 2008). While this ingenious setup is indeed more natural and yields important results, it is not controllable in the sense we needed for our present investigation. Here, we need a direct comparison of the control of attention in identical 2D and 3D scenes. As a consequence, we decided to use the 3D display.

In comparison with previous studies, we showed the stimuli for a longer presentation time. This decision could lead to problems, as at least the information of the 2D stimuli could already be captured in the first seconds. But as our main results were also observed when analyzing only the first 5 s, the long presentation time did not alter the results of the first seconds. Furthermore, it allowed a detailed investigation of potential time-related changes in the control of attention.

As we were interested in an influence of disparity that is due to a more general bottom-up process, our stimulus set was restricted to natural forest scenes that did not contain any man-made objects. Presenting man-made manipulable objects within reaching distance would have strengthened action-related top-down components. In a recent study it has been argued that relevant objects are also selected by models that rely on bottom-up mechanisms (Elazary & Itti, 2008). Here, the setup we used allows us to distinguish pure bottom-up components from components that arise due to correlations with higher level scene information by comparing noise images containing no scene content or only disparity information (pink/white noise) with natural images.

We found a correlation between disparity and the selection of fixation points, and a change in basic eye movement properties with the introduction of disparity information. Although the only difference between corresponding 2D and 3D images is the addition of disparity information, we cannot distinguish whether the disparity per se or the stimulus properties that were changed by the additional disparity information cause the effects we found. The addition of disparity information leads to a clear depth percept with, for example, the perception of surfaces due to depth differentiation and the perception of depth relations. The change in these properties between 2D and 3D presentations is largest for noise images, as 2D noise images do not include any cue about the depth relations within the image. This larger change could explain the larger effect in noise images compared to natural images. Thus, our setup allows us to say that the change in disparity information is the cause for any effects found between the 2D and 3D conditions. Whether disparity per se or other changes in image properties that are caused by the additional disparity information cause the effects we found must be evaluated in a follow-up study.

For a long time, it was believed that the dynamics of saccades are insensitive to high-level cognitive variables.
This belief was based on the strong relationship between peak velocity and saccade length, called “main sequence.” It was only known that state of alertness and age can affect the main sequence. But then, it was shown in a series of experiments that saccade dynamics also depend on task variables (Epelboim, 1998; Epelboim et al., 1995, 1997). One finding was that the main sequence changed when people had to look and tap a sequence of targets compared to only looking at the targets (Epelboim et al., 1997). In our study, we showed that even basic eye movement properties are affected by the type of stimulus. Further, there were some indications that saccade dynamics are different when looking at 3D compared to 2D images. Whether this difference arose because of the additional vergence movements or because of the more action-related perception of 3D images must be further evaluated in following studies.

In all conditions, the fixation rate increased with the introduction of new information such as disparity or higher order scene structures. The result suggests that more comprehensive images are explored faster. Supporting this suggestion, the duration of fixations is decreased in images containing more information. This finding supports the results of Unema, Pannasch, Joos, and Velichkovsky (2005) who reported that fixations are shorter in images with more objects. The overall faster exploration led to an increased explorative behavior of the subjects. This increase was largest between 2D and 3D noise images and could be explained by the high amount of additional information contained in the created surface regions. Saccade velocity increased with the introduction of 3D depth cues. In summary, we hypothesize that basic eye movement properties and explorative behavior depend on the amount of information in an image. This hypothesis should be further evaluated by comparing these features across stimuli containing a gradually increasing amount of information.

Concerning the time course of the basic eye movement properties, our results suggest that the exploration of an image is most pronounced in the first seconds of stimulus presentation. As in previous studies (Buswell, 1935; Unema et al., 2005), we found a continuous decrease of the fixation duration over time. Further, the saccade velocity and saccade length steadily decreased with a pronounced decrease in the first 5 s. Unema et al. (2005) also found a strong decrease of saccade length in the beginning of stimulus presentation, but only up to 2.5 s. This discrepancy could arise as in our study subjects first had to adapt to the 2D or 3D presentation, which on average took 1 to 2 s. The overall decreases occurred irrespective of the stimulus condition. In all conditions, the dynamics of saccades were also changed over time. In the beginning of stimulus presentation the dynamics was faster and the upper limit of mean velocity was higher. Thus, we suggest that in all stimulus presentations the speed of exploration decreases over time.

The additional disparity information in the 3D images necessitates vergence movements of the eye and therefore also binocular matching and fusion maintenance. This matching task did not prolong the fixation duration but reduced saccade length in 3D conditions. To further evaluate this suggestion, vergence movements should, as a next step, be directly analyzed on the data set. Such an analysis could also reveal whether the results from our natural image setup can be generalized to real-world behavior. For artificial stimuli, Steinman (2003) reported that the results often cannot be generalized.

In models of overt attention based on image saliency, saliency is a determining factor for the selection of fixation positions and thus consequently it plays a major causal role on oculomotor control. Therefore, one might expect that changes in the oculomotor strategies that may occur under different conditions should in principle be associated to changes in the properties of image saliency. Interestingly, we showed in our study that while basic eye movement properties and saccade dynamics were affected by the introduction of disparity information, there was no evidence for a change in the saliency of 2D image features. As a consequence, the changes we observed in the oculomotor characteristics cannot be accounted for by the saliency of 2D image features suggesting that ocular motor properties are at least not primarily dependent on saliency.

Concerning the image feature analysis, we did not find an effect of luminance and texture contrast saliency in noise images. This suggests that the effect in natural images could be caused by a higher order correlation with features that are present in natural images but absent in noise images. In accordance with our results, Einhäuser et al. (2006) showed that the saliency of contrast in natural images decreased with increased phase noise. From this finding, it could be hypothesized that the effect of the image features is not causal but only correlative. To test such a hypothesis, the effect of image features on unmodified natural images should be directly compared with the effect on the same but contrast-modified versions of the same images, as previously done for 2D images by Einhäuser and König (2003). Thus, although our data indicate that the correlations are not causal, we cannot definitely conclude this from the data. We did not directly address this hypothesis in the present study but were instead interested in whether the saliency effect can be found at all in 3D scenes.

There is an ongoing debate regarding the time course of the saliency of image features. The theory of bottom-up influences of image features suggests that the influence of these features is largest at the beginning of the stimulus presentation. And, indeed, Parkhurst et al. (2002) found such a time course. But three years later, Tatler et al. (2005) suggested that this finding was caused by a bias in the data and found a constant small correlation between fixation point selection and image features. Carmi and Itti (2006) argued again for a higher effect at the beginning of stimulus presentation, at least in movie clips. Our results support the finding of Tatler et al. (2005), as we found neither an increased saliency effect of luminance and
texture contrast at stimulus onset, nor a time trend over the 20-s presentation time. Thus, we suggest that the saliency effect of these features is constant over time. This result is an additional small indication against the causal role of these image features. Contrary to 2D scenes, 3D scenes contain the additional image features of disparity and disparity contrast. This additional information could change the saliency of features that are present in both 2D and 3D images. Here we showed, however, that the overall saliency and the saliency time course of the tested monocular features is comparable in 2D and 3D scenes. Thus, the bottom-up saliency model should also predict a substantial fraction of the allocation of attention in 3D scenes. We further suggest that in a 3D attention model, disparity could be included as a new salient image feature. In the first 5 s of presentation, the saliency effect of this feature was comparable to the saliency effect of luminance and texture contrast.

Whether the correlation between the selection of fixation locations and disparity is causal can indirectly be analyzed by evaluating the time course of the correlation and the saliency effect on noise images, in which no correlation with higher order structures is present, and on 2D natural images, where only monocular cues are present. Looking at the time course, the saliency of disparity depended on time. Participants first looked at closer locations before they shifted their gaze to locations further away. This finding is consistent with the hypothesis of selection for action. Objects that are closer to us have a higher impact, as on the one hand they could be more immediately dangerous and on the other hand they are within reaching distance and thus can be acted upon. As the saliency of disparity was not only present in 3D natural but also in 3D noise images, the influence could be interpreted as a general bottom-up awareness of locations that are closer. But it cannot be excluded that subjects saw object-like structures in the 3D noise images, as pure disparity information is sufficient for recognizing objects (Julesz, 1960). Supporting the idea of a general difference in the processing of locations in near and far space, previous studies showed dissociation between the neural representations of far and near space (Cowey, Small, & Ellis, 1994; Halligan & Marshall, 1991). Further, an enhanced processing of near space is supported by monkey studies that report an increased number of receptors for detecting stimuli close to the body (Rizzolatti & Camarda, 1987). In addition to the proposed general bottom-up component, the disparity effect may also be due to a top-down object-dependent component, as we found a reduced but similar disparity effect in 2D natural images. In these images, disparity cues were not directly present but they could be inferred from other monocular depth cues. Supporting evidence for such top-down effects on the representation of far and near space comes from neuroscientific studies. The size of receptive fields representing near space depended on reaching distance, as tool use can extend their size (Iriki, Tanaka, & Iwamura, 1996). Further, a neglect study reported that the radius of a near-space lesion effect can be extended by tool use (Berti & Frassinetti, 2000). Concluding, the distinction between near and far space not only depends on general cues like disparity but can also be altered by context. In our study, we found for the first 5 s a bias to fixations in near space that was present in images containing only disparity information (3DW) as well as in images that allowed the viewer to infer a distinction between near and far space despite the absence of disparity (2DN). Thus, the influence of disparity on the selection of fixation locations seems on the one hand to be casual and on the other hand to be caused by deduced depth information.

Given these results, we suggest that the control of attention is mediated by a selection-for-action process (Bekkering & Neggers, 2002). To further investigate this hypothesis, a follow-up study should include images showing objects that can be manipulated. If the saliency of disparity not only results from a general bias of higher attenuation of closer locations, but from a bias to look at objects that have a higher impact on our actions, these objects should be fixated first (Rothkopf et al., 2007). Further, including a condition that allows the viewer to grasp and interact with these objects should further enhance their saliency. Such extensions to our setup should allow the further investigation of a potential top-down component of selection for action.

Current saliency map models disregard the fact that humans operate in a 3D environment, as the models are only based on 2D image features and are only tested on 2D images. Our study has shown that even basic eye movement properties like saccade length or fixation rate changed with the addition of depth perception. Thus, we should be careful in generalizing findings from studies using 2D stimuli to statements about 3D environments. Concerning the saliency of image features, we showed that the saliency of monocular features is comparable in 2D and 3D presentations. The use of the 3D display enabled us to investigate binocular image features like disparity. This feature had a time-dependent saliency effect that in the first seconds is comparable in size to the saliency of luminance and texture contrast. In previous studies and in the saliency model, disparity was mainly disregarded. Extending the saliency model to include binocular features could improve performance of the current saliency models. In conclusion, we therefore promote the study of 3D scenes with the help of 3D displays for two reasons. First, 3D conditions provide a more natural and thus more generalizable setup, and second, depth cues seem to be as salient as well-studied monocular cues.

**Acknowledgments**

We thank Robert Märtin and Johannes Steger for the stimulus acquisition and the computation of the disparity maps, Boris Bernhard for providing the pink noise stimuli, Sonja Schall for recording additional subjects, and
Benjamin Tatler, Cliodhna Quigley, Alper Acik, and an anonymous reviewer for helpful comments on a previous version of this manuscript.

This work is related to the Perception on Purpose STREP project and was supported by EU Grant 027268.

Commercial relationships: none.
Corresponding author: Lina Jansen.
Email: lijansen@uos.de.
Address: Albrechtstr. 28, University of Osnabrück, IKW/NBP, 49069 Osnabrück, Germany.

References


