Philipps Universität Marburg

# Visual attention in the real world

**Dissertation**
**zur**
**Erlangung des Doktorgrades**
**der Naturwissenschaften**
(Dr. rer. nat.)

dem
Fachbereich Psychologie
der Philipps-Universität Marburg
vorgelegt von

**Bernard Marius ’t Hart**

Aus Stadskanaal, Niederlande
28-04-1977

Marburg/Lahn, August 2011

Vom Fachbereich Psychologie der Philipps-Universität Marburg als Dissertation am _________ angenommen.

Erstgutachter: Prof. Dr. Frank Rösler

Zweitgutachter: Prof. Dr. Wolfgang Einhäuser

Tag der mündlichen Prüfung am _________.

# Table of Contents

# Cumulus

## *Introduction*

Even though “everyone knows what attention is”, it has been studied widely since William James wrote these famous words, up to the present day. Here the focus is on *visual* attention: selecting some visual location(s) over others for prioritized processing. Most of the work on this topic has been done with laboratory experiments. The method of choice has often been eye-tracking where the direction of gaze is used as a proxy for visual attention. Though studies of visual attention in laboratory settings have been valuable, the external validity or real-world applicability of this work has not been tested often. With modern, wearable eye-trackers we can now assess visual attention in the real world. This allows validation of laboratory-generated theories and models, as well as measurements made in “setups” that are impossible – or at least very hard – to mimic in the laboratory. Two of the four studies described here are among the first to do this.

First, a feature-based model for the prediction of attention, and one of its mechanistic assumptions, is tested with naturalistic stimuli. Second, a comparison is made between laboratory- and real-world visual attention using virtually identical visual stimulation. Third, the effect of an implicit task on visual attention in the very common real-world activity of walking on a street is measured. Fourth, the effect of making hand movements on visual perception is studied. The first two studies focus more on the properties of the visual input, whereas the second two studies focus on the interaction between action and perception. These four studies cover several topics in real-world visual attention and show both the feasibility and necessity of studying perception and action under naturalistic conditions.

## Overt visual attention

Since the human eye has a fovea containing a local high density of photo receptors, we can inspect only a part of the visual field in high detail at any time. Consequently, the direction of gaze has to be changed if another point is to be inspected in high detail. As in most primates humans usually accomplish this through eye and head movements. Aside from the resolution distribution of the retina, there may be other 'bottlenecks' in the visual hierarchy, which restrict the processing of visual input.

Since the number of locations that can be looked at in a given amount of time is limited, choosing the right locations to inspect is essential for gathering the information necessary to complete any task that depends on visual input. The direction of gaze and the processes underlying its choice may both be called visual attention, or more precisely: overt visual attention. Though covert (cognitive) attention can be separated from the direction of gaze with some effort (Posner, 1980), the two are usually coupled and appear to share a common neural basis (Rizzolatti et al., 1987). In any case, by manipulating the task, its demands or the visual input, overt visual attention can be directed elsewhere. Hence, by measuring the direction of gaze the processes underlying visual attention can be studied.

Studies using the direction of gaze to assess visual attention have been done for decades (Buswell, 1935; Yarbus 1967). Participants in these studies usually have their heads restrained while they watch pictures. Here the study of visual attention is extended to more life-like situations. Several different issues in visual perception are investigated, which cover a wide range of topics. First, in the remainder of this introduction, some background on the different themes is provided and the four studies are briefly described, followed by a set of overarching conclusions and a discussion of some open questions. The next four chapters each deal with one of the studies. Finally, a summary in German, English and Dutch are provided.

## External validity

The purpose of vision – and indeed all sensory perception – is not to optimally represent stimuli, but to allow an organism to behave adequately given the situation it is in (Einhäuser & König, 2010). However, most of the research on visual attention to date has been performed with artificial stimuli in laboratory conditions. Real-world perception on the other hand usually involves multi-modal sensory input, and the information from the different senses may be converging or diverging, and occurs in a dynamically changing context and task-set. All the potential sources of differences between visual attention in traditional laboratory studies and in real life lead to two good reasons to perform real-life experiments.

First, there is the issue of external or ecological validity. Predictions from laboratory experiments (and models) should be validated in real-life situations (Einhäuser & König, 2010). Well-defined laboratory tasks may provide clear and repeatable results, but still be of little value in real-world situations because of low ecological validity. For example, a standard assessment test for cashiers produced performance rankings that were systematically different from rankings of actual productivity (Sackett et al., 1988). Optimum performance did not prove to be a good estimate of typical performance. In general, the desirability of (ecological) validity should be self-evident for any kind of science. Especially when studying something as complex as the human brain, the applicability of results depends on how well real-world situations are captured by experiments. Technological developments in wearable eye-tracking devices (e.g. Schneider et al., 2009; see Figure 3.1) now allow validation of laboratory studies on visual attention in real life.

Second, by carefully and systematically studying real-life behavior, new behavioral observations may be made, leading to new hypotheses to be tested in more readily controllable laboratory experiments. This approach has recently been dubbed 'cognitive ethology' (Kingston et al., 2008). Since behavior measured in laboratory tasks will be stereotypical for the task and context, the only way it may lead to the discovery of new paradigms is by accident or by a long and arduous search through all possible tasks ('task-space'). Observing naturalistic behavior in moderately free tasks can be seen as a heuristic which cuts the search through task-space short and thereby enables a faster development of the field.

These two reasons are complementary. Validation of laboratory results can be seen as moving research out of the laboratory into the world. Cognitive ethology can be seen as bringing real-world observations back to the laboratory. The experiments described here all fall in the category of testing laboratory results or models in more complex, even real-life, situations.

## Early visual processing

Preceding the visual perception of whole objects or scenes there are many brain areas that process visual information. The way visual information is processed at each stage determines what information is available for the next stage of processing. All stages of processing of visual information therefore may shape the visual world we perceive and hence what we may or may not pay attention to. This implies that although the selection of locations in visual space to attend may be cognitive, it still relies on the earlier stages of visual processing.

The distribution of rods and cones on the retina already results in a higher resolution at the fovea (Østerberg, 1935), implying a lower resolution in the periphery. In the retina the visual 'input' is processed and condensed by several types of retinal ganglion cells such as on- and off-center cells and color-opponency cells (e.g. Derrington et al., 1984). These cells' firing rates code for a contrast in luminance or color of one small area of visual space against its surroundings. This 'preprocessing' is continued after the output of the retina has been relayed over the lateral geniculate nucleus (LGN) to the primary visual cortex (V1) by so called "simple cells" and "complex cells", which code for oriented luminance contrasts of specific width (Hubel

& Wiesel, 1962). Similarly, other features of the visual input are coded in other areas. For example, perceived motion is coded in medial temporal area (V5/MT) (Tootell et al., 1995) and in the immediately anterior medial superior temporal area MST, visual flow-fields are coded for (Saito et al., 1986). MT and MST are areas in what is called the "parietal stream" or "dorsal stream" (Mishkin & Ungerleider, 1982; Goodale & Milner, 1992), which consists of many parietal areas up to somatosensory cortex. The dorsal stream supposedly plays a role in the visual guidance or planning of actions. The so-called "temporal stream" or "ventral stream", [leading to/consisting of] inferior temporal cortex, is supposedly involved in object recognition and the formation of long-term memory. A common hypothesis is that as activation spreads from early to late visual areas more and more complex features are extracted, or redundancy is reduced (Barlow, 1961) until the activity of a single cell codes for the presence of complete objects instead of low-level features (e.g.: Booth & Rolls, 1998, for recordings in humans; Quian Quiroga et al., 2005). This does not only require different kinds of visual features (such as shape and color) to be bound into coherent wholes, but needs to be robust against transformations of objects that do not change its identity, such as rotation. Beyond visual processing, representations of objects may encompass information from other sensory modalities as well (e.g. Amedi et al., 2005; Schall et al., 2009). Sensory perception is not a purely feedforward process, though. Activity in macaque V4 and in human V1, V2 and V3 is modulated by spatial attention (Moran & Desimone, 1985; Munneke et al., 2008). In healthy adults, the world that is perceived appears to be seamlessly integrated across modalities, even though the sensory organs relay very simple features to the brain in separate streams.

## Bottom-up models of visual attention

A model of visual attention is Feature Integration Theory (FIT; Treisman & Gelade, 1980). In this model, there are several more or less retinotopic feature maps that encode where there are interesting locations in the visual field, for that feature. Color is an example of a feature dimension. If a location has a different color from its surroundings that location stands out, and there is a high value in the feature map there. All feature maps are then combined into a single master map of locations. The highest peak in this map then receives attention via a winner-take-all mechanism. This model is based on physiological data on the processing of features as well as behavioral data from feature and conjunction search experiments. “Guided Search” (Wolfe et al., 1989) extends the notions of FIT to explain behavior when searching for triple conjunctions and allows top-down modulation of behavior. A later, computational variation on such a model, the Saliency Map model (Itti & Koch, 2000; see also: Koch & Ullman, 1985) was used to predict fixation locations (see Figure 1.1). With proper adjustments predictions can reach levels of more than 80% correct (Betz et al., 2010). Though these predictions are far from perfect, they are consistently above chance, demonstrating that low-level features, such as luminance contrast, color contrast and orientation contrast have elevated levels at the visual locations selected for fixations, even in real life (Schumann et al., 2008).

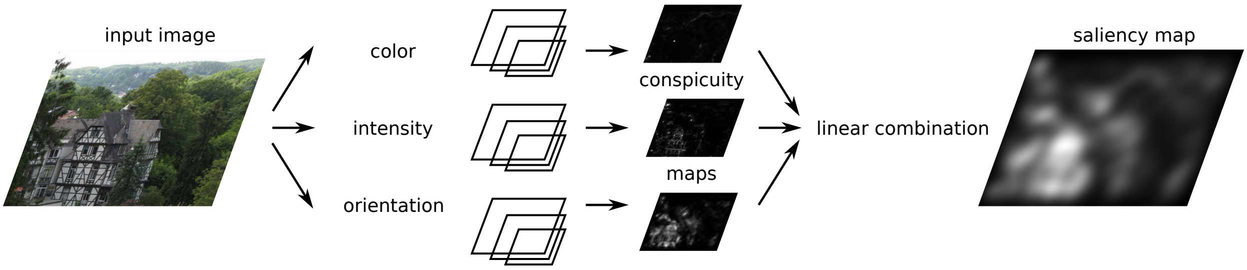

***Figure 1.1. Saliency Map Model with example.***

*For each of the three features color, intensity and orientation the center-surround contrast is calculated at different scales. After competition within each scale, the different scales are combined into a conspicuity map, which are in turn linearly combined to generate a saliency map (the brightness of the pixels codes for saliency). In the example, the model would predict more attention for the house in the foreground.*

The Saliency Map model has been augmented with additional features of various kinds (Einhäuser et al., 2009; Betz et al., 2010). Features that appear particularly good at predicting fixations are based on motion (Carmi & Itti, 2006), and hence require dynamic stimuli. This implies that salient movements induce high inter-observer consistency in attracting gaze. The high predictive value of perceived motion as a low-level feature, already suggests it may be relevant for real-life situations. Large-field motion is indeed used in walking (Callow & Lappe, 2008) and affects hand-trajectory in reaching movements (Saijo et al., 2005). Bodily movements affect the visual input in both tasks. In reaching, the position of the hand is continuously changed and this visual information may be used to correct the reaching trajectory. In walking a large expanding flow-field indicates speed and heading direction and can therefore also be used to verify the efficacy of movements or to adjust motor planning. Both types of tasks show that action and perception continuously interact, as is usually the case in real life.

## Top-down models of visual attention

Actions are usually embedded in a task, and this task may then also affect attention. The relevance of the task in determining attention is not a new notion. As early as 1935, Buswell used an elaborate setup for measuring gaze direction and with tedious data processing in the pre-computer era, found that the kind of task an observer is engaged in affects which parts of a scene are inspected and in which way. In this study an observer could freely view a photograph of a street scene with the Tribune Tower in Chicago or the same observer was asked to find a person standing behind one of the windows of the tower. A widely cited, and perhaps more thorough study by Yarbus (1967) confirmed this finding by asking an observer to answer several questions about Ilja Repin's painting "The Unexpected Visitor". Questions were for example "Give the ages of the people", "estimate how long the visitor had been away from the family" or "remember the positions of people and objects in the room", aside from a free-viewing condition. These different assignments resulted in distinctly different patterns of eye movements (see Figure 1.2). Both Buswell and Yarbus' experiments demonstrated that with a different task, observers inspect different objects in the scene and perhaps in a different way. This work has already shown that cognition, or ‘top-down’ processing plays a large role in determining the direction of gaze.

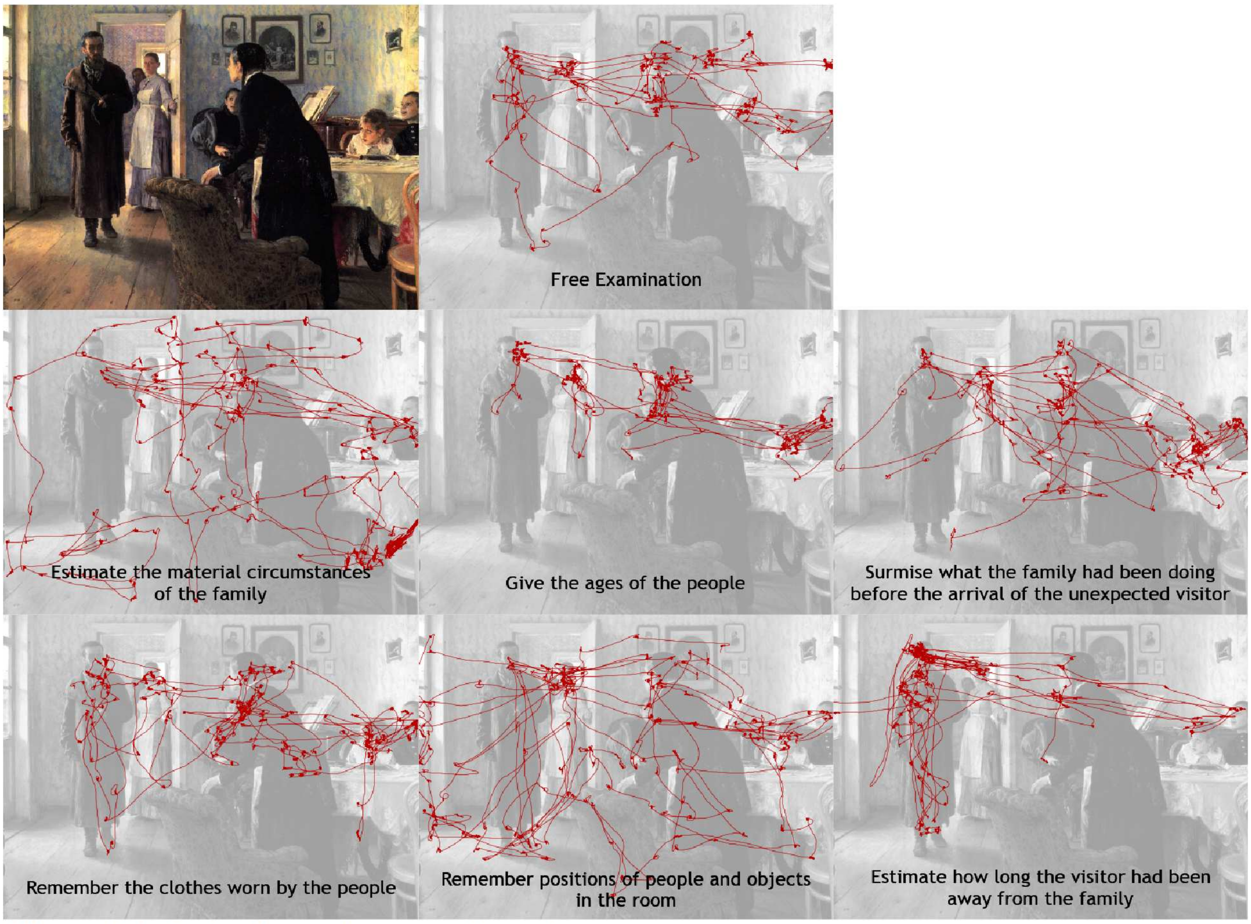
Free Examination
Estimate the material circumstances of the family
Give the ages of the people
Surmise what the family had been doing before the arrival of the unexpected visitor
Remember the clothes worn by the people
Remember positions of people and objects in the room
Estimate how long the visitor had been away from the family

***Figure 1.2. Yarbus' experiment replicated.***

*An observer's eye-movement patterns show the same differences between tasks as in the original experiment.*

Search provides a well-controllable task in terms of the target and its features and is thus widely used. When searching natural scenes, context provides priors to restrict gaze to areas likely to contain the target (Torralba, 2003) and the effect of task can completely override the effects of manipulations of low-level features (Henderson et al. 2007; Einhäuser et al., 2008a). Most task-oriented work in real-life settings has investigated direction of gaze in tasks such as sports (Hayhoe et al., 2005; Land & McLeod, 2000) food preparation (Land et al., 1999; Hayhoe & Ballard, 2005) or driving a car (Land, 1992; Land & Tatler, 2001; Kandil et al., 2009). Taken together these studies show that task is a better predictor of gaze than visual input by itself.

Apart from the problem that it is hard to model task or its effects in a generic way, (but see: Ballard & Hayhoe, 2009), it has also been shown in laboratory experiments (Posner, 1980) that some types of salient events of which participants know they contain no information, may nevertheless not be ignored. Salient, but non-predictive cues (non-predictive in terms of the laboratory task the participants are engaged in) still affect spatial attention. This means that given the right circumstances, top-down effects can surely override bottom-up effects, however the reverse may also occur.

If search targets are defined by their low-level features, the task may recruit bottom-up processing however to speed up performance. The “Guided Search” model tries to captures this (Wolfe et al., 1989; Wolfe, 2007) by adjusting the influence of features depending on task demands (see also: Navalpakkam & Itti, 2005; Peters & Itti, 2007). Regardless of whether these models are veridical or not, the integration of bottom-up and top-down influences on attention may capture real-world visual attention better than either alone.

### *Overview*

Here, the effects of low-level features and the task of walking on visual attention as well as the effect of making movements on perceptual interpretation of stimuli is studied in four separate experiments. These will be briefly discussed below.

## Study I: Color- and luminance-contrast effects add linearly

Laboratory tasks show effects of low-level features on attention (Shiffrin & Schneider, 1977; Schneider & Shiffrin, 1977; Treisman & Gelade, 1980). Typically, if a search target among homogeneous distractors differs on a single feature dimension, such as orientation or color, the number of distractors does not affect the time needed to find it. This phenomenon is called ‘pop-out’ and indicates that visual information in the periphery is processed and can even affect attention quite strongly. On the other hand, if two or more features define the search target, observers engage in serial search, indicated by search time increasing with the number of distractors. This shows that there is some kind of ‘bottleneck’ for processing visual stimuli, and that this has to be after the processing of low-level features. Several models intended to explain these behavioral results using physiologically plausible mechanisms have been proposed (Treisman & Gelade, 1980; Wolfe et al., 1989).

One classical model, the so-called “Saliency Map” (Koch & Ullman, 1985; Itti & Koch, 2000) that aims to explain sequential shifts of attention, assumes a linear addition (or an equivalent weighed averaging) of the effects of features on attention. More recently, however an alternative has been suggested that the maximum activity across several feature-based, retinotopically organized maps determines what location in visual space attracts most attention (Li, 2002), even in natural scenes (Lewis & Zhaoping, 2005). Both may be implemented in the brain. However, physiological evidence on functional connectivity on the level of single cells is hard to obtain.

A correlation between low-level features and visual attention has already been demonstrated in a real-world setting using a free-exploration task (Schuman et al., 2008). The theoretical merit of additive models has been discussed (Vincent et al., 2007) and the predictions of an optimal Bayesian and maximum model of human behavior have been compared (Vincent et al., 2009). A question that has remained open is whether maximum or additive models predict human behavior better.

In Study I of this thesis, eye-movements are measured while observers watch photographs of scenes where the contrast of two low-level features are manipulated independently along horizontal gradients. That is, color contrast and luminance contrast could increase to the right or to the left of the scene, or could stay at the original level. This means there was a 3 x 3 design, with one neutral condition with no changes made to the image, 4 conditions with single feature manipulations (color-contrast increasing to the right or to the left, as well luminance-contrast increasing to the right or to the left) and 4 conditions with combined feature manipulations (color-contrast and luminance contrast both increasing to the left or right, or increasing in opposite directions). Observers' eye-position was tracked in a free-viewing task. The horizontal eye-position in scenes with only a single feature manipulated was used to predict behavior in scenes with two gradients changed using an additive and maximum model. The predictive performance of the additive and maximum model were then compared.

In all stimuli, a viewing strategy could be observed: observers first look to the left of the image and then to the right. On top of these generic, task- or scene-induced effects, the individual feature-contrast gradients exerted a bottom-up influence as both models would predict. Observers directed gaze more to the side of the image with increased feature contrast, and this effect was stronger for luminance contrast than for color contrast.

The behavior in the four single feature conditions was used to predict the behavior in the four conditions with combined feature manipulations. In the additive model the prediction consists of the added effects found in the two single feature conditions that were combined in the double feature condition. Similarly, the maximum model used only the feature eliciting the strongest response across observers and images. Both models were used to predict the horizontal eye-position in the first five fixations. The additive model differed significantly only from the first fixations in two conditions, but did not differ significantly from the other 18 conditions (indicating good performance of the model). Using luminance contrast (which elicited the strongest effect in single feature conditions) to predict behavior in combined feature conditions results in a prediction of 4 of 20 fixations that do not differ significantly from the measured data, and using only color contrast results in the prediction of 8 of 20 fixations that does not differ significantly from the measured data. If horizontal eye-position is averaged over the first five fixations, the additive model does not differ from any of the 4 averages, while each single feature does.

The key result of Study I, better prediction of human behavior by additivity of features then by a maximum model, suggests that attention – insofar as it is based on visual input – may also employ additivity of features. Regardless of one's stance on whether attention is ultimately feature-based or task-based, models predicting attention or gaze under free-viewing conditions are constrained by these findings.

## Study II: Free exploration versus "free" viewing

Direct comparisons of overt visual attention in laboratory settings and real-world settings are scarce even though the assumption that laboratory conditions are a good model for natural vision has gone largely untested. Some first indication that the laboratory setup itself, with restrained head and stimulus presentation on a screen, can bias visual attention has been found previously (Tatler, 2007).

In a first attempt to directly compare laboratory- with real-life conditions, previous recordings of subjects freely exploring various environments are used in two laboratory tasks. These recordings consist of eye-in-head tracking data, as well as a first-person perspective movie. They provide a real-world 'free exploration' condition. In one laboratory condition 'continuous replay', the movies recorded with the head-centered camera are shown to observers. In another laboratory condition ('1s frame replay'), slide shows of equal duration to the movies are shown to the observers. The slide shows for the 1s frame replay condition are created by taking the first frame from each second of each movie and shuffling these frames so that each new slide show has an equal number of frames from each original movie in a random order. In both laboratory conditions, eye-in-head position is recorded, allowing a direct comparison with real-world eye-tracking data.

There are several differences between a laboratory and real setting that may cause differences in perception or behavior. First, participants remain immobile in the lab, since head movements are prevented with a forehead- and chin rest. Second, aside from visual stimuli, no other sensory input is given, though normally auditory, olfactory and perhaps tactile or other information would be integrated with the visual modality. Third, no interaction with the environment is possible in the laboratory. Some interaction with a visual environment may be provided in virtual reality laboratories, though this will usually affect the 'realism' of the visual input and the mode of interaction will not be the same as in real life (though see Ballard & Hayhoe, 2009). These differences with real life all apply to the continuous replay condition and may affect visual attention and hence the direction of gaze and eye-movements.

The 1s frame replay condition is a further step away from real-world visual input, even though showing static images for a short duration has been a common method to investigate 'real-life' visual attention. There are two further differences with the video replay condition. First, there is no motion left in the scenes, and second, all temporal context is removed. Even though one moves about through the world, the changes this induces in the visual input are small in comparison to the changes induced by the sudden onset of a new image. All these differences between real-life and laboratory situations may affect visual attention. By measuring eye-movements with equal visual input, the extent of the effect of these differences can be assessed.

The distribution of eye-in-head position over all observers has a different shape for each condition. Most notably, in the 1s frame replay condition, eye-positions are centered much more. This may have two causes. First, in the free exploration condition and the continuous replay condition, observers are confronted with a mostly expanding flow field. When observers foveate objects or other locations in such stimuli, tracking movements away from the center have to be made. These types of eye-movements are necessarily absent in the 1s frame replay condition. A second difference can be observed in the data; the onset of a new frame in the 1s frame replay condition triggers eye-movements back towards the center of the distribution. These movements could indicate that observers in this condition need to reorient themselves to the suddenly appearing scene. Such behavior would be absent in real life where sudden scene changes are rare. This already indicates that using suddenly presented, static images induces behavior that is qualitatively different from its natural counterpart.

A measure for similarity of behavior evoked by the different conditions, is to what degree the observers direct gaze at similar locations, or how ‘consistent’ the direction of gaze is between two observers. Consistency between two observers is defined here as 1 – d/m, where d is the euclidean distance between the two points the observers’ gaze is directed at and m is the length of the diagonal of the movie frames. Pairs of observers in the same laboratory condition have a higher inter-observer consistency than pairs of observers from different laboratory conditions. However, the inter-observer consistency within continuous replay is lower than within 1s frame replay. Additionally, when all participants in the free exploration condition are treated as one observer, the direction of gaze in free exploration is more consistent with the direction of gaze in continuous replay than the direction of gaze in 1s frame replay. This is a second indication that continuous replay may be a better model of real-world visual input for use in laboratory experiments.

The higher inter-observer consistency in 1s frame replay may be explained by the larger central bias found in this condition. This is confirmed by an analysis of Kullback-Leibler divergence within conditions, showing that the distributions of gaze are more similar in continuous replay compared to 1s frame replay, even when accounting for the 1 second periodicity in 1s frame replay. However, since Kullback-Leibler divergence cannot evaluate single gaze directions, measures based on euclidean distance remain best suited for estimating instantaneous inter-observer consistency, which is artificially high in 1s frame replay.

Artificially high inter-observer consistency may be problematic for real-world validity of results. This also raises the question of what the maximum inter-observer consistency is, which should be the upper limit for computational models for predicting gaze. Since inter-observer consistency will never reflect a perfect prediction of one observer by the other, its actual level should serve as the goal for the consistency between model predictions and human behavior when testing a model's validity. A rate of correct predictions higher than the inter-observer consistency may even be indicative of over-fitted models. Validation of models should ideally occur with behavioral data describing real-world behavior, and these results show that setting and context of data acquisition have an impact on what is measured. The same problem also affects models of other types of behavior, but here predictions of gaze made by the Saliency Map model are investigated.

If gaze direction is dependent on visual input, saliency should be elevated at the center of gaze relative to other locations in the visual world. Average saliency indeed shows a peak at the center of the frames recorded with the gaze camera, as compared to what is recorded with the head-centered camera. The shape of the 'peak' region in the head-centered saliency deviates as well: it is a horizontal streak above the mid line of the frames. The question is if this relationship between saliency and gaze is different in the laboratory conditions. Gaze-centered saliency maps were constructed for all three conditions using the frames of the head camera. As a baseline for comparison, shuffled gaze directions have been used as well. These have the same distribution, and should hence show how strong saliency is elevated at gaze if the two are unrelated. All three conditions show a stronger peak of saliency at gaze than baseline. Furthermore, in 1s frame replay it can be observed that at the onset of the 1 second interval, the relationship between gaze and saliency is about as strong as baseline which changes after about 400 ms. In other words, luminance contrast, color contrast and orientation contrast are higher at the direction of gaze, indicating a correlation between gaze and saliency.

How strong is this relationship? Discrimination of real from shuffled gaze directions by saliency is slightly better in continuous replay, compared both to 1s frame replay and free exploration. In a few movies however, the predictions are below chance level (50% correct) and never reach levels higher than 60%, If real fixations from free exploration have to be discriminated from shuffled ones, fixations from continuous replay do better than both model saliency and fixations from 1s frame replay. This again shows that continuous replay is a better stimulation mode for capturing life-like behavior than briefly displaying static images.

All these results indicate that real-world gaze is better captured by the continuous replay condition than by the 1s frame replay condition. This has implications for laboratory experiments studying natural gaze under free-viewing conditions: more lifelike stimuli than static images may improve results of laboratory work. Furthermore, the implicit task of walking on (uneven) terrain seems to have influenced the direction of gaze in the free exploration condition. Most studies on the role of task on visual attention have thus far explored the effects of explicit tasks. Study III will focus on the implicit task of negotiating uneven terrain instead.

## Study III: Eye- and head movements in real-life terrain negotiation

In real life, the effects of visual input on movements and vice versa are both present continuously. Visual attention is used to gather information in order to perform adequate actions, and movements one makes or plans to make similarly affect the direction of gaze. In Study II it was observed that participants that actually walk through the environment make much more downward eye-movements, presumably to coordinate walking to negotiate the terrain. Participants in the lab were immobile and a qualitative interpretation of the data suggests that they paid little attention to the path. It seems very likely that the terrain shown in the videos and slide shows is largely irrelevant for the task of watching pictures on a screen.

To test this explicitly, we had participants negotiate two paths of distinct regularity in the same visual environment. The paths used are in a local street (‘Hirschberg’) where a metal railing separates a continuously inclining cobbled road (the ‘road’ condition) from a sidewalk with irregularly placed steps (‘steps’ condition). Participants were asked to walk closely to either side of the metal railing both up and down. During these four walks, eye movements and a head-centered video were recorded. By determining the position of a reference point (‘vanishing point’) in the head-centered video, the orientation of the head in space can be calculated. When adding the eye-in-head orientation signal from the eye-tracker to this, a gaze-in-world signal is obtained.

A distribution of gaze-in-world direction, head-in-world orientation and eye-in-head orientation can then be analyzed. The horizontal and vertical coordinate of the averages are analyzed, as well as the horizontal and vertical spread, for the effect of the terrain and walking direction. If the interpretation of the distributions seen in the second experiment is true, there should be a difference in gaze-in-world direction between the two terrains. That is, on the more irregular path, participants should look downward more, or fixations on the path are closer to the participants' feet. Both options are not mutually exclusive, and both should affect the vertical coordinate of the average gaze direction. Several combinations of eye-in-head and head-in-world movements could be used to redirect gaze as needed for immediate terrain negotiation. These components are measured separately, and their contributions to direction of gaze can be investigated individually.

The distribution of gaze is indeed different on both terrains. There is a peak around the 0°,0° coordinate, which represents the direction of the vanishing point. Directly below this peak, is a second peak on both terrains, presumably indicating gaze used for attending the terrain being negotiated. This second peak is about 20° lower in steps as compared to road, which is accompanied by a higher vertical spread as well. This indicates that direction of gaze is changed to meet demands posed by terrain. The contributions of head-in-world orientation and eye-in-head orientation to the direction of gaze will be investigated next.

The shape distribution of head-in-world orientation as described by its vertical and horizontal spread does not significantly differ between the terrains, indicating a similar pattern of movements. On both terrains, the average head-in-world orientation indicates that people point their head a bit downward relative to the horizon, which has been found before (Guitton & Volle, 1987). However, the average head-in-world orientation is lower in steps than in road. This shows that head orientation is used to direct attention to the path by adjusting it in a constant manner, but not by making more or different movements.

Eye-in-head orientation is distributed more vertically on both terrains, contrary to what is found in the free exploration condition described in study 2. This suggests an effect of the visual environment or of the inclination of the path, which is equal in both conditions. Additionally, in the steps condition, the vertical spread of eye-in-head is larger than in road, and the average eye-in-head orientation is also lower in steps as compared to road. This shows that gaze is directed to the path more on steps by a general re-orientation of the eye as well as by a different pattern of eye movements.

Similar experiments have been conducted before, but these either used laboratory setups with highly impoverished visual environments (Patla & Vickers, 2003; Marigold & Patla, 2007), which may affect gaze, or the terrain types tested were located in different environments (Pelz & Rothkopf, 2007). To my knowledge, this is the first study of changes in visual attention induced by an implicit task performed in a constant real-world setting. The results show that attention and gaze are directed at locations in the world that are relevant for the task at hand, or, more precisely, to visual information that is useful for motor planning in walking. This visuomotor routine (Hollands & Marple-Horvat, 1996; Guitton & Volle, 1987) is a common example of an action-perception loop. In this case, the intended action determines what visual information is necessary, which determines where gaze is directed. The visual information gathered from the perceived scene is then used to optimize performance in the task at hand. That is, action is shaped by perception. The last study investigates if and when perception is shaped by action.

## Study IV: Action-to-perception transfer

Visual input naturally has a strong influence on actions. The high spatial resolution of vision makes visual information highly suitable for guiding goal-directed actions. For example, large lateral flow-fields induce changes in the reaching trajectory, which is called the manual following response (e.g. Saijo et al., 2005). A theoretical framework for the integration of action and perception is presented by the Theory of Event Coding (TEC; Hommel et al., 2001; Prinz, 1997). The theory states that the last stages of perception – perception events – and the first stages of motor planning – (intended) action events – share representations. All sensory input may affect action this way, and the theory also predicts that action events can in turn influence perception directly. An effect of motor-learning on later perception has already been shown (Casile & Giese, 2006; Hecht et al., 2001). Some first evidence that a concurrent effect of action on perception exists as well has also been found. However, either the visual stimuli were shaped by (previous) actions (Maruya et al., 2007) or hand-movements determined when the stimulus would be presented (Wohlschläger, 2000). This confounding effect of the participants' actions on the stimulus may have affected both studies' results. Additionally, the effects of prolonged stimulation and movements are still unknown.

A perceptual rivalry paradigm is used to investigate action-to-perception transfer (APT). Rivalry is a process where one of two or more alternative interpretations ('percepts') of a constant, but ambiguous stimulus is perceived at any one moment. Which percept is dominant keeps alternating over time (Blake & Logothetis, 2002). Previous work indicates that movements affect a visually perceived rivalrous stimulus similarly (Maruya et al., 2007; Wohlschläger, 2000). This would imply that movement signals (either proprioceptive signals, efference copies or motor plans) play a similar role as a sensory modality in integrating information.

The rivalrous stimulus used here consisted of moving dots that can be perceived as a cylinder rotating clockwise or a cylinder rotating counter-clockwise. While viewing this stimulus, participants simultaneously made unseen clockwise or counter-clockwise rotating movements with their right hand. These movements could be used to report the percept, either by making a movement congruent with the percept or incongruent with the percept. The movements could also be continuous, pre-defined movements (e.g. a block of clockwise movements). Percept was then reported by key-presses which allows for splitting the percepts in congruent and incongruent with the ongoing hand movement. The behavior from these four blocks can be analyzed in a 2 x 2 design, using movement type (motor instructed vs. motor report) and action-perception congruency (movements congruent vs. incongruent with the percept) as factors. Control conditions include catch blocks that used a disambiguated stimulus to test motor responses to changes in perception, as well as blocks where an unrelated (vertical) movement is made and blocks where no movements are made.

As dependent variable, the median length of the dominance durations of congruent and incongruent percepts is used. If hand movements have a similar effect on the resolution of ambiguity as sensory input from other modalities, the percepts congruent with the movement should be longer than the percepts incongruent with the movement. The dependency of any congruency effect on the type of movements (pre-defined or task-relevant) can also be investigated in this paradigm.

The data show an interaction between movement type and action-perception congruency. Within pre-defined movements, the dominance durations of percepts congruent with the movements did not differ from the dominance durations of percepts incongruent with the movements. However, when movements were used to report the current percept, the dominance durations of percepts congruent with the current movement were longer than dominance durations of percepts incongruent with the current movement.

These data confirm earlier findings demonstrating that action affects perception (Maruya et al., 2007; Wohlschläger, 2000). In addition, it is shown that actions have to be task-relevant to induce any effect on perception (Hommel, 2004). A model of rivalry proposes that two populations each code for one of the two percepts. These populations inhibit each other and the most active population determines the percept. Because of adaptation in the active population the percept eventually switches when the other population can take over (e.g. Lankheet, 2006). If the movements would have increased adaptation of the congruent percept, dominance duration in congruent percept tracking would be lower than those in incongruent percept tracking, which is the opposite of what we find. If this arguably simple model of rivalry is still correct, this would imply that the movements increase the inhibition of the percept incongruent with the movement. Cross-modal rivalry experiments find a similar congruency effect of non-ambiguous stimuli in one modality on the perception of ambiguous stimuli in another modality (Blake et al., 2004; van Ee et al., 2009). That actions influence perceptual rivalry in a way alike to other sensory modalities in cross-modal rivalry, confirms the proposed equivalence of perceptual and action events in TEC. Resolving ambiguity in one modality by using information from another modality, or in this case from the motor system, will most likely result in more stable perception in complex and noisy, real-world situations. Using only task-relevant action information to resolve ambiguity may be the most adaptive strategy for human behavior.

## Discussion

In a series of experiments attention and perception in more life-like situations have been studied. First, it has been shown that manipulations of low-level features direct attention in natural images, and that the effects of single features add linearly, as predicted by the Saliency Map model. Second, differences between laboratory and real-world attention have been quantified in a free exploration task. Third, it has been shown that naturally occurring implicit tasks guide visual attention. Lastly, self-produced, but unseen movements affect perception when perception is relevant for action.

In these four studies different influences on attention and perception have been investigated using stimuli and tasks that allow for, or explicitly study the interaction of several processes. Natural scenes have been combined with feature gradients, a first-person perspective on free exploration has been used in real-world settings and in the laboratory, walking a path has been combined with negotiating terrain and the perception of an ambiguous stimulus with the execution of several types of movements. Taken together, these studies provide insight into many interlocking subprocesses in human vision.

The first study shows that certain bottom-up models and their assumptions hold in viewing natural scenes. However, the second study demonstrates an effect of picture onset on the relationship between gaze direction and low-level features. The fourth study even shows that the perceptual interpretation of the same set of features is altered by making unseen hand movements. This could be seen as a top-down effect, not on attention in this case, but directly on perception. Furthermore, the third study demonstrates that even implicit tasks exert a top-down influence on attention. It appears that depending on context, the visual input may exert a stronger or weaker influence over attention, or perhaps context allows for exploratory behavior. In any case, the extreme positions in the bottom-up vs. top-down debate; completely ignoring peculiarities in the visual input and ignoring task demands altogether, are likely both suboptimal strategies in real life. If gaze is controlled by a mixture of bottom-up and top-down processes (Wolfe et al., 1989; Navalpakkam & Itti, 2005), it follows that in semi-constrained tasks, a part of all fixations is directed at salient locations not relevant for the task. An example could be the time spent waiting for water to boil while making tea. Many stimuli usually present in real-life would be more interesting to look at than the kettle. In the first study it has been shown that even in a free-viewing experiment there seem to be consistent top-down effects in the form of an image scanning strategy.

Since walking on a street presumably did not require the full attention of the participants in the third study, it could be expected that some fixations were used to explore the environment. If the horizontal spread of the gaze distribution (Figure 4.2) is used as an indicator for how much time the participants spent exploring the environment, there seems to be no appreciable difference between walking on the steps or on the road. However, if we interpret the gaze distribution as consisting of a lower, task-related part indicating how much the terrain was attended and an upper, exploration part, there may be a difference. In the group data, the peak of the upper part of the gaze distribution is about as high as the peak in the lower task-related part when walking on the road, but on the steps the upper peak is lower than the lower peak. This could indicate that participants explore the environment more when walking on the road. However, an alternative explanation is that attention is both given to the path as well as the end of the path and that the ratio of attention for both task-relevant locations is shifted. On the other hand the end of the path should be only a small part of the visual space, much smaller than the upper part of the gaze distribution. In any case, the data presented here is certainly not at odds with a view of visual attention being determined by an interplay of bottom-up and top-down processes. The eye-movement patterns in the first study seem to indicate that both types of processes are simultaneously active and the fourth study even indicates modulation of perceptual processes by motor signals.

The second study demonstrates that highly similar visual stimulation evokes dissimilar patterns of gaze, depending on the mode of presenting the visual information. The fourth study even demonstrates that the perceptual processes underlying the interpretation of a constant visual stimulus can be altered by concurrent actions. Both studies underline that despite the hierarchical nature of the visual system, coupling between areas is usually bidirectional. Consequently, in the dynamic, multi-modal situations encountered in real life, visual attention will rarely be controlled by a single process. It may be that effects found in laboratory experiments do not simply add up to accurately predict behavior. Instead, hitherto unmeasured interaction between many processes may guide real-world visual attention. Although this notion is not new, the experiments described here may lend it some credibility and underline the importance to study visual attention in the real-world as well as in laboratory setups.

The first study shows that feature-contrast gradients applied to natural stimuli affect direction of gaze in the direction predicted by classic computational models, such as the Saliency Map model (Koch & Ullman, 1985; Itti & Koch, 2000). In contrast to an effect equal to the maximum of the features (Lewis & Zhaoping, 2005), this model predicts linear addition of the effects of each feature, which has been confirmed by the data. This finding guides the construction of models that make use of multiple feature-based representations of the visual input.

The fact that the manipulation of features leads to a shift in attention may suggest a causal effect of features on attention. This is, however, not necessarily the case. Other work shows for example that manipulations of features have no effect on the direction of gaze when participants engage in a search task (Henderson et al., 2007; Einhäuser et al., 2008a) and that objects explain away the effects of features on attention (Einhäuser et al, 2008b). Objects can be perceived by their features, so that the correlation between gaze and features may depend fully on visual attention being directed at objects (but see, Naber et al., 2011). The stimuli used in the study presented here did not contain man-made objects, but the manipulation of the features may also have affected the perception or visibility of the natural objects (stones, trees, leaves, etc.) that were present in the stimuli. In other words, the effects that both objects and features have on attention raises the question if there is an interaction between these two effects, or whether objects override the effects of features.

Studying object-directed attention may also help shaping task-based models for predicting gaze. Object detection and classification can already be automated to some degree (face detection: Viola & Jones 2001; Dakin & Watt 2010) and gaze is likely to be directed at objects relevant for the task. Gaze may not only be directed at stationary objects – like steps on a path – but may also be directed at locations where task relevant objects are going to be, such as the point in space where a ball and racket will hit (Hayhoe et al., 2005) or the area around the tangent point of a curve in a road, which is relevant for controlling the trajectory of a car while driving it (Land, 1992; Kandil et al., 2009). Detecting task-relevant objects or predicting their location based on dynamic visual input, may be a first step in generating task-based predictions of gaze which can be verified against actual, measured gaze.

Such approaches to visual attention in explicit tasks require recording gaze in dynamic situations. As has been shown in the second study presented here, the effects that laboratory setups have on visual attention (see also: Tatler, 2007) can be a potentially confounding factor. Both of these arguments for doing real-world recordings stem from an intended or desirable validity of research for real-life situations. In the second study, visual attention in two laboratory paradigms have been compared with each other and with real-world recordings. The results clearly indicate that using dynamic, real-world movies instead of static images evokes behavior that is more alike to what is recorded in real-world environments.

If the effects of features and objects on real-world visual attention is to be studied, as has been argued above, the requirement of real-world stimuli and recording suggest two approaches for future research. First, applying gradients of luminance- and color-contrast to dynamic stimuli, recorded in the real world, can reveal the effect low-level features have on attention in more natural stimuli than static images. Second, using a wearable eye-tracker, interactions with objects can be recorded in real environments. By keeping the objects constant, and by varying task, environment or other contextual factors, the role that objects play in directing gaze in various natural settings can be quantified. Such paradigms will shed further light on real-world attention and behavior and may reveal further differences with laboratory situations. They may also lead to better models for predicting real-world gaze, as they combine bottom-up as well as top-down influences.

The assumption underlying the view of task as the main determinant of gaze, is that sensory perception serves to gather information necessary for adequate performance of the task at hand. Even when this task is only given implicitly, it has a distinct effect on gaze, as is shown in the third study. The steps may be considered objects relevant for the task of walking. The area of the visual field where the steps are, receives more visual attention compared to walking on terrain without steps (the road). This by itself suggests that humans engage in many tasks simultaneously, as walking and talking are usually easily combined by most people. During free exploration of various environments in the second study, a large amount of visual attention is still directed at the environment. In the third study, however, gaze is mostly directed at the terrain or at the end of the path. This suggests first that free exploration of the environment may be an implicit task that humans engage in to gather information to enhance performance in the future. Second, it suggests the possibility that when the amount of visual information needed for negotiating terrain decreases, gaze can be used for free exploration more. One way to reduce the amount of visual information needed for negotiating terrain may be to memorize or learn the terrain. If walking the same irregular path several times results in gaze being directed at the path less, this would be a further indication that real-life gaze serves to gather information for adequate performance of tasks. However, if practice does not result in gaze being directed upward, away from path more, but instead remains equally fixed on the irregular path, this does not mean that the terrain is not learned. It may instead indicate that visual exploration of the environment is not an alternative to the task of negotiating terrain, or that the environment is learned just as much as the path. Learning the environment could remove the necessity to explore it visually so that attention can be directed at safely negotiating terrain in the present. In short, learning seems to be a good candidate paradigm to further clarify the effects of tasks in real-world situations and may shed light on the functional role, if there is any, of free exploration of the environment.

The possibility that learning the environment may free up attention for path negotiation is hard to distinguish from not learning the terrain, but both learning the terrain and learning the environment imply a role of long-term memory in directing gaze in real-world situations. Short-term memory effects have already been shown, in so called “deictic pointers” (Ballard et al., 1995; Ballard et al., 1997) used in a simple construction task. Deictic pointers supposedly make it easier to look at a point in visual space repeatedly. This contrasts with so-called “inhibition of return”, observed in search tasks in the laboratory, where participants are less likely to revisit the same point in visual space than random other locations. Given that the objects don’t move, this makes sense in a single search. However, in a real life search, we may often inspect the same locations as experience may have taught us that we are likely to have misplaced items in these locations. In general, real-life- or laboratory tasks where rewards are more likely to appear at certain locations than at others, should demonstrate long-term memory effects that may be labeled “facilitation of return”. Such learning effects could be seen as more specific versions of priors (Torralba, 2003). Studying visual search in real-life would reveal if behavior in agreement with facilitation of return does or does not occur. This would clarify if inhibition of return is a general principle of the visual system or if it is only engaged in within very specific laboratory search tasks.

The third experiment assumes that task guides attention and that what is then perceived affects performance of the task. This process keeps repeating itself, forming a perception-action loop. What has been shown in the fourth study is that self-produced, unseen movements affect the perceptual interpretation of constant visual stimuli directly. That is, the visual stimulus is not changed by movements made by the participants, and nevertheless these movements affect how the stimulus is perceived. The disambiguation of constant but ambiguous stimuli is similarly affected by what is perceived in other sensory modalities. This is in accordance with the Theory of Event Coding (TEC; Hommel et al., 2001; Prinz, 1997) that states that perceived visual events are represented in areas closely related to intended motor events. Motor intentions can affect perceived visual events according to TEC. In the experiment presented here, the movements were not only intended, but actually executed simultaneously with reporting perception. Consequently, efference copies and proprioceptive feedback as well as motor intentions were available continuously in all conditions with hand movements and the experiment cannot dissociate which of these signals caused the effects. An interesting follow-up study would be to investigate the perception of ambiguous stimuli during the planning of a movement. If a motor intention by itself suffices to alter the perception of a visual stimulus, this implies a role for intended motor events, but this would still have to be contrasted with simultaneously executed movements to quantify if there is a role for efference copies or proprioceptive feedback as well. In any case, the effect found in the fourth study shows that for interpreting noisy or unclear information in a certain modality, the brain does not only use information from other modalities, but signals from the motor system too.

Buswell (1935) used a search task to compare to a free-viewing ‘task’ which demonstrated an effect of instruction on visual attention. Interestingly, behavior observed in tasks such as feature search and conjunction search have led to “bottom-up” models of attention. The validity of these models is now under heavy debate and task is suggested as the prime or even sole cause of shifts in visual attention (Ballard & Hayhoe, 2009). Although it is unlikely that features play a causal role in directing attention in all tasks, they may be used in real-life search tasks. An ecologically valid example would be search for berries or other natural objects defined by relatively simple features. To the best of my knowledge, no data on real-life search experiments has been published to date. Hence, real-life search would be a potentially fruitful paradigm for future studies.

With these experiments a wide range of topics in visual attention in naturalistic situations is covered. By using natural stimuli and tasks, and conducting experiments out of the lab, results obtained with laboratory tasks have been tested for external and ecological validity. Some behavior, such as walking through a street, is impossible, or at least very hard to investigate in the lab. Hence, measurements performed in the actual environment are very useful to study psychological constructs, such as attention. Though the range of topics does not cover all aspect of visual attention by far, this thesis makes clear that doing experiments in more natural settings is not only feasible but even necessary if the implications of experiments are to reach beyond the walls of the laboratory.

## References

Amedi, A., Von Kriegstein, K., Van Atteveldt, N.M., Beauchamp, M.S., Naumer, M.J. (2005). Functional imaging of human crossmodal identification and object recognition. Exp Brain Res, 166, 559-571.

Ballard, D.H., Hayhoe, M.M. (2009). Modelling the role of task in the control of gaze. Vis Cog, 17(6), 1185-1204.

Ballard, D.H., Hayhoe, M., Pelz, J.B. (1995). Memory representations in natural tasks. J Cogn Neurosci, 7(1), 66-80.

Ballard, D.H., Hayhoe, M., Pook, P.K., Rao, R.P.N. (1997). Deictic codes for the embodiment of cognition. Behav Brain Sci, 20(4), 723-742.

Barlow, H.B. (1961). Possible principles underlying the transformations of sensory messages. In: Rosenblith, W.A. (Ed.), Sensory communication. Endicott House: MIT Press, 217-234.

Betz, T., Kietzmann, T.C., Wilming, N., König P. (2010). Investigating task-dependent top-down effects on visual attention. J Vis, 10(3):15, 1-14.

Blake, R., Logothetis, N.K. (2002). Visual Competition. Nat Rev Neurosci, 3(1): 13-21.

Blake, R., Sobel, K. V., James, T. W. (2004). Neural synergy between kinetic vision and touch. Psychol Sci, 15, 397-402.

Booth M.C., Rolls E.T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cereb Cortex, 8(6), 510-23.

Buswell, G.T. (1935). How people look at pictures: A study of the psychology of perception in art. Chicago: University of Chicago Press.

Callow, D., Lappe, M. (2008). Efficient encoding of natural optic flow. Netw Comput Neural Syst, 19(3), 183-212.

Carmi, R., Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. Vis Res, 46, 4333-4345.

Casile, A., Giese, M.A. (2006). Nonvisual motor training influences biological motion perception. Curr Biol, 16, 69-74.

Dakin, S.C., Watt, R.J. (2009). Biological “bar codes” in human faces. J Vis, 9(4):2, 1-10.

Derrington, A.M., Krauskopf, J., Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. J Physiol, 357, 241-265.

Einhäuser, W., Rutishauser, U., Koch, C. (2008a). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. J Vis, 8(2):2, 1-19.

Einhäuser, W., Spain, M., Perona, P. (2008b).Objects predict fixations better than early saliency. J Vis, 8(14):18, 1-26.

Einhäuser, W., Schumann, F., Vockeroth, J., Bartl, K., Cerf, M., Harel, J., Schneider, E., König, P. (2009). Distinct roles for eye and head movements in selecting salient image parts during natural exploration. Ann N Y Acad Sci, 1164, 188-193.

Einhäuser, W., König, P. (2010). Getting real – sensory processing of natural stimuli. Curr Opin Neurobiol, 20(3), 389-395.

Goodale M.A., Milner A.D. (1992). Separate visual pathways for perception and action. Trends Neurosci, 15(1), 20-25.

Guitton, D., Volle, M. (1987). Gaze control in humans: Eye-head coordination during orienting movements to targets within and beyond the oculomotor range. J Neurophysiol, 58(3), 427-459.

Hayhoe, M., Mennie, N., Sullivan, B., Gorgos, K. (2005). The role of internal models and prediction in catching balls. Proc Conf AAAI Artif Intell 2005 Fall Symposium.

Hayhoe, M., Ballard, D. (2005). Eye movements in natural behavior. Trends Cogn Sci, 9(4), 188-194.

Hecht, H., Vogt, S., Prinz, W. (2001). Motor learning enhances perceptual judgment: a case for action-perception transfer. Psychol Res, 65, 3-14.

Henderson, J.M., Brockmole, J.R., Castelhano, M.S., Mack, M. (2007). Visual saliency does not account for eye-movements during visual search in real-world scenes. In: R.P.G. van Gompel, M.H Fischer, W.S. Murray, R.L. Hill (Eds.), Eye movement research: Insights into mind and brain. Oxford: Elsevier, 437-562.

Hollands, M.A., Marple-Horvat, D.E. (1996). Visually guided stepping under conditions of step cycle-related denial of visual information. Exp Brain Res, 109, 343-356.

Hommel, B. (2004). Event files: feature binding in and across perception and action. Trends Cogn Sci, 8, 494-500.

Hommel, B., Müsseler, J., Aschersleben, G., Prinz, W. (2001). The Theory of Event Coding (TEC): A framework for perception and action planning. Behav Brain Sci, 24(5), 849-937.

Hubel, D.H., Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol, 160, 106-154.

Itti, L., Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res, 40, 1489-1506.

Kandil, F.I., Rotter, A., Lappe, M. (2009). Driving is smoother and more stable when using the tangent point. J Vis, 9(1):11, 1-11.

Kingston, A., Smilek., D., Eastwood, J.D. (2008). Cognitive Ethology: A new approach for studying human cognition. Br J Psychol, 99(3), 317-340.

Koch, C., Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. Hum Neurobiol, 4, 219-227.

Land, M.F. (1992). Predicting eye-head coordination during driving. Nature, 359(6393), 318-320.

Land, M.F., McLeod, P. (2000). From eye movements to actions: how batsmen hit the ball. Nat Neurosci, 3, 1340-1345.

Land, M., Mennie, N., Rusted, J. (1999). The role of vision and eye movements in the control of activities of daily living. Perception, 28, 1311-1328.

Land, M.F., Tatler, B.W. (2001). Steering with the head: The visual strategy of a racing driver. Curr Biol, 11, 1215-1220.

Lankheet, M.J.M. (2006). Unraveling adaptation and mutual inhibition in perceptual rivalry. J Vis, 6(4):1, 304-310.

Lewis, A., Zhaoping, L. (2005). Saliency from natural scene statistics. Abstract Viewer/Itinerary planner. Washington DC: Society for Neuroscience. Program No. 821.11.

Li, Z. (2002). A saliency map in primary visual cortex. Trends Cog Sci, 6(1), 9-16.

Marigold, D.S., Patla, A.E. (2007). Gaze fixation patterns for negotiating complex ground terrain. Neuroscience, 144, 302-313.

Maruya, K., Yang, E., Blake, R. (2007). Voluntary action influences visual competition. Psychol Sci, 18, 1090-1098.

Mishkin, M., Ungerleider, L.G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. Behav Brain Res, 6(1), 57-77.

Moran, J., Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. Science, 229(4715), 782-784.

Munneke, J., Heslenfeld, D.J., Theeuwes, J. (2008). Directing attention to a location in space results in retinotopic activation in primary visual cortex. Brain Res, 1222, 184-191.

Naber, M., Carlson, T.A., Verstraten, F.A.J., Einhäuser, W. (2011). Perceptual benefits of objecthood. J Vis, 11(4):8, 1-9.

Navalpakkam, V., Itti, L. (2005). Modelling the influence of task on attention. Vis Res, 45(2), 205-231.

Østerberg, G. (1935). Topography of the layer of rods and cones in the human retina. Acta Ophthalmol, Suppl. 13(6), 1-102.

Patla, A.E., Vickers J.N. (2003). How far ahead do we look when required to step on specific locations in the travel during locomotion? Exp Brain Res, 148, 133-138.

Pelz, J.B., Rothkopf, C. (2007). Oculomotor behavior in natural and man-made environments. In: R.P.G. van Gompel, M.H Fischer, W.S. Murray, R.L. Hill (Eds.), Eye movement research: Insights into mind and brain. Oxford: Elsevier, 661-676.

Peters, R.J., Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In: Proc. IEEE (CVPR).

Posner, M.I. (1980). Orienting of attention. Q J Exp Psychol, 32, 3-25.

Prinz, W. (1997). Perception and action planning. Eur J Cogn Psychol, 9(2), 129-154.

Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C., Fried, I. (2005). Invariant visual representation by single neurons in the human brain. Nature, 435(7045), 1102-1107.

Rizzolatti, G., Riggio, L., Dascola, I., Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. Neuropsychologia, 25(1A), 31-30.

Sackett, P.R., Zedeck, S., Fogli, L. (1988). Relations between measures of typical and maximum job performance. J Appl Psychol, 73(3), 482-486.

Saijo, N., Murakami, I., Nishida, S., Gomi, H. (2005). Large-field visual motion directly induces an involutary rapid manual following response. J Neurosci, 25(20), 4941-4951.

Saito, H.-A., Yukie, M., Tanaka, K., Hikosaka, K., Fukada, Y., Iwai, E. (1986). Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey. J Neurosci, 6(1), 145-157.

Schall, S., Quigley, C., Onat, S., König, P. (2009). Visual stimulus locking of EEG is modulated by temporal congruency of auditory stimuli. Exp Brain Res, 198(2-3), 137-151.

Schneider, E., Villgrattner T., Vockeroth J., Bartl K., Kohlbecher S., Bardins S., Ulbrich H., Brandt T. (2009). EyeSeeCam: an eye movement-driven head camera for the examination of natural visual exploration. Ann N Y Acad Sci, 1164, 461-467.

Schneider, W., Shiffrin, R.M. (1977). Controlled and automatic human information processing: I. Detection, search and attention. Psychol Rev, 84(1), 1-66.

Schumann, F., Einhäuser, W., Vockeroth, J., Bartl, K., Schneider, E., König, P. (2008). Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. J Vis, 8(14):12, 1-17.

Shiffrin, R.M., Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. Psychol Rev, 84(2), 127-190.

Tatler, B.W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. J Vis, 7(14):4, 1-17.

Tootell, R.B.H., Reppas, J.B., Kwong, K.K., Malach R., Born, R.T., Brady, T.J., Rosen, B.R., Belliveau, J.W. (1995). Functional Analysis of Human MT and Related Visual Cortical Areas Using Magnetic Resonance Imaging. J Neurosci 15(4), 3215-3230.

Torralba, A. (2003). Contextual priming for object detection. Int J Comput Vis, 53(2), 169-191.

Treisman A., Gelade G. (1980). A feature integration theory of attention. Cogn Psychol, 12, 97–136.

van Ee, R., van Boxtel, J. J., Parker, A. L., Alais, D. (2009). Multisensory congruency as a mechanism for attentional control over perceptual selection. J Neurosci, 29, 11641-11649.

Vincent, B.T., Troscianko, T., Gilchrist, I.D. (2007). Investigating a space-variant weighted salience account of visual selection. Vis Res, 47(13), 1809-1820.

Vincent, B.T., Baddely, R.J., Troscianko, T., Gilchrist, I.D. (2009). Optimal feature integration in visual search. J Vis, 9(5):15, 1-11.

Viola, P., Jones, M.J. (2001). Rapid object detection using a boosted cascade of simple features. Comput Vis Pattern Recog, 1, 511-518.

Wohlschläger, A. (2000). Visual motion priming by invisible actions. Vis Res, 40, 925-930.

Wolfe J.M., Cave K.R., Franzel S.L. (1989) Guided search: An alternative to the feature integration model for visual search. J Exp Psychol Hum Percept Perform, 15(3), 419-433.

Wolfe, J.M. (2007). Guided search 4.0: Current progress with a model of visual search. In: W. Gray (Ed.) Integrated models of cognitive cystems. New York: Oxford, 99-119.

Yarbus, A.L. (1967). Eye movements and vision (B. Haigh, Trans.). New York: Plenum.

# Experiments

Study I:

Engmann, S., ’t Hart, B.M., Sieren, T., Onat, S., König, P. and Einhäuser, W. (2009). Saliency on a natural-scene background: Effects of color- and luminance-contrast add linearly. Attent Percept Psychophys, 71(6):1337-52.

Study II:

’t Hart, B.M., Vockeroth, J., Schumann, F., Bartl, K., Schneider, E., König, P. and Einhäuser, W. (2009). Gaze allocation in natural stimuli: comparing free exploration to head-fixed viewing conditions. Vis Cog, 17(6+7), 1132-1158.

Study III:

’t Hart, B.M. and Einhäuser, W. (submitted). Mind the step: complementary roles for eye-in-head and head-in-world orientation when negotiating a real-life path.

Study IV:

Beets, I.A.M., ’t Hart, B.M., Rösler, F., Henriques, D.Y.P., Einhäuser, W., & Fiehler, K. (2010). Online action-to-perception transfer: only percept-dependent action affects perception. Vis Res, 50(24), 2633-2641.

## *Study I*

*Saliency on a natural-scene background: Effects of color- and luminance-contrast add linearly*

Published as:

Engmann, S., ’t Hart, B.M., Sieren, T., Onat, S., König, P. and Einhäuser, W. (2009). Saliency on a natural-scene background: Effects of color- and luminance-contrast add linearly. Attent Percept Psychophys, 71(6):1337-52.

# Saliency on a natural-scene background: Effects of color- and luminance-contrast add linearly

### *Abstract*

In natural vision, shifts in spatial attention are associated with shifts of gaze. Computational models of such "overt" attention typically use the concept of a "saliency map": normalized maps of center-surround differences are computed for individual stimulus features and added linearly to obtain the saliency map. While the predictions of such models correlate with fixated locations better than chance, their mechanistic assumptions are less well investigated. Here we test one key assumption: do effects of different features add linearly or according to a max-type of interaction? We measure the eye-position of observers viewing natural stimuli, whose luminance-contrast and/or color-contrast (saturation) increase gradually towards one side. We find that these "feature gradients" bias fixations towards regions of high contrasts. When two contrast gradients (color and luminance) are superimposed, linear summation of their individual effects predicts their combined effect. This demonstrates that the interaction of color- and luminance-contrast with respect to human overt attention is – irrespective of the precise model – consistent with the assumption of linearity, but not with a max-type interaction of these features.

## Introduction

While inspecting complex natural scenes, human observers sequentially allocate attention to subsets of the stimulus (James, 1890). Under natural conditions, shifts in attention are typically associated with shifts of gaze (Rizzolatti, Raggio, Dascola & Umilta, 1987). Several factors guide this "overt" attention (Buswell, 1935; Yarbus, 1967), such as the task, the observer's experience, and the features of the stimulus. Models of the latter, "bottom-up", factors are often based on the concept of a so-called saliency map (Koch & Ullman, 1985): various feature channels (luminance, color, orientation, etc.) are analyzed independently, local center-surround filters yield maps of differences (“contrasts”) in these features, and these maps are added up. Following the saliency-map literature, such maps in a single feature are referred to as “conspicuity” maps. These conspicuity maps are then added linearly across features to obtain the saliency map, which represents the likelihood of a location to be attended. Various studies have demonstrated that implementations of this model predict human fixations in natural scenes at levels above chance (Itti & Koch, 2000; Parkhurst, Law & Niebur, 2002; Peters, Iyer, Itti & Koch, 2005; Tatler, Baddeley & Gilchrist, 2005). In addition, luminance-contrast is significantly elevated at fixation points (Krieger, Rentschler, Hauske, Schill & Zetzsche, 2000; Mannan, Ruddock & Wooding, 1997; Reinagel & Zador, 1999). This correlative effect of contrast depends, however, on spatial frequency (Mannan et al., 1997, Tatler et al., 2005) and acts mostly indirectly through correlations to higher order scene structure (Einhäuser & König, 2003), which may include texture contrast (Parkhurst & Niebur, 2004), edge density (Baddeley & Tatler, 2006) or objects (Elazary & Itti, 2008; Einhäuser, Spain, & Perona, 2008) and faces (Cerf, Harel, Einhäuser, & Koch, 2008). In sum, while some correlative prediction performance of saliency map models for humans freely viewing natural scenes under laboratory conditions and without a specific search task stands mostly undisputed (Parkhurst et al., 2002; Peters et al., 2005), recent evidence has substantially undermined their causal and mechanistic implications.

Despite a large body of data on the neural representation of saliency (Gottlieb, Kusunoki & Goldberg, 1998; Horwitz & Newsome, 1999; Mazer & Gallant, 2003; Kustov & Robinson, 1996; McPeek & Keller, 2002; Posner & Petersen, 1990; Thompson, Bichot & Schall, 1997), the mechanistic principles underlying its computation are less well understood. Koch and Ullman’s (1985) model is founded on neural principles, but does not make any explicit reference to the nature of interactions between feature-channels. In contrast, most later saliency-map implementations (Itti, 2005; Itti & Koch, 2000; Peters et al., 2005) make the critical assumption that feature effects add linearly. First, the conspicuity maps for each feature are linearly summed, and second, possible dependencies between features are neglected when obtaining the final saliency map. In addition, most models of visual attention that are *not* based on the saliency map still implicitly share the assumption of linearity (Wolfe, Butcher, Lee & Hyle, 2003). Several studies test this assumption using well controlled, albeit artificial, stimuli. Using grids of bars in a matching task, Nothdurft (2000) finds that different features are additive, though their interaction may be sub-linear. Along these lines, for the features of color and orientation, Li (2002) contradicts the assumption of linearity and instead proposes the overall saliency of an item to be defined by the most salient feature alone. This implies a maximum operation rather than a linear summation across features to compute saliency. Recently, research from the same lab has suggested that this maximum operation might also apply to human overt attention in natural scenes (Lewis & Zhaoping, 2005) and suggested a computation of saliency as early in the visual hierarchy as V1. In contrast, Navalpakkam & Itti (2005) have argued that linear summation is more compatible with performance in conjunction search experiments. Complementary to the question under which conditions low-level features influence fixations at all, it has remained open how the effects of different features interact. Irrespective of whether the features' effects are causal or correlative, the answer will constrain models of attention.

Besides linearity, the independence of different feature channels comprises the second major assumption of most saliency models. In a discrimination task on grating stimuli, Morrone, Denti and Spinelli (2002) find that the features of color and luminance recruit independent attention channels. However, the extent to which such results can be transferred to natural stimuli – where higher order dependencies between features not only exist, but are also exploited by the visual system (Golz & MacLeod, 2002) – remains to be investigated. When it comes to natural scenes, stimulus features are not independent, but highly correlated. In the context of overt attention, Baddeley & Tatler (2006) show that conditioned on edge-density, other feature maps have little predictive power, that is, one feature can "explain away" the effect of others. Consequently, when measuring attention in natural scenes directly, such stimulus-inherent correlations need to be considered.

Here we combine the usage of natural scenes, with modifications that are independent along the two stimulus dimensions under investigation (color and luminance). We adopt a previously proposed paradigm (Einhäuser, Rutishauser, Frady, Nadler, König, & Koch, 2006) to bias attention by increasing contrast towards one side of the stimulus ("feature gradients"). We compare effects on fixated locations of gradients in color contrast, which is modulated by varying saturation, and in luminance contrast to the effect of the feature gradients applied simultaneously. This allows us to test directly how well a linear interaction of color and luminance contrast predicts their combined effect against a natural scene background.

## Methods

### *Participants*

Eight students of the Philipps University Marburg (3 female, 5 male, age: 20-27, mean: 22.3) participated in the study. All participants had normal or corrected-to-normal vision and normal color vision as assessed by the Ishihara 16-plate color blindness test. They were naïve to the purpose of the study and had not previously viewed the stimuli used. All procedures conformed to national and institutional guidelines for experiments on human observers, and to the Declaration of Helsinki. All participants gave informed written consent for participation in this study and were paid as compensation.

### Experimental Setup

Experiments were conducted in a dark room with negligible ambient light levels. Stimuli were presented using a 19.7-inch EIZO FlexScan F77S CRT monitor located at 85 cm distance from the participant, and the stimulus subtended an angle of 26°x18°. The display resolution was set to 1280 x 1024 pixels and its refresh rate to 100 Hz. The monitor was characterized ("calibrated") using a PR-650 spectrometer (Photo Research, Chatsworth, CA) and – for low luminance values – a S370 photometer (UDT instruments, San Diego, CA). Gun CIE coordinates of the monitor were at x=0.610, y=0.339 (red), x=0.282, y=0.601 (green), and x=0.151, y=0.065 (blue), maximum luminance at 36.9 cd/m$^2$; luminance of the dark screen (black) was at 0.001 cd/m$^2$.

During the experiment, observers' eye position was recorded at 2000Hz using an infrared, non-invasive Eyelink-2000 eye tracking system (SR Research Ltd., Mississauga, Ontario, Canada). Standard procedures, as recommended by the manufacturer, were used to calibrate the eye tracker, and to validate the eye position. In brief, 13 fixation points were presented before each experimental block to compute the mapping from eye-tracker signal to screen coordinates. The calibration was then verified with a similar display and was 0.4° RMS on average and never larger than 1°. Before each trial observers were asked to fixate a fixation point in the center of the screen for at least 300 ms. If they failed to do so within 5 s, the eye tracker was recalibrated.

All stimulus presentation and eye position recording was programmed in Matlab (MathWorks, Natick, MA, USA), using its psychophysics and eyelink toolbox extensions (Brainard, 1997; Cornelissen, Peters, & Palmer, 2002, http://psychtoolbox.org; Pelli, 1997). The data was preprocessed in Python 2.5 (http://www.python.org) and statistical analysis was performed in R 2.5.1 (http://www.R-project.org).

### Stimulus Database

All stimuli were based on a set of 90 photographs of natural scenes selected from the Zurich Natural Image database (Einhäuser, Kruse, Hoffmann & König, 2006), which are available from the authors at http://www.klab.caltech.edu/~wet/ZurichNatDB.tar.gz. The images depict natural

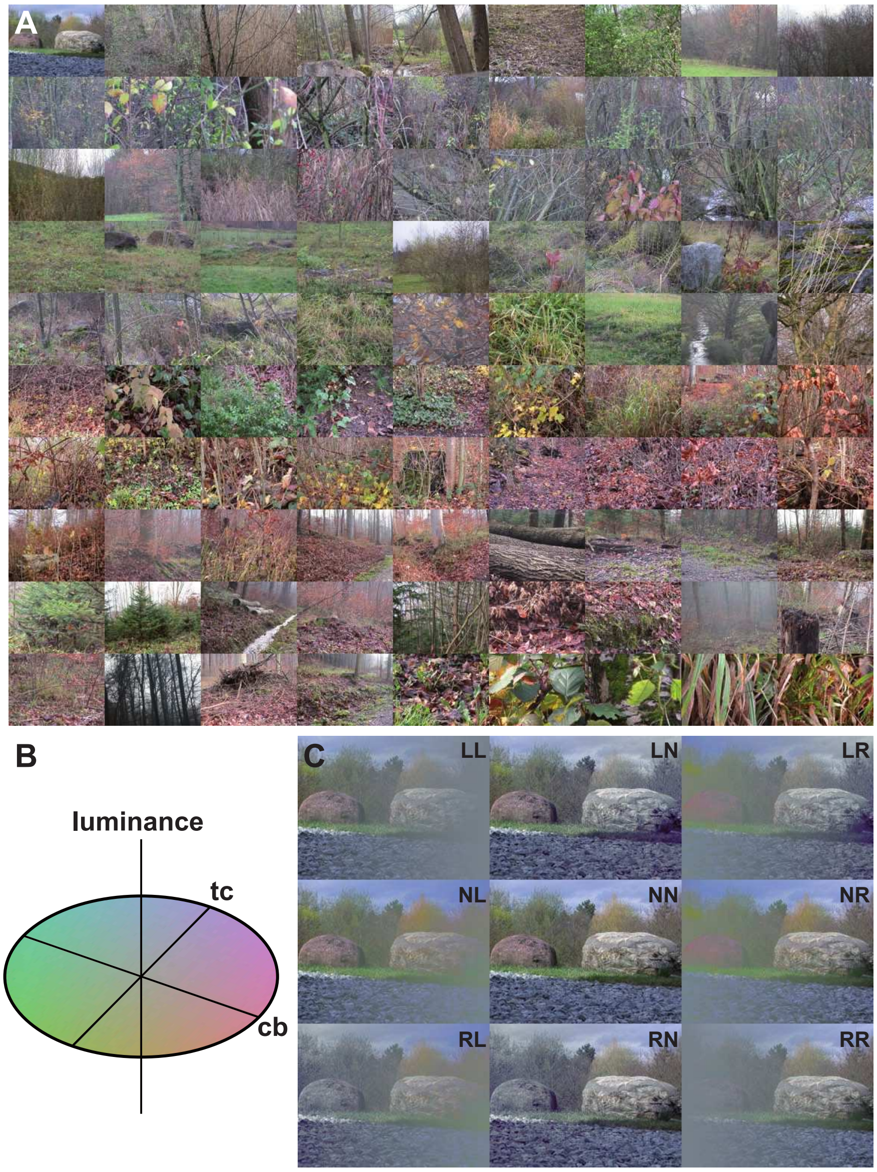

***Figure 2.1 Stimuli.***

*A) The 90 stimuli of the Zürich Natural Image Database used for the experiment. B) Schematic representation of the DKL color space. C) Columns: modification of luminance contrast; increase to left, none, to right; rows: modification of color-contrast; increase to left, none, to right. Letters denote condition abbreviation (gradient increasing to Left, Right or Neutral, first letter color, second luminance contrast).*

outdoor scenes, which only rarely contain isolated nameable objects or man-made artifacts (figure 2.1A). The images were captured using a digital camera (3.3 Mega pixel color mosaic CCD, Nikon Coolpix 995, Tokyo, Japan) with high quality settings. The stimuli were stored at a resolution of 2048 x 1536 pixels and color-depth of 24 bits in RGB format. To fit the screen resolution, images were down-sampled to 1280 x 960 pixels using bicubic interpolation in Matlab, and presented at the center of the 1280 x 1024 pixel screen.

### *Color Space*

Stimuli were characterized and modified in the DKL color space (Derrington, Krauskopf, & Lennie, 1984; figure 2.1B). This space is defined physiologically using the relative excitations of the 3 types of retinal cones. It is spanned by the orthogonal axes luminance, “constant blue” (cb, the difference between L and M cone excitations) and “tritanopic confusion” (tc, L + M - S cone excitations). Hue in DKL space is given by the azimuth, luminance by the respective axis and saturation by the projection on an isoluminant plane.

In DKL space, we defined luminance-contrast (LC) as variation along the space's luminance axis. Color contrast (CC) – as used in saliency map models – is inspired by the excitation of color-opponent cells in retina and thalamus. Hence it scales linearly with saturation and we modified CC by varying saturation. The mapping from DKL space uses the known parameters of the screen’s guns, in particular correcting for their non-linearities (“gamma”). Since the camera parameters were unknown, they were assumed to be the inverse of the screen. This guaranteed that an unmodified stimulus looks natural and all stimuli fit within the gamut of the screen.

### *Stimulus Modification: Feature Gradients*

To modify the stimulus features of interest (LC and CC) without introducing novel local image structure, we adapted the feature gradient technique introduced in Einhäuser, Rutishauser et al. (2006). Here, images were first converted into DKL-color space. To modify luminance contrast, we first subtracted the mean image luminance $<I^0>$ from the luminance values $I^0(x,y)$ of the original image. We then multiplied the luminance with a value depending on the horizontal position ("gradient"). For contrast increase to the right ("R"), this factor ranged linearly from 0 on the left to 1 on the right, and the converse held for contrast increase to the left ("L"). Finally, the original mean value was added:

Modification "R": $I(x,y) = x/w\ (I^0(x,y)-<I^0>) + <I^0>$

Modification "L": $I(x,y) = (1-x/w)\ (I^0(x,y)-<I^0>) + <I^0>$

where w denotes the image width (w=1280 pixels). Intuitively, the low-end of the gradient reduces the contrast to 0 as it clamps all luminance values to the mean image luminance ($I(x,y)=<I^0>$) at the high-end the image, and thus the contrast, remains unaffected ($I(x,y)= I^0(x,y)$). Both is most easily exemplified by an image only consisting of an equal number of black and white pixels. In the appendix we provide a detailed analysis as to how the gradient definition relates to common definitions of luminance contrast. As a consequence of the orthogonality of the DKL space, this modification did not affect physical color at any point (neither hue nor saturation).

To modify color-contrast, we similarly subtracted the means along the tc and cb axes, multiplied the result by the gradient from 0 to 1 ("R") or 1 to 0 ("L"), and shifted back to the original mean:

Modification "R":

$T(x,y) = x/w\ (T^0(x,y)-<T^0>) + <T^0>$

$$C(x,y) = x/w\ (C^0(x,y)-\langle C^0\rangle) + \langle C^0\rangle$$

Modification “L”:

$$T(x,y) = (1-x/w)\ (T^0(x,y)-\langle T^0\rangle) + \langle T^0\rangle$$

$$C(x,y) = (1-x/w)\ (C^0(x,y)-\langle C^0\rangle) + \langle C^0\rangle$$

where C and T denote the values along the cb and tc axis respectively, superscript 0 the original image and <.> the image mean as above. This varied the saturation of each pixel from 0 to its original value across the image. Intuitively, the usage of saturation as proxy for color contrast can be understood by considering an isoluminant red-green grating, which at 0 saturation would be a mere gray patch (0 color-contrast) and would take maximum color contrast whenever saturation is at 100%. To formalize this, we demonstrate in the appendix that this modification affects color conspicuity in the expected way.

Taking advantage of the orthogonality of DKL space, both gradients could be combined without interaction on the physical stimulus. The modified stimuli were converted back to RGB-space using the screen’s gun’s specifications (figure 2.1C).

### Notation for modifications

As shorthand notation, we denote conditions by two-letter abbreviations, where the first characterizes the color modification, the second the luminance modification, with “L” implying contrast increase to the left, “R” to the right and “N” no modification. For example, LN denotes a stimulus solely modified in color with gradient increasing to the left and no changes in luminance, while NN denotes an unmodified stimulus (figure 2.1C). Where there is no risk of ambiguity, the same abbreviations are also used to denote the corresponding effect sizes. We will refer to the conditions, in which a single gradient is applied (LN, NL, RN, NR) as “single-feature conditions”, to the conditions, in which two gradients are superimposed (LL, RR, LR, RL) as “dual-feature conditions”. For part of the analysis, we consider the effects of each modification relative to the modulation of eye position in the unmodified condition (NN). As short-hand notation, we used brackets [.] to denote subtraction of NN (e.g., [LN] := LN-NN).

### *Paradigm*

For all observers, each of the 90 images was presented in each of the 9 conditions exactly once. The experiment was split in 9 blocks of 90 trials. Trials were balanced such that per block each image appeared once and each condition 10 times. Since pilot experiments demonstrated little change in effect after 2 s, each stimulus was presented for 2 s. A trial started with fixation cue at the center of the screen. As soon as the participant's gaze was steady on this cue for at least 300 ms, stimulus presentation was triggered. Observers were instructed to "study the images carefully", be "free to move [their] eyes naturally", and to "reduce head movements as much as possible". None of our previous studies, which used the same instruction of "studying images carefully", showed any evidence that this induced a top-down bias. To the contrary, eye movement patterns are indistinguishable from an explicit instruction of "free viewing" (Steinwender & König, unpublished observations).

### *Data analysis*

#### Fixations

The main body of the analysis was based on periods of fixation, which accounted for 77.3% of total data. Fixations were defined by the default algorithm implemented in the Eyelink system as periods between saccades. Saccades were therefore defined as movements that exceed an acceleration threshold (9500 deg/s$^2$) and a velocity threshold (35 deg/s). Although no explicit lower limit for the duration of a fixation was used, 96.1% of fixations lasted longer than 100 ms. The initial central fixation for each stimulus originated from the fixation cue and was not used for any analysis.

### Statistical analysis

Since in 64.5% of the 2-s trials there were at least 5 fixations, but 6 or more fixations were only reached in 35.2%, we restricted fixation analysis to the first 5. For each observer the average horizontal coordinate of each of the first five fixations was calculated. A linear model ANOVA was performed with this dependent variable using fixation number (1 - 5), color contrast condition (L,N,R) and luminance contrast condition (L,N,R) as factors. Linear model ANOVA's are also performed over the two sets of single feature data where the data with manipulations in the other feature is left out (only using the data from LN, NN and RN or from NL, NN and NR). To see how the effect of the single feature manipulations develop over time, post-hoc t-tests were done on the average horizontal coordinate for each fixation, comparing LN with RN and NL with NR.

## Results

### *Number of fixations*

We recorded eye-movements in 8 observers while they were viewing natural scenes upon which a gradient in luminance-contrast (LC), color-contrast (CC), or both was superimposed. For each of the 9 conditions defined by the directions of those gradients we recorded 90 trials in each observer. On average, observers made 5.2 fixations on each stimulus. We did not find evidence that this number was dependent on the luminance condition, the color condition, nor their interaction (all $p > 0.07$). Consequently, we could directly compare different conditions on the basis of fixation locations.

### Fixation maps

In the image of figure 2.2A color contrast increased to the right (condition: “RN”). In this example, the observers’ fixations exhibited a bias towards the right, the side of increased color contrast. To visualize this effect across all observers and images, we computed an average fixation map for each condition. That is, we histogrammed fixated locations and aggregated the histograms over all fixations (excluding the initial, central one), observers and images (figure 2.2B). In all conditions the center of mass of these maps was slightly (1.2° to 1.5°) above the midline. That is, the horizontal gradient had little effect on vertical eye position. In contrast, the horizontal location depended on the condition: for unmodified images, there was a slight (0.3°) bias to the left. If both gradients pointed to the left (LL), however, the center of mass was shifted 2.5° to the left, if both gradients pointed to the right (RR), the shift was 2.7° to the right. For single-feature gradients (LN, RN, NL, NR) the center-of-mass shifts were smaller, but always to the higher (color- or luminance) contrast side (0.5°, 0.9°, 1.5° and 1.7°, respectively). The incongruent gradients showed a slight bias towards the higher luminance-contrast, consistent with the somewhat larger effect of this feature as compared to color. This first qualitative and aggregate analysis of horizontal eye position was suggestive of a superposition between the effects of color and luminance-contrast gradients on horizontal eye-position, on which the further quantitative analysis is based.

### Overall effect of gradients

We performed a 3-way ANOVA to characterize the dependence of average horizontal fixation location on fixation number (1-5), luminance contrast condition (L,N,R) and color contrast condition (L,N,R). Each factor had a significant effect (all $p<0.0001$; $F(4,315)=33.3$; $F(2,315)=78.0$; $F(2,315)=20.5$, respectively). There were no two-way interactions between

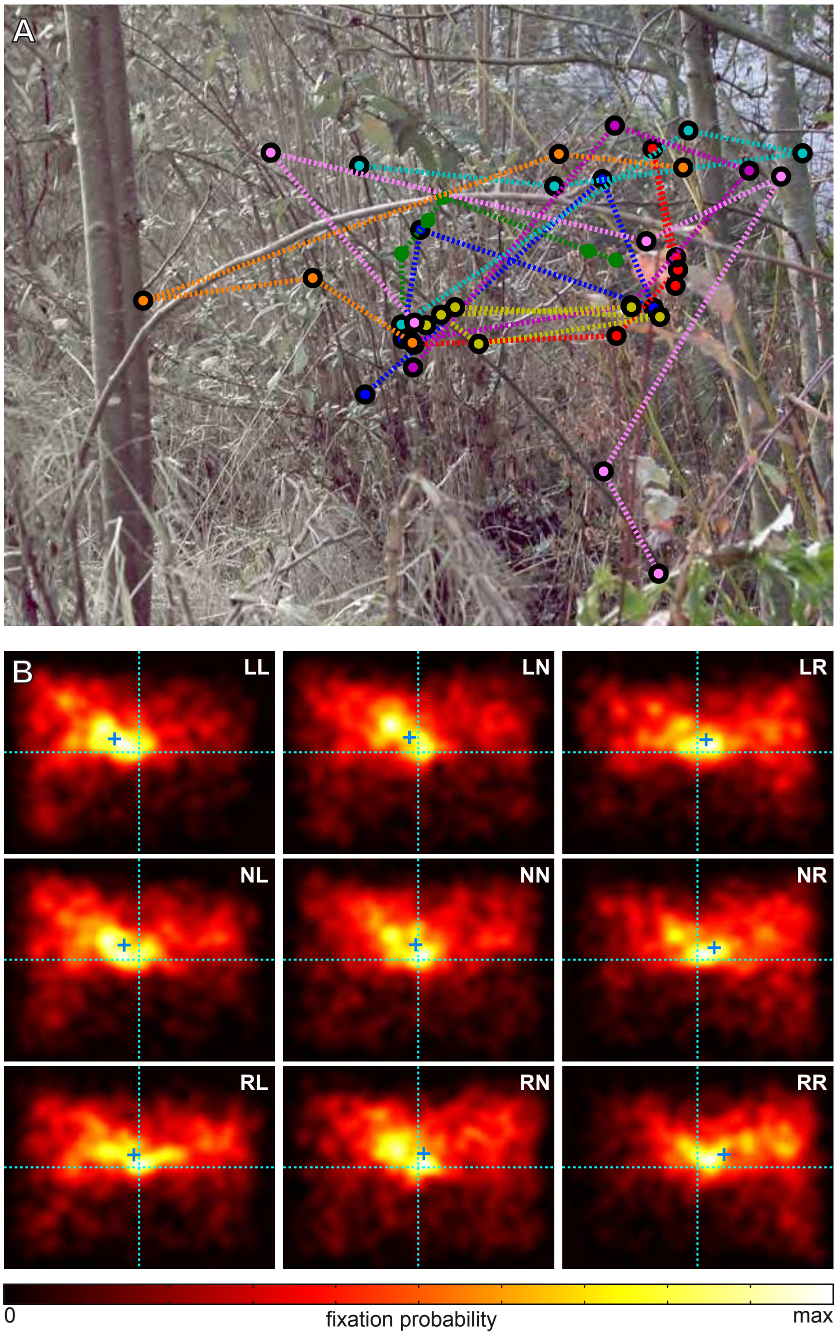

***Figure 2.2 Effect of modifications.***

*A) Example stimulus with color contrast increase to the right (RN), fixations of all observers superimposed, color identifies observer. B) Average fixations maps (spatial distribution of fixated location) for each condition, sorted as in figure 1C. For display, maps are smoothed with a 27 pixel (0.5 deg at the center) wide Gaussian kernel; extension of each map corresponds to the full image size, dashed lines indicate mid-lines, cyan crosses center-of-mass location.*

luminance- and color contrast (F(4,315)=0.18, p=0.95), between color contrast and fixation number (F(8,315)=0.38, p=0.93) or between luminance contrast and fixation number (F(8,315)=1.31, p=0.24). There was no three-way interaction between all three factors (F(16,315)=0.063, p=1.00). Hence we could analyze the effects separately.

### *Single-feature conditions*

### Color contrast

First we analyzed the effect of single-feature modifications, whether color- and luminance contrast gradients alone induce biases in fixated locations. To quantify this bias, we compared the average horizontal eye position at each fixation in condition RN to LN (figure 2.3A). Taking the conditions NN, RN and LN into account, there were main effects of color (F(2,105)=5.38, p=0.006) and of fixation number (F(4,105)=17.30, p<0.0001), but there was no interaction (F(8,105)=0.21, p=0.99). Post-hoc paired t-tests comparing the effects of gradients to the left (LN) and gradients to the right (RN) by individual showed a significant effect for all fixations tested (1st fixation: t(7)=3.24, p=0.01; 2nd: t(7)=6.61, p=0.0003; 3rd: t(7)=3.75, p=0.007; 4th: t(7)=4.68, p=0.002; 5th: t(7)=3.00, p=0.02). This demonstrated a robust and prolonged effect of the color contrast gradient on fixation location.

### Luminance contrast

For the gradients in luminance contrast we observed a similar pattern as for color contrast modifications (figure 2.3B). Considering conditions NL, NN and NR, there was a main effect of luminance contrast (F(2,105)=24.29, p<0.0001) and of fixation number (F(4,105)=9.04, p<0.0001), and no interaction (F(8,105)=0.50, p=0.86). Post-hoc paired t-tests showed a

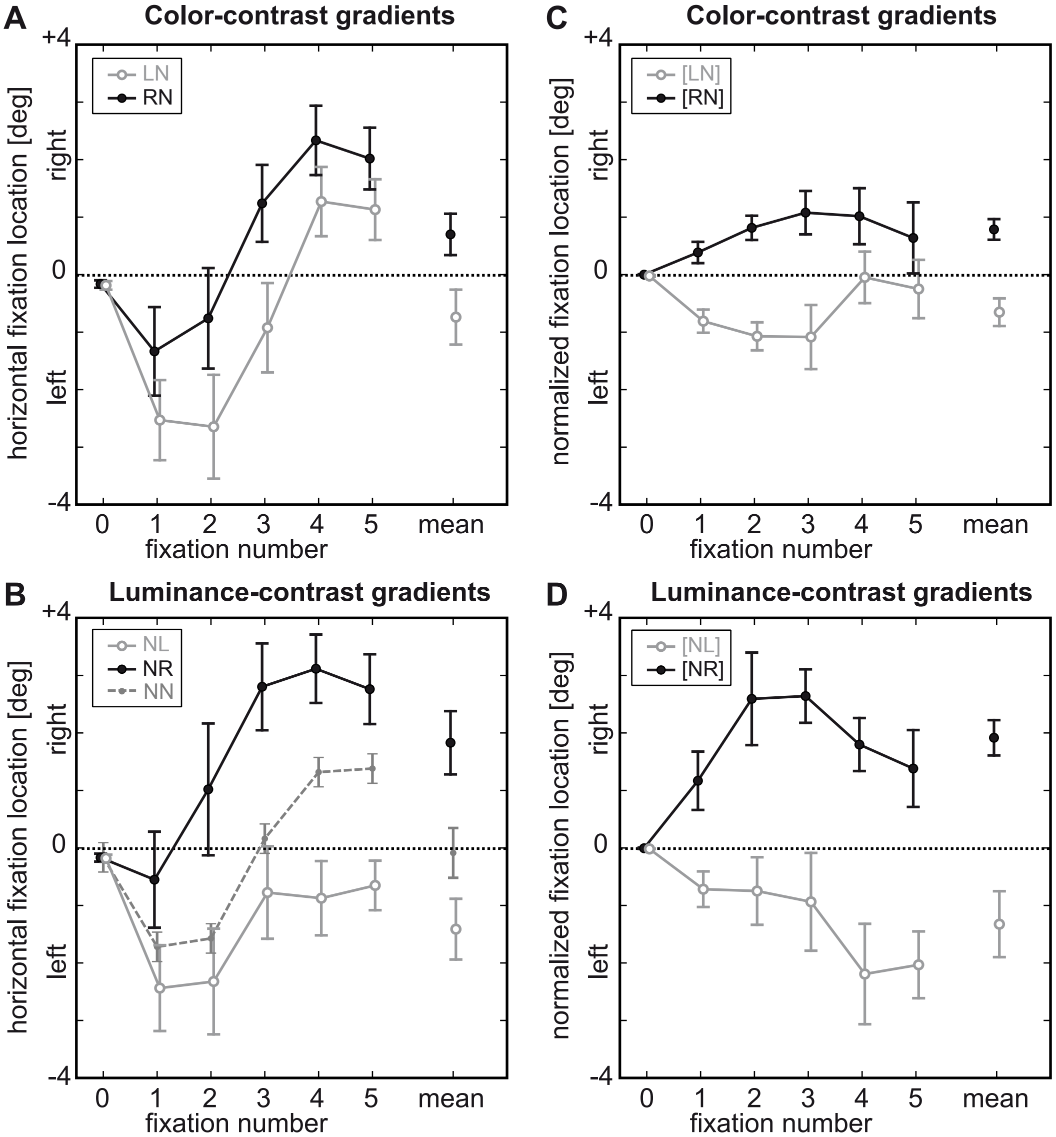

***Figure 2.3 Single-feature gradients.***

*A) Effect of color-contrast gradients. Mean +/- SEM over subjects of horizontal fixation location, positive values to the right, negative values to the left of screen center. Black: condition RN; gray: condition LN. The 0th (initial) fixation, which starts before stimulus onset is central by instruction, was not used for analysis. Note that all statistics are based on paired tests, while the standard errors of unnormalized locations include differences in general observer biases. Overlap in errorbars thus does not contradict a significant effect in paired tests. B) Effects of luminance-contrast gradients; black: NR; gray solid: NL; gray dashed: general bias without modification (NN) for comparison (omitted in panel A). C) Normalized effect of color-contrast gradients, black: [RN]=RN-NN; gray: [LN]=LN-NN. D) Normalized effect of luminance-contrast gradients, black: [NR]=NR-NN; gray: [NL]=NL-NN* significant effect starting at the first fixation (1st: t(7)=2.49, p=0.042; 2nd: t(7)=2.70, p=0.03; 3rd: t(7)=2.99, p=0.02; 4th: t(7)=4.38, p=0.003; 5th: t(7)=5.76, p=0.0007). This showed – consistent with our earlier results (Einhäuser, Rutishauser, et al., 2006) - that gradients in luminance contrast induce robust biases in fixated locations.

### Normalized analysis

The condition NN showed a modulation with fixation number (figure 2.3B). To measure the effects that gradients have on top of this general bias, we normalized horizontal fixation locations by subtracting the respective values of the NN condition. The normalized data showed the reported effects even more pronounced, both for color (figure 2.3C) and luminance-contrast (figure 2.3D). In all cases and for all fixations, single-feature gradients biased the condition in the direction of higher (color- or luminance-) contrasts relative to the general bias, which is revealed by condition NN.

### Average position

So far we had analyzed the data separated by fixation number. The mean positions exhibit the biases in the direction consistent with the gradient (rightmost data points in each panel of figure 2.3), whose significance had already been quantified by the aforementioned 2-way-ANOVA main effects of color and luminance-contrast, respectively. Other averaging schemes, e.g. weighing fixations with their duration or using all data including periods of saccades, yielded the same result.

In sum, the single-feature gradients induced robust biases, especially relative to a neutral (NN) condition, which held for the average eye-position but also for individual fixations.

### Dual-feature conditions

The dual-feature conditions examined the interaction between the effects of color contrast and luminance contrast. If the effects of luminance and color contrast added linearly, there are the following predictions as to how the effects of superimposed gradients can be computed from the single-feature gradients with a correction for the unmodified (NN) condition:

(1) LL ~ LN+NL-NN

(2) RR ~ RN+NR-NN

(3) RL ~ RN+NL-NN

(4) LR ~ LN+NR-NN

Linearity now predicts that the left-hand sides ("data") are statistically indistinguishable from the right-hand sides ("model"). The difference between model and data was tested by means of a two-sided paired t-test over observers; first considering the average effect over all images, but separated by fixation number. The right hand sides of all relations were indistinguishable from the respective left-hand sides for any of fixations 2 to 5 ($p_{min}$=0.25, table 1, gray shaded rows). Furthermore, for relations 1 and 4 this also held for the first fixation (p=0.81 and p=0.55, respectively). Hence the dual-feature data is consistent with linear summation of single-feature effects – both for congruent (figure 2.4A) and incongruent (figure 2.4B) gradients.

When considering the average fixation location rather than individual fixations (rightmost data point in each panel of figure 2.4), even the remaining deviations from a linear model vanished: for all conditions, the linear models' predictions were indistinguishable from the corresponding data ($p_{min}$=0.20, table 1 rightmost column). An alternative representation of these data again considered them subtractively normalized with respect to the unmodified condition. With the shorthand notation [.] for NN subtraction, relations (1) to (4) could be rewritten as

(1') [LL] ~ [LN]+[NL],

(2') [RR] ~ [RN]+[NR],

(3') [RL] ~ [RN]+[NL], and

(4') [LR] ~ [LN]+[NR],

respectively. Plotting the thus normalized data and model more evidently visualized their time-course over fixations relative to the general bias (figure 2.4C,D). Since the representations were mathematically equivalent (which is seen by subtracting NN from each side of the latter equations), the statistics (table 1) were unaffected.

To evaluate our statistical power, we tested the individual right-hand side summands as alternative models (white rows in table 1). For these controls, the average fixation location was always statistically different from the left-hand sides. This indicates that we had sufficient statistical power to find a deviation of model from data, if there were any. For the analysis of individual fixations, the results are less clear, especially in case of incongruent gradients. Nonetheless, with few exceptions, the linear model was in general more constistent with the dual-feature data than any individual single-feature effect even for individual fixations (table 1). Hence the compatibility between linear-summation model and dual-feature data cannot be attributed to a lack of statistical power.

### Alternative model: Max norm

So far we have merely argued that a linear model is *consistent* with our data and that we would have sufficient power to discriminate linearity from each feature alone. A maximum operation presents a frequently proposed alternative model. It predicts that the combined effect of two features corresponds to the larger individual effect. That is, the combined effect is predicted to have the magnitude of the larger individual effect and to also point in the direction of the effect with larger magnitude. Let a and b be the individual normalized effects (e.g., a=[LN], b=[NL]),

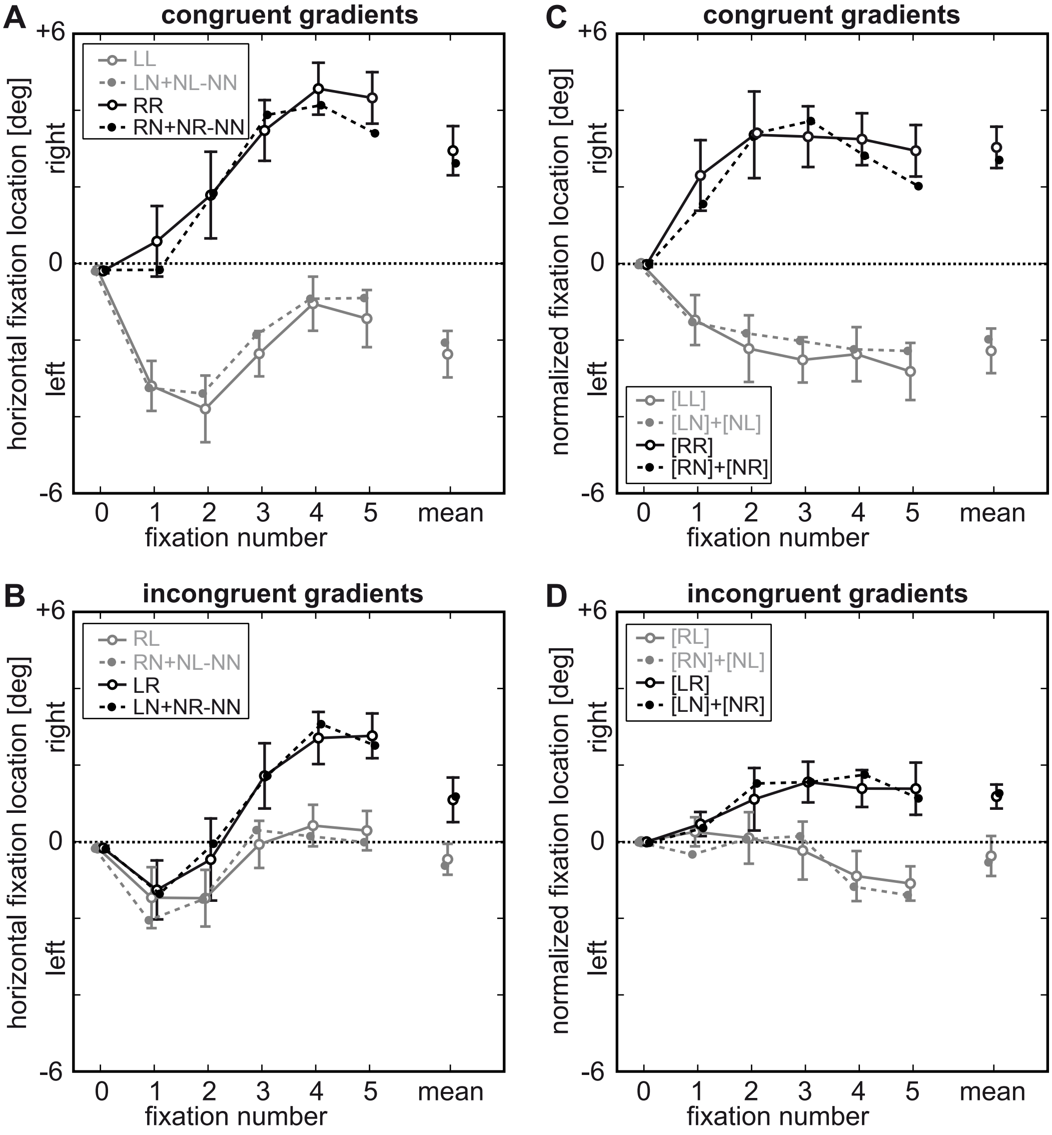

***Figure 2.4 Dual-feature gradients.***

*A) Mean +/- SEM for dual-feature conditions with same direction of gradients ("congruent gradients"); solid black: RR; solid gray: LL. Dashed lines denote predictions from single-feature trials, dashed black: RN+NR-NN; dashed gray: LN+NL-NN. B) Mean +/- SEM for dual-feature conditions with opposing directions of gradients ("incongruent gradients"); solid black: LR; solid gray: LR. Dashed lines denote predictions from single-feature trials, dashed black: LN+NR-NN; dashed gray: LN+NR-NN. C-D) Analogous to A and B, using normalized data instead.*

then the predicted combined normalized effect f (e.g., f prediction for [LL]) is given by

(5a) $f(a,b) = \max(|a|,|b|)\ \mathrm{sign}(a)$ if $|a|>|b|$

(5b) $f(a,b) = \max(|a|,|b|)\ \mathrm{sign}(b)$ if $|b|>|a|$

Where sign(x) = 1 f. x>0 and sign(x)=-1 f. x<0. For ease of notation, we denote f(a,b) as defined in equation (5) as smax(a,b) (for *s*igned *max*imum).

As for the linear model we tested whether the mean fixation location in the dual-feature gradients was distinguishable from this max-norm prediction using paired t-tests across observers:

(6) [LL] ~ smax([LN],[NL])

(7) [RR] ~ smax([RN],[NR])

(8) [RL] ~ smax([RN],[NL])

(9) [LR] ~ smax([LN],[NR])

In all but one case, we found significant differences between the model and the data ([LL]: t(7)=6.53, p=0.0003; [RR]: t(7)=2.71, p=0.03; [RL]: t(7)=1.36, p=0.22; [RL]: t(7)=6.56, p=0.0003). In conclusion, whereas the linear combination of single-feature effects was indistinguishable from the dual-feature data in all four cases, the maximum norm was significantly different in 3 out of 4. This confirmed not only that our statistical power sufficed to exclude alternative models, but also clearly demonstrated that a linear addition of single feature effects better explained the data than a max-norm model.

### *Image-by-image results*

Up to here, we have considered aggregate data across images. This was motivated by the fact that the images primarily serve as “background” and image structure by itself probably had a substantial effect on fixation allocation. To quantify this, we analyzed data averaged over subjects and fixations for each image individually. For the congruent gradients 85/90 (LL) and

| Data | Model | 1st fixation | 2nd fixation | 3rd fixation | 4th fixation | 5th fixation | Mean |
|---|---|---|---|---|---|---|---|
| [LL] | [LN]+[NL] | ***0.81*** | ***0.27*** | ***0.53*** | ***0.85*** | ***0.51*** | ***0.20*** |
|  | [LN] | ***0.07*** | ***0.08*** | 0.04 | 0.003 | 0.006 | 0.004 |
|  | [NL] | 0.005 | 0.002 | 0.007 | ***0.66*** | ***0.25*** | 0.0003 |
| [RR] | [RN]+[NR] | 0.006 | ***0.94*** | ***0.57*** | ***0.56*** | ***0.27*** | ***0.46*** |
|  | [RN] | 0.01 | 0.009 | 0.02 | 0.01 | 0.004 | 0.006 |
|  | [NR] | 0.0007 | ***0.16*** | ***0.15*** | 0.02 | 0.046 | 0.02 |
| [RL] | [RN]+[NL] | 0.005 | ***0.94*** | ***0.55*** | ***0.70*** | ***0.71*** | ***0.48*** |
|  | [RN] | **0.75** | **0.41** | **0.16** | 0.001 | 0.007 | 0.03 |
|  | [NL] | 0.005 | **0.055** | **0.056** | 0.001 | 0.043 | 0.001 |
| [LR] | [LN]+[NR] | ***0.55*** | ***0.25*** | ***0.9997*** | ***0.51*** | ***0.71*** | ***0.74*** |
|  | [LN] | 0.02 | 0.02 | 0.01 | 0.02 | 0.002 | 0.005 |
|  | [NR] | 0.04 | 0.0004 | 0.02 | ***0.23*** | ***0.994*** | 0.0003 |

***Table 1.***

*Analysis as to how well the linear "model" from the single feature conditions deviates from the "data" obtained in the respective dual feature condition. Each entry denotes the p-value of a paired t-test. Linearity predicts that there is no evidence for differences of data from model in the gray shaded rows. Significant effects in the other rows ("control models") show that we would have sufficient power to recognize a deviation if it would occur. Note that there two equivalent ways of testing, using either the normalized or the raw positions. For example, the distance between [LL] and [LN]+[NL] is equivalent to the distance between LL and LN+NL-NN, as is directly seen by adding NN on each side of the latter relation.*

89/90 (RR) images showed a bias to the left and right (relative to NN), respectively. For the single-feature conditions, this bias to the higher contrast side was slightly less pronounced (LN: 60/90, RN: 60/90, NL:75/90, NR: 82/90), but the fraction was still significantly above chance (all $p < 0.003$, sign-tests). Consequently, the biases were robust across images. Finally, we tested the prediction of the two models (linear addition and max-norm) on these image-wise data. In all cases, the linear model was indistinguishable from the data ([LL]: t(89)=1.65, p=0.10; [RR]: t(89)=1.02, p=0.31; [RL]: t(89)=0.77, p=0.45; [LR]: t(89)=0.89, p=0.38), while the max-norm model showed significant differences for the congruent dual-feature gradients ([LL]: t(89)=6.08, $p=3\times10^{-8}$; [RR]: t(89)=5.70, $p=1\times10^{-7}$; [RL]: t(89)=0.05, p=0.96; [LR]: t(89)=0.67, p=0.51). Hence the image-by-image analysis confirmed the main finding: dual-feature effects were consistent with a linear addition of single-feature effects in all conditions. In contrast, a max-norm was only consistent with the data when individual effects were too small to clearly distinguish between the models. Our results therefore provided clear evidence that the interaction of color and luminance contrast on a natural-scene background is more consistent with linear summation than with a maximum operation.

## Discussion

The present study investigates human overt attention on natural-scene background. We demonstrate that luminance- and color-contrast gradients that are superimposed over a scene affect the selection of fixation points: fixations are biased towards regions of high contrasts. Most notably the combined effects of luminance- and color-contrast gradients are consistent with a linear summation of feature effects, but not with a maximum operation.

The effects of gradients operate on top of a general bias in viewing direction when inspecting unmodified stimuli (condition NN), which starts to the left and then rebounds to the right of the midline. Although this is not the aim of the present study, it might be interesting to speculate whether this bias reflects a general strategy, possibly related to reading direction, as observed for other attentional phenomena, such as inhibition-of-return (Spalek & Hammad, 2005).

In order to encourage subjects to pay attention to the stimuli, we asked them only to “study the images carefully”. We had used this instruction in earlier studies and expect it to bias fixation allocation in a “bottom-up” driven mode, and to operate in sharp contrast to explicit top-down tasks, such as search (Henderson, Brockmole, Castelhano, & Mack, 2007; Einhäuser, Rutishauser & Koch, 2008). A recent experiment (Steinwender & König, unpublished observations) indeed shows that “study carefully” yields the same result with respect to low-level features as the explicit instruction of “free-viewing”, whereas for example “subjective assessment” yields distinct fixation behavior. Although, we cannot exclude that the size of the effects for single-feature conditions depends on the particular choice of instruction, we clearly see a bottom-up (i.e., feature-driven) component. In the present context, we build on this observation of a systematic shift of fixation locations induced by single-feature gradients. The prediction of linear interaction of different features is tested by comparing these measured shifts to those measured in dual-feature conditions. This test is therefore independent of the size of single-feature effects, as long as they are different from 0 and sufficiently small to avoid the image-boundaries to come into play for the combined effects. In particular, it does not depend on whether or not the effect of LC and/or CC modification itself is linear in gradient strength, although we observed linearity at least for LC earlier (Einhäuser, Rutishauser, et al., 2006). Consequently, as long as the instruction allows for shifts in the single-feature conditions that are sufficiently robust for the comparison to dual-feature effects, their precise size is not critical, nor is the exact choice of instruction.

Since Koch and Ullman’s (1985) original proposal, the saliency map model has repeatedly been used to predict fixation behavior in natural scenes (Itti & Koch, 2000; Parkhurst et al., 2002; Peters et al., 2005; Tatler et al., 2005). In all these studies, however, prediction remains well below the optimum for any bottom-up model: the optimal prediction a bottom-up model (i.e., a model taking into account only the current stimulus’ features) can be expected to achieve is the level of mutual inter-observer prediction (Peters et al., 2005). Furthermore, the reasonable success of predictions on the system level neither implies causality nor does it provide support for the model’s mechanistic assumptions. This raises the question of the extent to which individual features are indeed correlated to overt attention. With respect to luminance-contrast, various studies (Reinagel & Zador, 1999; Krieger et al., 2000) have found this feature to be elevated at fixation points. Depending on presentation conditions, however, the correlative effect of luminance-contrast is only observed after correcting for general biases in fixation pattern, depends on spatial frequency (Mannan, Ruddock & Wooding, 1996, 1997; Einhäuser & König, 2003; Tatler et al., 2005) and its size depends on the image material used (Privitera & Stark, 2000; Parkhurst et al., 2002). In addition, the effect of luminance-contrast is often small compared to other luminance-related features, such as "edge density" (Mannan et al., 1996), texture contrast (Einhäuser & König, 2003; Parkhurst & Niebur, 2004), higher order geometric kernels (Priviterra, Fujita, Chernyak & Stark, 2005) and image-category specific features (Privitera & Stark, 2000). With respect to the relative effects of the features under investigation here, Tatler et al. (2005) find luminance-contrast and "edge-content" to contribute consistently more strongly to human fixation than "chromaticity" and luminance itself. Since measuring the additivity of features is the main aim of the present study, our single-features had to fulfill two conditions: they had to be sufficiently large and robust to allow statistical analysis of their interaction (in the limit of no effect, all summation schemes are equally valid), but to be sufficiently small that image boundaries do not artificially cut the dual-feature effect. Therefore we chose gradients that induce a robust effect for single-feature conditions. The fact that at least the effect of luminance gradients is linear in gradient slope (Einhäuser, Rutishauser et al., 2006) renders it likely that our results on linearity can be generalized to weaker contrast changes, as found in natural contrast variations.

Most of the aforementioned studies measure the influence of each feature in its natural context. This, however, does not allow the isolation of the effects of each feature. If a feature were correlated with higher-order structure in natural scenes, increased fixation probability might result from higher-order structure or from correlation to other features, rather than from the feature itself (Baddeley & Tatler, 2006). To overcome this confound, Einhäuser and König (2003) locally increase or decrease luminance-contrast in natural scenes. They find that the effect of reduced local contrast attracts human attention, and conclude that this is inconsistent with saliency-map model predictions. Although Parkhurst & Niebur (2004) reconcile this particular finding with saliency map models by incorporating higher order contrasts, local modifications are suboptimal in the present experimental context.

Strong local modifications introduce local deviations from global context, which are likely to attract attention. This is most evidently seen in the phenomenon of pop-out (Treisman & Gelade, 1980) and has recently entered the saliency map literature as the notion of "surprise", an information-theoretic measure of deviations from the temporal context (Itti & Baldi, 2005). This issue of local deviations from context becomes especially prominent when the applied modifications extend beyond the naturally occurring range of the feature. To avoid this potential confound in analyzing the interaction between features, we use large-scale gradients rather than local modifications. This procedure neither introduces local deviations, nor does it modify higher-order contrasts *locally*.

Obviously, the contrast gradient does not leave higher-order structures unaffected, e.g. reducing contrast will also reduce edge density (if there is zero contrast, there are also no edges) and affect texture contrast. In any case, as we compare the effects of LC and CC in isolation to their combined effects, correlations to higher order structure *within a feature channel* would not confound our findings. One needs to ensure, however, that modifying LC does not affect CC and vice versa. By using definitions of LC and CC that are orthogonal in DKL-space this requirement is fulfilled, although it is conceivable that *perceived* LC varies with CC and vice versa. Although our gradients may affect higher order structure to some extent, their large scale, as well as the physical independence of the modified features, mean that the linearity of LC and CC effects is also likely to hold in the natural context.

The rationale for using natural scenes quasi as a "background" for the observed effects is two-fold. First, the effect of the gradients is independent as to whether the scene is perceived as natural, at least as long as the amplitude spectrum is conserved (Einhäuser, Rutishauser, et al., 2006). Second, if we would use a noise background instead, it could be argued that the interaction would be different if objects distract from the superimposed low-level effects. Hence observing a linear interaction of color and luminance contrast on – or maybe despite – the natural scene background, strengthens our argument. Our data do, however, not address the issue of whether or not feature biases that are inherent in a scene affect fixated locations. Tatler (2007) has argued that those biases do not influence fixation. Similarly, our data are agnostic with respect as to whether features like color and luminance naturally occurring in natural scenes drive attention causally, and thus do not contradict the large body of recent work that fails to find a causal effect under realistic conditions (see below).

Any model of attention that incorporates different features needs a mechanism to appropriately combine those features. Contemporary implementations of saliency maps usually solve this issue by using a sophisticated normalization scheme to achieve comparable saliency measures for each individual feature (see Itti & Koch (2000) for a thorough discussion of normalization schemes). Subsequently these models linearly combine the resulting “conspicuity maps” into the final saliency map. Here we directly measure the individual effects (“conspicuities”) of each feature by using single-feature conditions and then test whether linearity between these effects holds. We find that color- and luminance-contrast interact linearly. Using a model based on psychophysical and physiological data, Li (2002) proposed that the saliency of an item is given by “the saliency of its most salient component”. This implies a maximum operation. Lewis and Zhaoping (2005) suggested that this model might also be applicable to the interaction of color and orientation in human overt attention for natural scenes. Provided the different features and different methodology, our data does not contradict these findings directly. Instead, it will be an interesting issue for future research, whether our results can be extended to other features, such as color and orientation, on a natural scene background. For the case of color and luminance-contrast, however, our data clearly falsifies the max-norm hypothesis.

Since the saliency map model was originally designed as a purely bottom-up model of attention, by construction it does not capture top-down influences such as the observer's experience or the task. The task plays a decisive role for human overt attention in inspecting pictures (Buswell, 1935; Yarbus, 1967) or search displays (Bacon & Egeth, 1997) or in everyday activities (Land & Hayhoe, 2001). When memorizing objects, for example, observers tend to replicate their own scan-paths, a feature not adequately captured by bottom-up saliency alone (Foulsham & Underwood, 2008).

Visual search constitutes a task frequently used to quantify the performance of attention models. Predictive performance of the original bottom-up saliency map model reduces or vanishes in search tasks (Einhäuser, Rutishauser & Koch, 2008; Henderson, Brockmole, Castelhano, & Mack, 2007), but inclusion of contextual or task-dependent information can improve saliency-map algorithms (Navalpakkam & Itti, 2005; Oliva, Torralba, Castelhano & Henderson, 2003; Torralba, 2003). For evaluating the performance of saliency-map-type models in predicting search in natural scenes, the intuitive strategy to fixate the point of highest saliency is usually suboptimal; instead the discriminability between target and distractor on the basis of the full map should be utilized (Vincent, Troscianko & Gilchrist, 2007; Gao, Mahadevan, & Vasconcelos, 2008). For specific search tasks, such as searching a pedestrian in a street scene, the task-modulated prior alone may predict search patterns better than bottom-up signals (Torralba, Oliva, Castelhano & Henderson, 2006). This approach, however, requires the prior distribution of potential target locations to be non-uniform and known. Such knowledge may be learned from scene statistics, and joint learning of bottom-up and top-down saliency in a

Bayesian framework seems a promising approach (Zhang, Tong, Marks, Shan, & Cottrell, 2008).

Visual search models often use the selective up-regulation of target features (Wolfe, Cave, & Franzel, 1989; Pomplun, 2006), of the corresponding visual filters (Rao, Zelinsky, Hayhoe, & Ballard, 2002) or statistical knowledge of target location (Najemnik & Geisler, 2005) to predict human performance. Rao et al.'s (2002) model bears some similarity to Itti & Koch's (2000) saliency map, but instead of adding different feature maps linearly, it computes a single map, which is modulated based on the distance to the target template, rather than treating features individually. As Navalpakkam & Itti (2005) have pointed out, this approach predicts that search for targets differing in one feature (pop-out) should be as efficient as conjunction search, contrary to experimental evidence (Treisman & Gelade, 1980). While not contesting the approach of Rao et al. (2002) per se, this argues in favor of different feature channels that need to be appropriately integrated.

We are well aware that the seeming mechanistic implications of the saliency map model have to be interpreted with care. In fact, we consider it likely that its predictive power for fixations stems entirely from correlations of its constituents to higher-order structure inherent in natural scenes, such as "interesting objects" (Elazary & Itti, 2008; Einhäuser, Spain & Perona, 2008). Furthermore, we are just beginning to understand how context and top-down information can be integrated in computational models of attention. Nevertheless, the original, purely bottom-up model is widely used and – up to now – other models that reach similar correlations to fixation probability (under the constraints of free-viewing, laboratory setup, etc.) are rare. Independent of the precise model and its prediction on a systems' level, a sound understanding of human attention on a mechanistic level, will always require a rigorous test of its assumptions. Irrespective of the exact nature of a future model that finally supersedes the saliency map for scene prediction, it will be constrained by the present finding: effects of color and luminance –

under laboratory conditions and on a natural scene background – add linearly. The extent to which this finding transfers to other low-level features and to spatial distributions of higher-order scene structures thus remains an exciting issue for future research, no matter one's hold on the original saliency map.

### Acknowledgements

The work was financially supported by the Deutsche Forschungsgemeinschaft Research Training Group 885 – NeuroAct (BMtH, WE), the German Academic Exchange Service (SE), and by IST-027268-POP (Perception On Purpose, PK). The authors are grateful to K. Libertus and H.-P. Frey for technical assistance, and to C. Quigley for editorial assistance.

## *Appendix A: Effects of modifications on luminance-contrast, saliency, and color conspicuity*

Here we address the relation of our proposed modifications to common definitions of contrast, feature conspicuity, and saliency. There are plenty of possible ways to define LC (Peli, 1997). Most definitions are originally based on the comparison of a single foreground intensity with a single background intensity, such as Weber contrast (the difference of foreground and background divided by the background) or Michelson contrast (the difference of foreground and background divided by their sum; Michelson, 1927) and have been extended to account for arbitrary stimuli. Amongst the possible variants, we here focus on those that are of common use in the context of eye-tracking and attention studies:

(1) The standard-deviation of luminance in a local patch divided by the image mean (e.g., Reinagel & Zador, 1999)

(2) The standard-deviation of intensity in a local patch divided by the patch mean (also suggested – and dismissed as suboptimal in the present context – by Reinagel & Zador, 1999)

(3) The difference of maximum and minimum in a local patch divided by their sum in the same local patch. This is a definition most closely related to Michelson contrast. Note that Mannan et al.'s (1996) usage of the mean of intensity in a local patch as foreground and the mean of the image as background in their calculation of Michelson contrast is a measure of luminance rather than of luminance-contrast in the present context. The denominator is also commonly scaled or replaced by the mean (akin to Weber contrast) or by the maximum alone.

We compute all these contrasts for the luminance channel of our images in DKL space, which we shift and scale to range from 0 to 1 (rather than from -1 to 1), and for squared patches of width 24 pixels (corresponding to 0.5° at the screen center). Comparing the conditions in which luminance-contrast increases to left, right or remains unmodified for a single image (the one of figure 2.2A) and averaging over rows, we see the intended effect of modification clearly, and the differences between the various contrast definitions are minute (figure 2.5A to C). Note, that – by definition – the luminance-contrast profile is not affected by modifications to the color-contrast, e.g., the conditions LL, NL and NR have the same luminance-contrast profile. The example profiles show that highly noticeable structures, such as the tree on the left-hand side of the example image, are still visible, although the modification dominates this contrast profile.

To quantify how "unnatural" the contrast modifications were, we assessed the additional variation of contrast introduced by the gradients, using the contrast-definition (1) of above. In the unmodified condition, the mean contrast within an image amounts to 0.61±0.13 (mean±sd across images). As one would expect by construction, this value is about halved for modifications, no matter whether the increase is to the left (LL,NL,RL) or to the right (LR,NR,RR) with values of 0.30±0.07 in both cases. It should be noted, however, that a *large-scale* single-feature modification of contrast always biases towards the higher contrast side, no matter if the gradient decreases or increases the contrast relative to unmodified (Einhäuser, Rutishauser et al., 2006), which is different from local modifications (Einhäuser & König, 2003). More importantly, the gradients lower the variation of contrast within each image, quantified as standard-deviation of contrast values, only by about 25% (0.33±0.10 for unmodified, for 0.25±0.06 for gradients to the right and 0.25±0.07 for gradients to the left). This indicates that a sufficient amount of image-inherent variability remains in the low-level features, which could – in principle – drive attention. The fact that the gradient nonetheless dominates the fixation allocation is consistent with a minute (or absent) effect of image-inherent low-level features.

Next we consider the effect of our LC modifications on the Itti & Koch (2000) model for

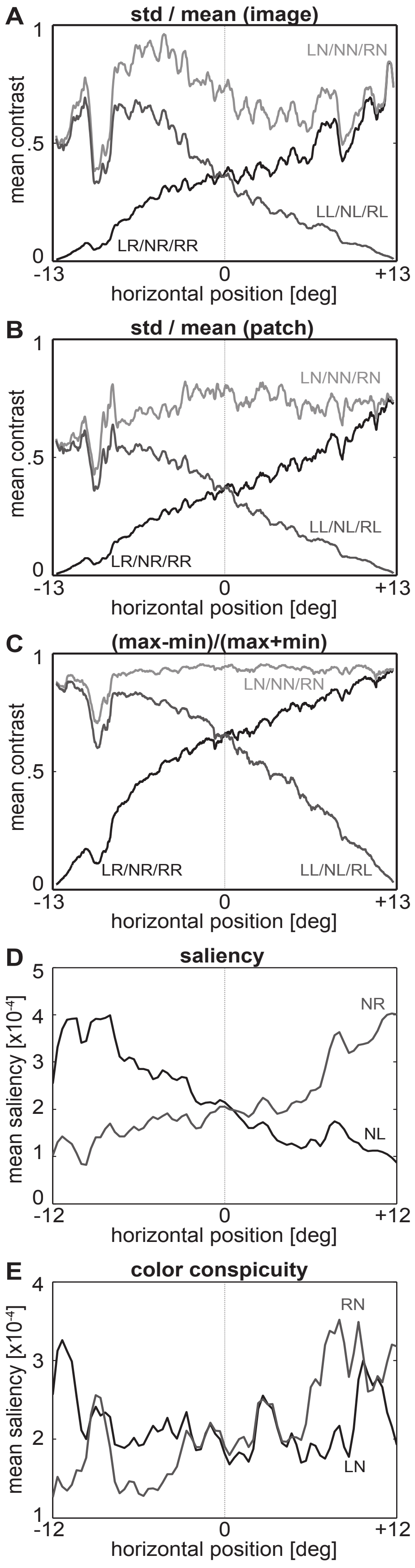

***Figure 2.5 Effect of gradients on common definitions of contrast, conspicuity and saliency.***

*A-C) Effect of luminance-contrast gradient on different definitions of LC along horizontal scanline for the example image of figure 2A, averaged over image rows. A) standard-deviation of luminance in a 1°x1° patch divided by image mean, B) standard-deviation of luminance in a patch divided by patch mean, C) difference between maximum and minimum luminance in a patch divided by their sum. D) saliency according to Itti & Koch (2000), maps linearly normalized to unit integral E) color conspicuity according to Itti & Koch (2000), maps linearly normalized to unit integral. Note that maps in panels D and E have a lower resolution and additional cut-off at the image boundary.*

visual saliency. For the model we use the implementation provided at http://ilab.usc.edu with no normalization, but otherwise default parameter settings. To be closer to the typical scenario for the application of these algorithms, we here use the image in the RGB version sent to the screen rather than the original DKL-definition, i.e. luminance is non-linearly scaled. By performing the same analysis as for the contrast definitions, we find that our LC modifications strongly modulate model saliency in the expected direction: saliency increased to the right in the NR condition and increases to the left in the NL condition (figure 2.4D).

Finally, we address the effect of our color modification on the color channel of the saliency map model with the same settings as above. As expected we find color conspicuity to increase to the right in the RN condition and to the left in the LN condition (figure 2.5E). The original image structure, however, is more conserved than in the luminance case, and the luminance-conspicuity dominates the overall saliency map with default weighing (not shown). That is, the effect of color modification on image structure is weaker, consistent with the slightly weaker bias induced by color. While this bias-difference is worth investigating, in particular with respect to feature-weighing schemes for saliency maps, it is not of relevance for the present paper. When both gradients induce a robust fixation bias (figure 2.3), we can compare their effects (figure 2.4). In this appendix we verified that the modifications leading to these effects are indeed consistent with common definitions of luminance-contrast and color conspicuity.

### *Appendix B: Raw eye position*

In order to be independent of the fixation definition, we repeated our main analysis, using 1-ms bins instead of individual fixations. Since subsequent time-points fell on the same fixation and were thus not independent, we cannot perform the equivalent statistical analysis. Instead we used paired t-tests to test for the significance of difference, but adjusted the alpha level to match an expected false discovery rate (FDR) of 5%, using the procedure proposed by Benjamini and Hochberg (1995). A result was called significant if it fell below this adjusted level (denoted as$FDR_{0.05}$). For color-only gradients, we found a significant difference between LN and RN ($p < 0.019 = FDR_{0.05}$) on 773 sample points between 364 ms and 1198 ms. Similarly, for luminance-contrast-only gradients, there was a significant difference between NL and NR ($p <0.036 = FDR_{0.05}$) on 1435 sample points between 117 ms and 2000 ms. This confirmed that during the majority of the presentation time gradients affect eye position. At an expected false-discovery rate of 0.05, LL was at no time point different from LN+NL–NN (figure 2.6A, gray). Subtracting the NN condition on both sides by construction did not alter the results, i.e. [LL] was not different from [LN]+[NL] (figure 2.6C). Neither was [RR] different from [RN]+[NR] anywhere (figure 6A,C black). Similarly, the incongruent gradient data did not exhibit significant differences from their respective models at any time point, [LR] was indistinguishable from [NR]+[LN] (figure 6B,D black) and [RL] from [NL]+[RN] (figure 6B,D, gray). In sum, the analysis of the raw eye position data confirmed the fixation analysis, ruling out that the observed effects depend on the definition or timing of fixations.

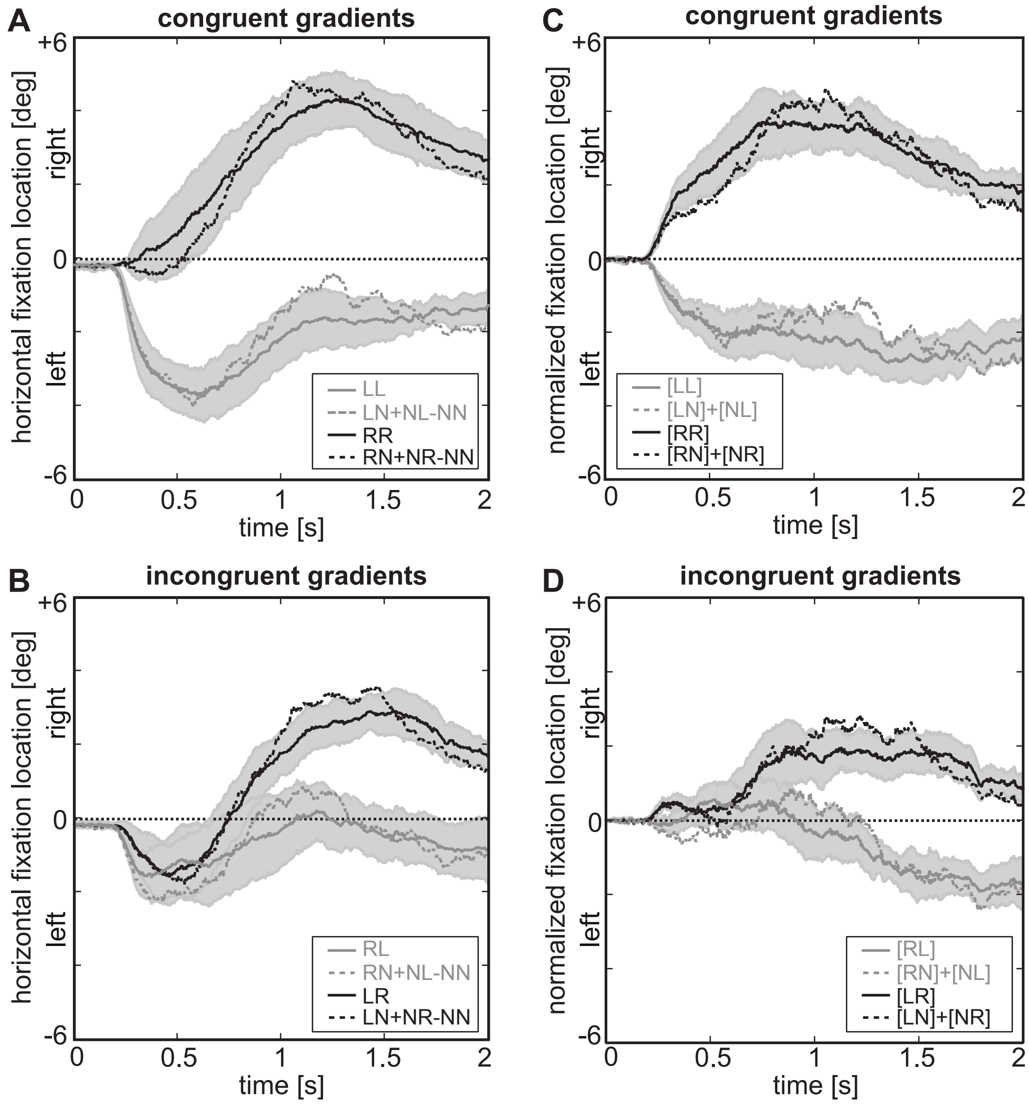

***Figure 2.6 Analysis over time.***

*Analogous to figure 2.4, using time into trial rather than fixation number as parameter, shaded areas denote SEM of data.*

## References

Bacon, W. J. & Egeth H. E. (1997). Goal-directed guidance of attention: evidence from conjunctive visual search. *Journal of Experimental Psychology: Human Perception and Performance,* **23(4)**, 948-961.

Baddeley, R.J., & Tatler, B.W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research*, **46(18)**, 2824-2833.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* **57**, 289-300.

Brainard, D.H. (1997). The Psychophysics Toolbox. *Spatial Vision,* **10**, 433-436.

Buswell, G. T. (1935). *How people look at pictures. A study of the psychology of perception in art*. Chicago, Illinois: The University of Chicago Press.

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing (NIPS),* **20**, 241-248.

Cornelissen, F.W., Peters, E.M., & Palmer, J. (2002). The Eyelink Toolbox: eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers,* **34**, 613-7.

Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *Journal of Physiology*, **357**, 241-265.

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience,* **17(5)**, 1089-1097.

Einhäuser, W., Kruse, W., Hoffmann, K.-P. & König P. (2006). Differences of Monkey and Human Overt Attention under Natural Conditions. *Vision Research,* **46**, 1194-1209.

Einhäuser, W., Rutishauser, U., Frady, E.P., Nadler, S., König, P, & Koch, C. (2006). The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. *Journal of Vision*, *6*(11):1, 1148-1158.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, **8(2):2**, 1-19.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14):18, 1-26.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, **8(3):3**, 1-15.

Foulsham, T. & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, ***8*(2)**:6, 1-17.

Gao, D., Mahadevan, V., & Vasconcelos, N. (2008). On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, ***8*(7):**13, 1-18.

Golz, J., & MacLeod, D. I. (2002). Influence of scene statistics on colour constancy. *Nature,* **415(6872)**, 637-640.

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature,* **391**, 481-484.

Henderson, J.M., Brockmole, J.R., Castelhano, M.S., & Mack, M. (2007). Visual Saliency does not account for Eye-Movements during Visual Search in Real-World Scenes. In R. van Gompel, M. Fischer, W. Murray, R. Hill (Eds.), *Eye Movement Research: Insights into Mind and Brain.* Amsterdam: Elsevier.

Horwitz, G. D., & Newsome, W. T. (1999). Separate Signals for Target Selection and Movement Specification in the Superior Colliculus. *Science,* **284**, 1158-1161.

Itti, L. (2005). Quantifying the Contribution of Low-Level Saliency to Human Eye Movements in Dynamic Scenes. *Visual Cognition,* **12(6)**, 1093-1123.

Itti, L., & Baldi, P. (2005). A Principled Approach to Detecting Surprising Events in Video. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research ,* **40**, 1489-1506.

James, W. (1890). *Principles of Psychology*. New York: Holt.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology,* **4**, 219-227.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision,* **13**, 201-214.

Kustov, A. A., & Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. *Nature,* **384**, 74-77.

Land, M. F. & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research,* **41(25-26)**, 3559-3565.

Lewis, A., & Zhaoping, L. (2005). Saliency from Natural Scene Statistics. *Abstract Viewer/Itinerary planner. Washington DC: Society for Neuroscience.* Program No. 821.11.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Science,* **6**, 9-16.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision,* **10(3)**, 165-188.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997). Fixation patterns made during brief examination of two-dimensional images. *Perception,* **26(8)**, 1059-1072.

Mazer, J. A., & Gallant, J. L. (2003). Goal-related activity in V4 during free viewing visual search. Evidence for a ventral stream visual salience map. *Neuron,* **40(6)**, 1241-1250.

McPeek, R. M., & Keller, E. L. (2002). Superior Colliculus Activity related to Concurrent Processing of Saccade Goals in a Visual Search Task. *Journal of Neurophysiology,* **87**, 1805-1815.

Michelson, A. A. (1927). *Studies In Optics*. Chicago: University of Chicago Press.

Morrone, M. C., Denti, V., & Spinelli, D. (2002). Color- and luminance-contrasts attract independent attention. *Current Biology,* **12(13)**, 1134-1137.

Najemnik, J. & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature,* **434(7031)**, 387-391.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research,* **45(2)**, 205-231.

Nothdurft, H. (2000). Salience from feature contrast: additivity across dimensions. *Vision Research,* **40(10-12)**, 1183-1201.

Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003). Top-Down Control of Visual Attention in Object Detection. *IEEE Proceedings of the International Conference on Image Processing,* **1**, 253-256.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research,* **42**, 107-123.

Parkhurst, D., & Niebur, E. (2004). Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience,* **19(3)**, 783-789.

Peli, E. (1997). In search of a contrast metric: matching the percieved contrast of Gabor patches at different phases and bandwidths. *Vision Research,* **37(23)**, 3217-3224.

Pelli, D.G. (1997) The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision* **10**, 437-442.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research,* **45**, 2397-2416.

Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research,* **46(12)**, 1886-1900.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience,* **13**, 25-42.

Privitera, C. M., & Stark, L. W. (2000). Algorithms for Defining Visual Regions-of-Interest: Comparision with Eye Fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **22(9)**, 970-982.

Privitera, C. M., Fujita, T., Chernyak D., & Stark, L. W. (2005). On the discriminability of hROIs, human visually selected regions-of-interest. *Biological Cybernetics,* **93(2)**, 141-152.

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M. & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research,* **42(11)**, 1447-1463.

Reinagel, P., & Zador, A. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems,* **10**, 341-350.

Rizzolatti, G., Raggio, L., Dascola, I., & Umilta, C. (1987). Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia,* **25**, 31-40.

Robinson, D. L., & Petersen, S. E. (1992). The pulvinar and visual salience. *Trends in Neuroscience,* **15**, 127-132.

Spalek, T.M., & Hammad, S. (2005) The left-to-right bias in inhibition of return is due to the direction of reading. *Psychological Science,* **16(1)**:15-18.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research,* **45**, 643-659.

Tatler, B. W. (2007). The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision,* 7(14), 4, 1-17.

Thompson, K. G., Bichot, N. P., & Schall, J. D. (1997). Dissociation of visual discrimination from saccade programming in macaque frontal eye field. *Journal of Neurophysiology,* **77**, 1046-1050.

Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America A - Optics, Image Science and Vision,* **20(7)**, 1407-1418.

Torralba A., Oliva, A., Castelhano, M.S., & Henderson, J.M. (2006) Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review.* 13(4):766-786.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology,* **12**, 97-136.

Vincent, B. T., Troscianko, T., & Gilchrist, I. D. (2007). Investigating a space-variant weighted salience account of visual selection. *Vision Research,* **47(13)**, 1809-1820.

Wolfe, J. M., Butcher, S. J., Lee, C., & Hyle, M. (2003). Changing Your Mind: On the Contributions of Top-Down and Bottom-Up Guidance in Visual Search for Feature Singletons. *Journal of Experimental Psychology: Human Perception and Performance,* **29(2)**, 483-502.

Wolfe, J. M., Cave, K. R. & Franzel, S. L. (1989). Guided search: an alternative to the feature integration model of visual search. *Journal of Experimental Psychology: Human Perception and Performance,* **15(3)**, 419-433.

Yarbus, A. L. (1967). *Eye movements and vision.* Haigh, B. (trans.). New York: Plenum Press.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7):32, 1-20.

## *Study II*

*Gaze allocation in natural stimuli: comparing free exploration to head-fixed viewing conditions*

Published as:

’t Hart, B.M., Vockeroth, J., Schumann, F., Bartl, K., Schneider, E., König, P. and Einhäuser, W. (2009). Gaze allocation in natural stimuli: comparing free exploration to head-fixed viewing conditions. Vis Cog, 17(6+7), 1132-1158.

# Gaze allocation in natural stimuli: comparing free exploration to head-fixed viewing conditions

### Abstract

"Natural" gaze is typically measured by tracking eye positions during scene presentation in laboratory settings. How informative are such investigations for real-world conditions? Using a mobile eye-tracking setup ("EyeSeeCam"), we measure gaze during free exploration of various in- and outdoor environments, while simultaneously recording head-centered videos. Here, we replay these videos in a laboratory setup. Half of the laboratory observers view the movies continuously, half as sequences of static 1-second frames. We find a bias of eye position to the stimulus center, which is strongest in the 1-s-frame replay condition. As a consequence, inter-observer consistency is highest in this condition, though not fully explained by spatial bias alone. This leaves room for image-specific bottom-up models to predict gaze beyond generic biases. Indeed, the "saliency map" predicts eye position in all conditions, and best for continuous replay. Continuous replay predicts real-world gaze better than 1-s-frame replay does. In conclusion, experiments and models benefit from preserving the spatial statistics and temporal continuity of natural stimuli to improve their validity for real-world gaze behavior.

## Introduction

The question as to which factors guide human eye movements under natural conditions has puzzled researchers for decades. Many have approached this issue by showing observers complex natural photographs or pictures, while tracking their eye position (Buswell, 1935; Yarbus, 1967; Mannan, Ruddock, & Wooding, 1996; Reinagel & Zador, 1999; Krieger, Rentschler, Hauske, Schill & Zetsche, 2000; Privitera & Stark, 2000; Parkhurst, Law, & Niebur, 2002; Peters, Iyer, Itti, & Koch, 2005; Tatler, Baddeley, & Gilchrist, 2005; Baddeley & Tatler, 2006). Such passive-viewing approaches were extended to showing systematically modified photographs (Einhäuser & König, 2003; Kayser, Nielsen, & Logothetis, 2006) or movies (Tosi, Meccacci, Pasquali, 1997; Itti, 2005; Carmi & Itti, 2006), and combined with computer-game, simulator or virtual environment settings (Hayhoe, Ballard, Triesch, Shinoda, Alvar, & Sullivan, 2002; Peters & Itti, 2008). To what extent such laboratory data are informative for unrestrained gaze allocation during natural behavior has, however, remained largely unaddressed.

Eye-tracking in natural scenes has motivated stimulus-driven ("bottom-up") attention models. Most are rooted in the concept of the saliency map (Koch & Ullman, 1985; Itti, Koch, & Niebur, 1998): the stimulus is filtered in different channels (color, luminance, orientation) to obtain maps of feature contrasts. These are added across spatial scales to a saliency map, the peak of which is predicted to draw most attention. Although not originally designed for fixation prediction or to operate on natural scenes, a location's saliency and its probability to be fixated are (weakly) correlated (Parkhurst et al., 2002; Peters et al., 2005). However, the model's features do not necessarily drive gaze causally (Einhäuser & König, 2003), but rather act through mutual correlations to higher level scene content (Elazary & Itti, 2008; Einhäuser, Spain, & Perona, 2008a). Furthermore, if observers search for a template, the correlation vanishes entirely (Henderson, Brockmole, Castelhano, & Mack, 2006) and immediately (Einhäuser, Rutishauser, & Koch, 2008b). In a virtual reality setting top-down signals supersede bottom-up saliency for gaze allocation (Rothkopf, Ballard, & Hayhoe, 2007). Consequently, recent extensions of the saliency map include task-specific information (Navalpakkam & Itti, 2005). The saliency map's lack of causality and its breakdown for specific tasks has spurred another question: Can *any* purely bottom-up model have predictive power for gaze allocation during real-world behavior?

Despite the advantages of well-controlled settings and stimuli, eye-movement studies in the laboratory typically suffer from several constraints. First, stimuli must be chosen to be adequate for the natural situation. In particular, biases already present in the stimuli, such as preference for specific features at the center of gaze (Tatler, 2007), must be matched to the natural setting under investigation. Second, laboratory recordings typically restrain the observer, suppressing head-movement components of gaze allocation, which are particularly relevant for large gaze-shifts (Stahl, 1999) and a major source of inter-individual differences (Fuller, 1992). Third, vestibular stimuli are typically not matched to the visual input. Finally, the display's limited resolution and extent provide an artificial frame of reference. These constraints necessitate a quantification of differences between laboratory and natural settings.

Several studies measured real-world eye-movement behavior for specific tasks, including driving (Land & Lee, 1994; Land & Tatler, 2001), food preparation (Land, Mennie & Rusted, 1999; Land & Hayhoe, 2001), and a variety of sports (Land & McLeod, 2000; Fairchild, Johnson, Babcock, & Pelz, 2001; Hayhoe, Mannie, Sullivan & Gorgos, 2005; Chajka, Hayhoe, Sullivan, Pelz, Mennie, & Droll, 2006). In contrast, gaze allocation during free exploration of natural settings has rarely been addressed, although its laboratory homologue, "free-viewing", provides the typical test-bed for bottom-up models.

Virtual reality presents a step towards real-world scenarios that preserves the controllability of laboratory settings. Jovancevic, Sullivan and Hayhoe (2006) measured eye-movement patterns evoked by surprisingly occurring colliders (other pedestrians) during task performance, and established the scheduling scheme between the bottom-up (collider events) and the top-down (task) signals. Their results were in remarkable agreement with an optimal scheme for such scheduling proposed in earlier theoretical work (Sprague & Ballard, 2003; Sprague, Ballard, & Robinson, 2007): observers minimize the expected cost of not making a specific eye movement. Unlike static displays, virtual-environment settings readily allow for naturalistic tasks. In contrast to truly real-world experiments, the same trial can be replicated and the environment can be controlled, allowing quantitative manipulations and thus assessment of the interaction between task, stimulus and context (Rothkopf et al., 2007). Since the statistics of a truly natural environment may influence gaze and in turn gaze-shifts affect the selected subset of stimuli to operate upon, the controllability of the environment is, however, virtue *and* challenge. Nonetheless, virtual reality complements truly real-world experiments and will help bridging the gap between laboratory and actual reality.

In earlier work, we used a mobile setup ("EyeSeeCam", Schneider, Bartl, Bardins, Dera, Boning & Brandt, 2005; Schneider et al., in press) to record large amounts of gaze-centered and head-centered movies during free exploration. Eye-head-coordination analysis suggested a profound influence of non-saccadic eye movements to gaze-centered stimulus statistics (Einhäuser et al., 2007; Einhäuser, Moeller et al., in press), and the spatial statistics of features at the center of gaze transferred the concept of feature saliency from the laboratory to the real world (Schumann et al., 2008; Einhäuser, Schumann et al. in press). A direct comparison between free-exploration and laboratory data with the same visual input has, however, yet to be performed.

Here we compare gaze-allocation behavior during free exploration and during replay of videos in a standard head-fixed setup. Laboratory stimuli are taken from head-centered movies recorded simultaneously with the free-exploration data. This ensures that eye-in-head movements operate on the same visual stimuli in all conditions. We dissociate effects on eye movements arising from the visual stimulus alone from effects specific to either free exploration (e.g., resulting from vestibular input) or the laboratory (e.g., resulting from head restraints). To dissociate effects of stimulus continuity, we show static frames from the videos to a second set of laboratory observers. By using data from these three conditions, free-exploration, continuous replay and 1-s-frame replay, we address three topics. First, does the spatial distribution of gaze relative to the head differ in active free exploration and head-restrained free viewing? Second, are stimulus locations that are preferentially fixated, consistent within and between the laboratory and real-world conditions? Finally, does the correlation of fixation probability with saliency map values, which here exemplifies a typical bottom-up model of attention, transfer from static images, to movies and to the real world?

## Methods

### *Free Exploration*

Gaze-centered and head-centered videos were recorded using the “EyeSeeCam” setup, which is described in detail elsewhere (Schneider et al., 2005; Schneider, Bartl, Dera, Boning, Wagner & Brandt 2006; Brandt, Glasauer & Schneider, 2006; Vockeroth, Bardins, Bartl, Dera, Schneider, 2007), as is the recording procedure and the stimulus material (Schumann et al., 2008). In brief, an eye-tracking system attached to swimming goggles controls a gaze-centered camera, while an identical camera (head camera) is fixed to the observer’s forehead. Both cameras had a resolution of 752x432 pixels, covered a visual angle of 60° x 41° and recorded digital video at 25Hz. In the present study, fifteen 90-s excerpts from the head-centered movies recorded in Schumann et al. (2008) were selected as stimuli for the laboratory experiments (Fig.3.1A). These data were recorded in different environments - German residential areas (4), Munich downtown (3), hospital indoor, forest, open field, scientific conference indoor, park, university building indoor, Californian desert and beach. Movies were recorded in color through an IEEE-1394 (“firewire”) interface using Bayer-encoding with no compression. The total number of observers for recording all used movies was 6, but each movie stemmed from a continuous recording in one particular observer. All observers were accustomed to wearing the setup and instructed to “behave naturally”. The eye-position data were reconstructed from the motor-control signals of the gaze-centered camera. These data defined the gaze relative to the head-centered stimuli for the free-exploration condition.

The EyeSeeCam records eye-movements at 192 Hz. By interpolating and subsampling the signal to 25Hz we obtain one sample per frame. We conservatively defined a saccade as any fast movement (velocity>35 deg/s, acceleration>4000deg/s$^2$). By this definition 7.2%±2.1% (mean and sd across movies) of samples contained saccades. Excluding all frames for which at least 10% of samples in the frame or adjacent frames were saccades, has no qualitative effect on results.

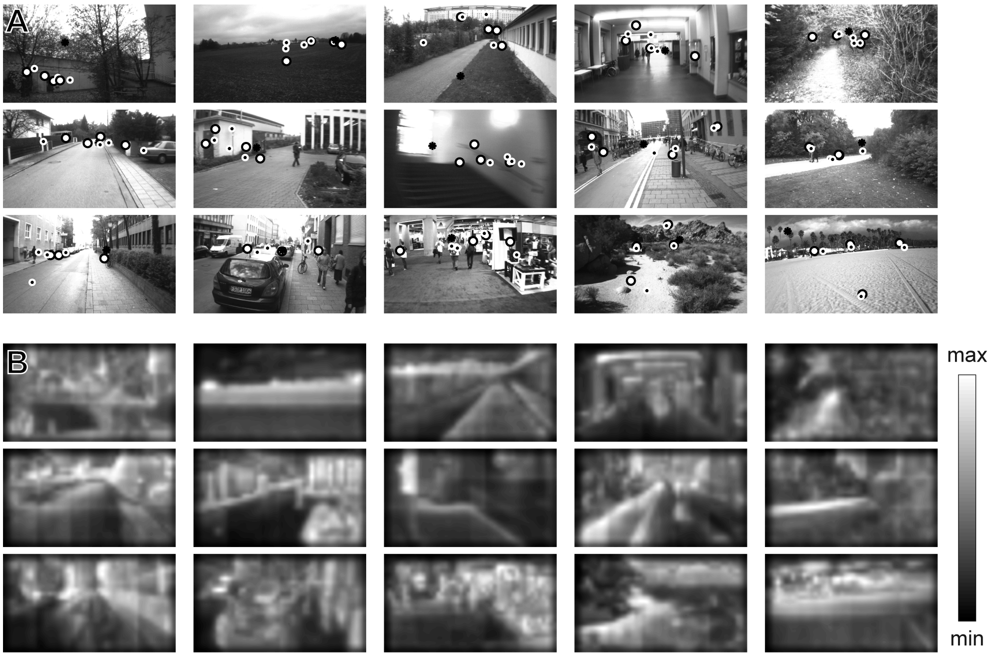

***Figure 3.1***

*(A) Frames appearing at t=20s of each of the 90s movie clips. Markers denote eye positions during presentation of the frame, black star: free exploration, black circles: continuous replay (median eye position during 40ms of frame presentation); white circles: 1-s-frame replay (median eye position 280ms-320ms after stimulus onset). Note that the videos were recorded and displayed in color. All video material is available from the authors.(B) Saliency maps for the frames shown in panel A.*

### Setup

In the laboratory experiments stimuli were presented in a dark room on a FlexScan F77S (EIZO, Hakusan, Ishikawa, Japan) 19.7' CRT monitor at 48cm distance. Four-pixel wide fringes were cropped from the stimuli at each side, resulting in a resolution of 744x424 pixels. Stimuli were scaled using bilinear interpolation to the central 1280x730 pixels of the 1280x1024 pixels wide screen, thus covering a visual angle of 45° x 26°. This is smaller than the cameras' field-of-view, being the maximum within equipment's constraints at the time of recording. The monitor's frame rate was 100Hz, an integer multiplier of the 25Hz at which movies were recorded and presented. Maximum luminance of the monitor was 37cd/m$^2$, the minimum below 0.01 cd/m$^2$. Since the precise characteristics of the cameras were unknown, screen settings were chosen such that the color movies appeared natural. In particular, the mapping from pixel-values to luminance was non-linear (gamma of 2.9), and no attempt was made to match the displayed colors physically to the real-world (x/y CIE-coordinates of the screen's guns: 0.610/0.339, 0.282/0.601, 0.151/0.065). Since most measures used in the context of saliency maps are insensitive to monotonic scaling, this restriction should not substantially affect our results.

Throughout the laboratory experiments each observer's eye position was recorded at 2000Hz using an Eyelink-2000 device (SR Research, Mississauga, Ontario, Canada). Calibration and validation procedures followed manufacturer's recommendation, using a 13-point grid on the effective display area. Eye positions outside the stimulus and blinks were discarded (2.5% of data). To avoid confounds by different dynamics of fixations and eye movements among the conditions, we did not explicitly analyze fixations. Instead, the median eye position during each 40ms frame or – for the 1-s frame conditions and all temporally resolved analyses – the eye position at each sample point was used. In 4.6%±1.1% of the frames saccades occurred. Excluding these data has virtually no effect on the results.

All presentation, eye-movement recording and analysis used Matlab (MathWorks, Nattick, MA) with its psychophysics and Eyelink toolbox extensions (Brainard, 1997; Pelli, 1997; Cornelissen, Peters, & Palmer, 2002; http://pyschtoolbox.org). All observers had normal or corrected-to-normal vision and normal color vision, as assessed by a 16—plate Ishihara test and gave written informed consent to participation. All procedures conformed with national and institutional guidelines and the Declaration of Helsinki.

### *Continuous replay condition*

In the continuous-replay condition 4 male observers (age: 21-26) watched a total of fifteen 90-s excerpts of the head-centered free-exploration videos. The observers' heads were stabilized with the chin-rest and forehead rest of the Eyelink system. Each observer viewed all movies in random order. Between the movies, observers could rest and the eye tracker was recalibrated. Observers received written instructions prior to the experiment that they were to "watch short video clips", that they should keep their heads "as still as possible", and that they were allowed and encouraged to "move their eyes naturally". The latter was added in distinction to the calibration phases, in which points and crosses had to be fixated.

### *1-s-frame replay condition*

Four additional observers (2 male, 2 female; age: 21-27) participated in the 1-s-frame replay condition. To create stimuli for this condition, the first frame of each second in each movie was selected, yielding 90 frames per movie and 15x90=1350 frames in total. These were randomly rearranged to 15 new 90-s sequences of 1-s still frames, such that each sequence contained 6 frames from each original movie. The same sequences were used for all observers, but the presentation order of sequences was randomized. All presentation- and setup parameters as well as instructions were otherwise identical to the continuous replay condition.

Taken together, we obtained for each video eye-position data of 9 different observers, four from each laboratory condition and one from the free-exploration data. Although the free-exploration videos were recorded by different observers, we will refer to them as a single "free-exploration observer".

### *Model Saliency Maps*

As a prototypical model of bottom-up guidance of eye movements, we used Itti & Koch's (2000) saliency map (http://ilab.usc.edu). Parameters were unchanged, except that max-norm normalization of the saliency map and randomness were switched off. For analysis the saliency map of each frame was normalized to range from 0 to 1 (Fig.3.1B).

### *Eye-position maps*

Average maps of eye positions relative to the image were computed by binning the eye position at each time-point at the image's resolution (744x424) and adding the resulting maps. For display purposes (Fig.3.2A-C), the maps were smoothed by averaging the 49x49pixel (3.1°x3.1° in the lab, 4.8°x4.8° in free exploration) neighborhood of each pixel and truncated with 24 pixels to each side. The binning (kernel size for smoothing) was thus matched in image coordinates (pixels) rather than in world-coordinates (degrees), and was uncritical for the location of the maxima. For each observer and movie, the spatial spread in horizontal direction and vertical direction ($\sigma_x$ and $\sigma_y$, respectively) was computed as the standard deviation of the respective eye-position coordinate, either pooling all 2000 data points per second, or – for time-resolved analysis of the 1-s-frame condition – for each of the 2000 time-points separately.

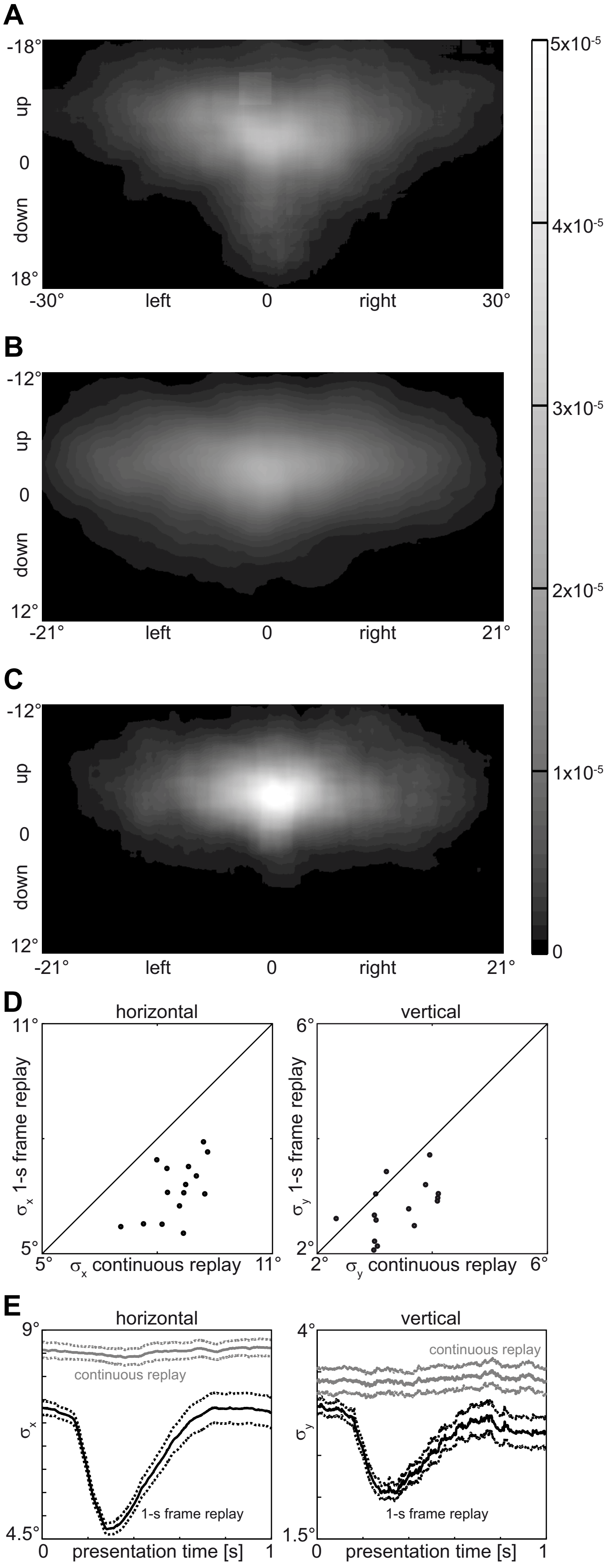

***Figure 3.2***

*(A)-(C) Density of gaze allocations averaged over all movies and observers, binned at image resolution and smoothed with a mean-filter of 49 pixels width. Maps show the full valid field of view. (FoV of the camera of 744x424 pixels cropped by 25 pixels half filter size at each end). Note that the colormap is brightened compared to the standard matlab gray colormap (e.g., figure 1) for better reproduction of low values. (A) free exploration, (B) continuous replay, (C) 1-s-frame replay. (D) Standard deviation of eye position in each movie averaged over 1-s-frame replay observers (y-axis) and continuous replay observers (x-axis) respectively. Each data point represents one movie clip. Left: horizontal eye position, right: vertical eye position. (E) Time-course of eye-position's standard deviation during 1-s-frame replay (black), continuous replay (gray) as reference. Data averaged over observers; lines denote mean ± s.e.m. over movies. Left: horizontal eye position, right: vertical eye position.*

### Spatial Consistency

Spatial consistency between each pair of observers' eye positions was measured by their Euclidian distance at each time-point. As display size differed from the camera field-of-view for free exploration, we obtain a dimensionless measure by dividing the Euclidian distance by the image diagonal. This value is subtracted from 1 to obtain a consistency measure, which is zero for maximally inconsistent observers, and 1 for identical observers. Since the spatial distribution of eye positions is a priori unknown, there is no analytic expression for the spatial consistency to be expected at random. To estimate the part of the consistency that is stimulus-independent and caused by generic spatial biases, we computed a random-reassignment baseline for each observer as follows: the eye-positions of each one-second interval in each movie were randomly attributed to another randomly chosen one-second interval from the same movie. Shuffling across rather than within movies slightly lowers the baseline, as it is less likely to hit the same or a nearby frame, but qualitatively the observed effects are independent of the baseline choice.

### Comparing fixation distributions

If there are some equally salient or relevant items spread through the scene, which are visited by all observers but in different order, the Euclidian measure would report low degrees of consistency. "Consistency" in an image-related interpretation might therefore be underestimated, especially for static displays. Kullback-Leibler (KL) divergence is an alternative measure of the generic similarity between fixation distributions of different observers and its variation over time. Following Tatler et al. (2005), we binned the display in squares (here: 16x16 pixels) and defined the probability P(x,y) of fixation from this histogram F(x,y) as

$$P(x.y)=\frac{(F(x,y)+\varepsilon)}{\sum_{x',y'}(F(x',y')+\varepsilon)}$$

with $\varepsilon=10^{-3}$. The KL divergence is then given by

$$KL = -P_a(x,y)\log P_b(x,y) + P_a(x,y)\log P_a(x,y) = P_a \log\left(\frac{P_a(x,y)}{P_b(x,y)}\right)$$

where $P_a$ is the eye-position distribution of a given observer, and $P_b$ the eye-position distribution of *all* other observers in the same condition. This measure cannot be used for the instantaneous comparison of fixated locations in any given frame, as the distribution cannot be estimated from a small number of eye positions; here Euclidian distance remains the most straightforward measure complementing the KL analysis.

### Signal-detection analysis (ROC)

For the analysis of saliency, we use the same baseline as for spatial consistency. This avoids any confound from shared spatial biases of stimuli and observer (e.g., Tatler, et al., 2005; Tatler, 2007) as the baseline and the actual data share condition, setup and observer biases. Hence any differences between the distribution of saliency at actual gaze and at baseline locations are then guaranteed to result from stimulus-specific effects. Using signal-detection analysis, we quantified how well saliency map values discriminate actual eye position from baseline locations. For a given detection threshold, we obtained the rate of hits (true values above threshold divided by number of all true values) and false alarms (baseline values above threshold divided by number of all baseline values). By varying the threshold from the minimum to the maximum of the values, we obtained the receiver operator characteristic (ROC) curve of hit rate versus false alarm rate. The area under this curve (AUC) quantifies how well saliency discriminates eye positions from baseline locations. Note that this usage of the ROC is somewhat different from earlier eye tracking studies (Peters et al., 2005; Tatler et al., 2005), where in an individual image it is asked how well saliency can discriminate fixated from non-fixated locations. Here, with only one or two fixations per subject and frame in free exploration and continuous replay, the AUC estimate from this frame-based measure would be inaccurate.

Instead, we here measured how well saliency discriminates eye positions from baseline locations within observers, movies and conditions to allow a comparison of the quality of model saliency map predictions between different experimental conditions. As this procedure does not allow tailoring parameters of the decision process to individual images, it returns lower numerical values and is the more conservative measure.

### *Empirical Saliency Maps*

To test how well laboratory data predicts free-exploration data, we defined empirical saliency maps: For each data sample, we created a map by placing a Gaussian of 2° standard-deviation centered at the eye position. These maps were added for all data samples and observers to get two empirical saliency maps for both laboratory conditions in each 40ms period of each movie (one frame in the continuous replay condition). The empirical maps were then normalized to range from 0 to 1, and the same signal-detection-theory analysis as for the model saliency maps was used.

### *Gaze-centered Average Saliency Maps*

To visualize the relationship between gaze allocation and saliency, we computed average saliency maps that were centered at gaze. For each observer, each model saliency map was shifted to align the gaze location with the center of the to-be-created average saliency map. The saliency values of all pixels in the same new pixel were added. To obtain the mean map for each observer, each coordinate of the map was divided by the number of pixels that went into it after all frames were summed. Finally, we averaged maps over observers. For display, the results were cropped to the central part corresponding to the size of the original map. For this analysis, the median gaze position in a 40ms time window (corresponding to 1 frame in the continuous condition) was used. Again, the same analysis was performed for a random-reassignment baseline that paired frames and gaze at random. Depending on the strength of the relationship between gaze and saliency one predicts a peak at the center of the new average saliency maps. If image-specific saliency contributes to gaze allocation, this peak should be more pronounced for the actual than for the baseline data. Unlike other analysis techniques, the average maps do not only use the peak or specific statistics, but consider the entire map's values. This has the advantage that if several points have similar salience but only one attracts attention this is reflected in the map. As all analyses that use the saliency values, rather than the position of peaks, the measure is, however, sensitive to non-linear scaling of individual maps. Hence we use it for visualization and to verify the analysis qualitatively, whereas quantification is based on signal-detection-theory measures, which are insensitive to any strictly monotonic scaling of the saliency value.

## Results

We compare human gaze allocation during three different conditions; free exploration, continuous replay and 1-s-frame replay of head-centered movies. In laboratory conditions, fixations (periods not containing saccades or blinks) account for 93.2%±1.1% of time (1-s-frame replay) and 92.4%±2.9% (continuous replay), respectively, with no significant difference between the two conditions (t-test, t(6)=0.53, p=0.61). In free exploration, such periods account for 94.8%± 0.9% of samples, and 72.3% ± 4.7% of frames correspond entirely to a fixation or slow (i.e., non-saccadic) movement.

### *Spatial distribution of eye positions*

We measure the spatial distribution of eye positions in the head-centered coordinate frame. In all three conditions gaze-allocations show a spatial bias towards the upper center of the visual field. The peak is above the vertical midline, 3.5° in free exploration (Fig.3.2A), 2.7° in continuous replay (Fig.3.2B), and 3.0° in 1-s-frame replay (Fig.3.2C). Horizontally, the peak is displaced slightly to the left in free exploration (0.3°) and continuous replay (1.1°), and virtually at the center for 1-s-frame replay (0.1°). In sum, in all conditions the eyes direct gaze about 3° upward on average, but show little bias sideward.

As display sizes were matched in the two laboratory conditions, the height of the peaks can be compared directly to measure how pronounced a spatial bias is. With $5.6\text{x}10^{-5}$ of data in the maximum bin (Fig.3.2C), the maximum in 1-s-frame replay is more than twice as high as in continuous replay (Fig.3.2B, $2.5\text{x}10^{-5}$). This is a first indication that the spatial bias is enhanced by the discontinuous presentation in 1-s-frame replay. The spatial spread in horizontal and vertical direction in both conditions for each movie quantifies this further. In all 15 movies the standard deviation of horizontal eye position is larger in continuous replay than in 1-s-frame replay (Fig.3.2D, left). A sign-test shows that this fraction of movies is significant ($p=6.1\text{x}10^{-5}$). Similarly, the spread in vertical direction is larger in continuous replay in 12 out of 15 movies (Fig.3.2D, right), also a significant fraction (p=0.04, sign-test). Consequently, gaze allocation is spatially more constrained for static than for continuous presentation.

We analyze the time-course of spatial spread from the onset of each new stimulus in the 1-s-frame condition. We find a dip at 283ms (horizontal) and 338ms (vertical) after the onset of a new frame (Fig.3.2E, black). As baseline, the same spatial spread for the continuous replay, for which the time-point relative to the 1-s intervals has no particular meaning, shows little variation and is above the value for 1-s-frame replay (Fig.3.2E, gray). This shows that the spatial bias in 1-s-frame replay is most pronounced about 300ms after stimulus onset, even though there is no blank or enforced fixation between subsequent frames. These 300ms may reflect the time needed for processing new visual information and making an according eye movement. We interpret the stronger spatial bias during discontinuous presentation along with its time-course as evidence that the central bias often observed in laboratory experiments is to a large degree a consequence of "resetting" eyes to the stimulus center when new information is onset.

### Consistency between observers

Consistency between individuals is a necessary – though not sufficient - prerequisite for bottom-up models to predict gaze allocation from stimulus statistics alone. Hence we test how consistent different observers are in their eye position, and to what degree this consistency is explained by common spatial biases. Measuring consistency by the average pair-wise Euclidian distance between two observers (Fig.3.3A), we find the highest consistency within the 1-s-frame replay observers with 86.7%±1.0% (mean±sd across the 6 pairs), which is significantly larger than within the continuous-replay observers, who reach 82.6%±1.4% (t(10)=5.97, p=1.4x10$^{-4}$, t-test), and across the laboratory conditions of 81.1%±2.3% (t(20)=5.63,p=1.7x10$^{-5}$). The free-exploration observer is slightly more consistent with the continuous-replay observers (82.6% ±1.6%) than with the 1-s-frame replay observers (81.9%±1.0%), but this difference fails to reach significance (t(6)=0.76, p=0.48). In summary, observers in 1-s-frame replay are more consistent among each other than with observers in other conditions or observers within and across other conditions.

Given the stronger spatial bias in 1-s-frame replay, is the larger consistency in this condition an effect of spatial bias alone? We compare the within-condition consistencies to a random-reassignment baseline, which randomly shuffles seconds of presentation within each observer and movie. This baseline reflects the image-independent contribution to consistency, i.e. the contribution of the spatial bias. In both laboratory conditions, the mean baseline consistency remains significantly below the values obtained on the actual data, with 83.4% ±1.2% (t(10)=-5.25, p=3.8x10$^{-4}$) for 1-s-frame replay and 77.9%±2.3% (t(10)=-4.38, p=0.001) for continuous replay. The baseline consistency significantly depends on condition (t(10)=5.22, p=3.9x10$^{-4}$), suggesting that the difference between consistency in the two conditions is - at least in part - a consequence of the different strength of spatial bias. Furthermore, the baseline values are smaller than the actual values for all pairs of observers in both conditions, a fraction (6/6) that is significant even without taking the absolute values into account (p=0.03, sign-test). Hence, inter-observer consistency by itself is not a consequence of spatial bias alone, but contains a stimulus-specific component. There is a stimulus-driven component to gaze allocation that is shared among observers. This effect, however, operates only in addition to consistencies imposed by generic spatial biases, which are consequences of shared preferred heading direction, setup and presentation mode.

As the spatial bias exhibits a pronounced time-course during a 1-s-frame presentation, we complemented our analysis of spatial consistency time resolved. We find that consistency peaks at around the same time (at 89.9%, 342ms after stimulus onset, Fig.3.3B black) as the spatial bias. The baseline, which reflects the consistency imposed by spatial biases, peaks slightly earlier (282ms, Fig.3.3B gray) and stays consistently below the actual curve. Analyzing the continuous-replay conditions analogously does not show any pronounced peak, neither in baseline nor in actual data.

From the Euclidian consistency measure alone it is unclear to what extent the seemingly increased consistency in 1-s-frame replay results from the smaller spatial spread (Fig.3.2E). To directly compare the spatial distributions of eye positions we measure KL-divergence between an observer’s distribution and the distribution of all other observers in the same condition. The time-course averaged over the four 1-s-frame observers is very similar to the results from the Euclidian distance: consistency increases first during viewing but quickly returns to baseline (Fig.3.3C, black). Since the time-locking is arbitrary in the continuous-replay condition, we

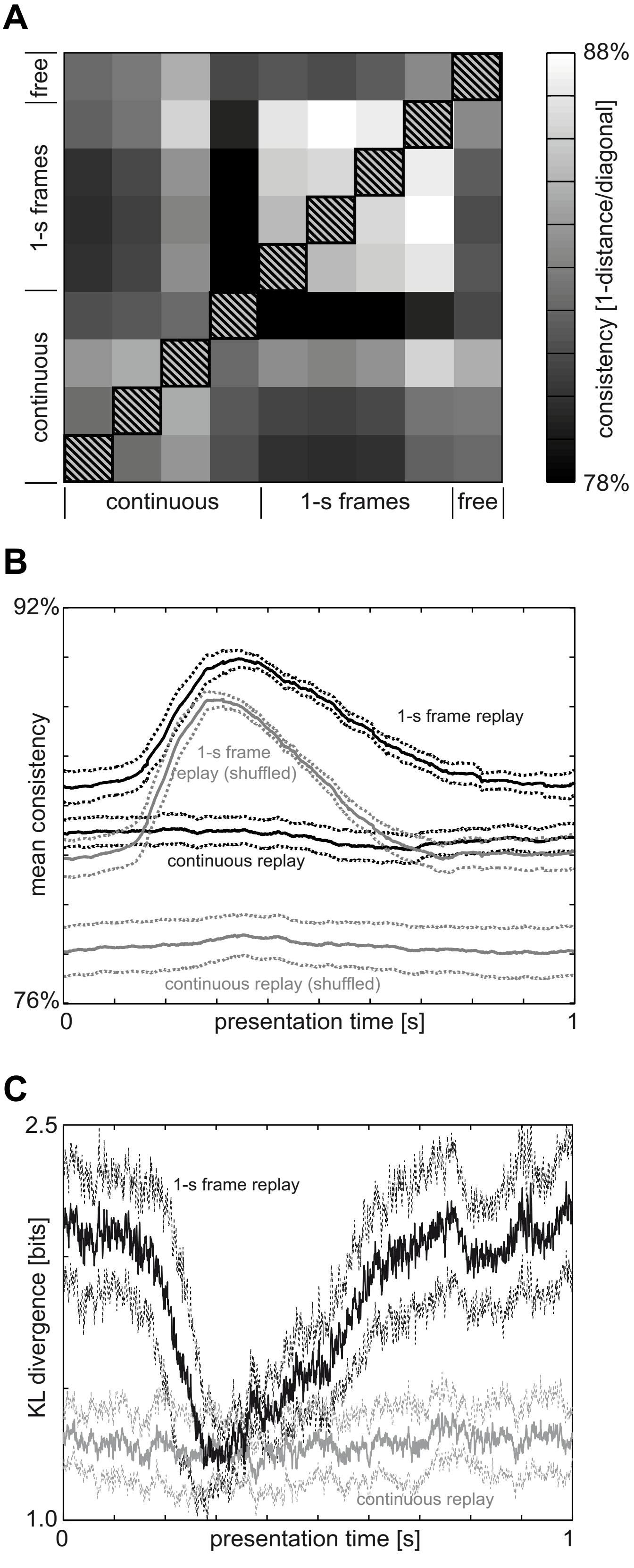

***Figure 3.3***

*A) Pairwise consistency measure of eye position for continuous replay (left 4 bins), 1-s-frame replay (middle bins) and free exploration. Data averaged over observers and movies; diagonal at 100% by definition. (B) Consistency measure over presentation time for 1-s-frame replay condition (upper black trace). Continuous replay (lower black trace) depicted for comparison. Stimulus-independent effects (random-reassignment baseline) presented in gray. All data averaged over subjects, mean ± s.e.m over movies depicted. (C) KL-divergence as alternative measure of consistency of fixation distributions. Black: mean±sem KL-divergence for the four 1-s-frame observers; gray: continuous-replay observers. Note that lower KL implies higher consistency and absolute values are irrelevant in the present context as they depend on discretization.*

again see a flat line. Remarkably, the consistency in the continuous-replay condition is higher (lower KL) on average and 1-s replay reaches comparable levels only between about 300 and 400ms after stimulus onset. This implies that the difference in Euclidian distance is fully explained by the larger spatial spread during continuous replay.

### *Average Saliency Maps*

As prototypical bottom-up model, we use Itti & Koch's (2000) saliency map. To take the entire map's values into account, we average the model saliency maps in a gaze-centered reference frame. The stimuli themselves (in the head-centered reference frame) do not exhibit a pronounced central peak (Fig.3.4, upper left), rather a stripe of highest saliency 91 pixels (9.0° in free exploration) above the midline. In gaze-centered coordinates, the peak is centered and found 9 pixels below the center of gaze by the gaze-camera's image (Fig.3.4, top-row, 2nd panel from left), and 14 pixels below, when using the motor commands and the images of the head-centered camera as in the remainder of the analysis (3rd panel). These values correspond to 0.9° and 1.4° in free-exploration, respectively. Although the peaks are rather broad with the 90th percentile spanning about half the camera's visual field (Fig.3.4), they are substantially narrower than in the baseline condition (3rd from right, Fig.3.4 top row). The continuous-replay result (top right panels) is qualitatively similar to free exploration: a peak centered at the center of gaze that is wide, but substantially narrower than baseline. Averaged over the full 1-s presentation the peak in 1-s-frame replay (middle row, left) is wider than in continuous replay, but narrower than the baseline (bottom row, left). Time-resolved analysis shows little variation for the baseline over presentation time (bottom row), which is similar in shape and peak position for all presentation conditions. In contrast, the peak for the actual data in 1-s-frame replay starts narrowing in the 200ms-400ms interval and remains at an about constant width for the remainder (Fig.3.4, middle row). This is a first indication, that saliency becomes more predictive of eye positions in the

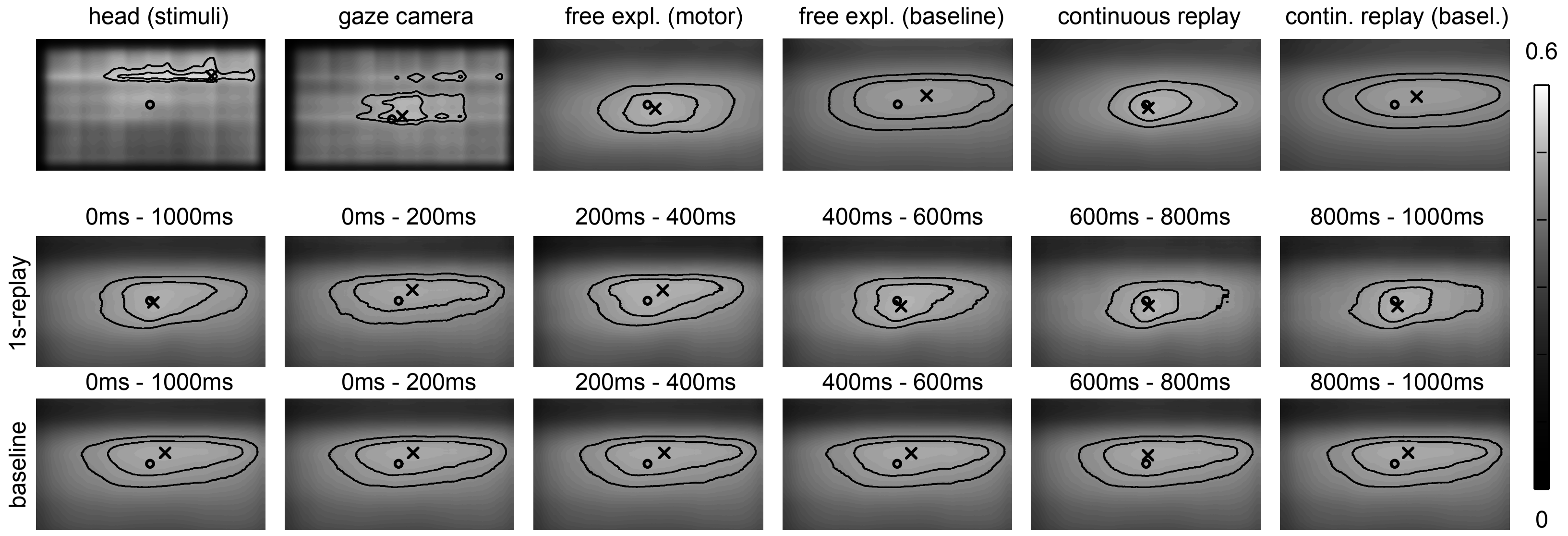

***Figure 3.4***

*Average saliency maps centered at center of gaze in various conditions. Circle: center of gaze direction (head direction in upper left); Cross: location of maximum; black lines: 95% and 90% contour lines. Top row, from the left: Stimuli (head-centered camera), gaze-centered camera (note that the center of gaze is slightly offset from the center of the image), free exploration (reconstructed from camera control signal), random-reassignment baseline for free exploration, continuous replay condition (all observers and movies), random-reassignment baseline for continuous replay. Middle row: 1-s-frame replay condition, full presentation time; presentation time in 200ms intervals after presentation onset. Bottom row: Random-reassignment baseline to middle row. Note that the difference between the gaze-centered camera and the free-exploration data reconstructed from the motor control is partly due to visual impression. The factual difference arises from the unavoidable cropping when reconstructing the gaze-centered map from the head-centered camera during a fixation at high eccentricity. Together with a potential slight misalignment of the cameras this likely causes the slight (0.5°) difference in estimated peak location.*

course of the presentation. In summary, this first qualitative analysis suggests that mode saliency is best centered at gaze for continuous replay, and has a distinct time-course for static presentations.

### *Predicting fixations by model saliency*

To quantify how saliency relates to eye positions in the different conditions we use model saliency map values to discriminate eye positions from baseline locations, i.e. to "predict" eye positions. "Prediction" here does not imply causality, and it is well conceivable that correlations to other - hidden - scene properties explain away the effects. The area under the curve (AUC) of the receiver operator characteristics (ROC) serves as measure. It reaches 50% if saliency cannot discriminate eye positions from baseline locations and 100% for perfect discrimination. In free exploration, 12/15 movies have AUCs above chance, a significant fraction (p=0.04, sign-test), with the mean AUC 53.2%±3.4% also significantly exceeding chance (t(14)=3.64, p=0.003). In the laboratory conditions, there are 60 pairs of observers and movies, of which 56 in each condition exceed chance ($p=9.1x10^{-13}$, sign-test). When averaging across observers, 13 (1-s replay) and 14 (continuous replay) movies exceed chance, and the means across movies are also significantly larger than chance in both conditions with 53.1%±2.0% (t(14)=6.12, $p=2.7x10^{-5}$) and 55.2%±2.7% (t(14)=7.58,$p=2.6x10^{-6}$), respectively. Comparing the conditions, saliency predicts eye positions better in continuous replay than in 1-s-frame replay (t(14)=4.00,p=0.001, paired t-test, Fig.3.5A), and slightly better in continuous replay than in free exploration (t(14)=2.29, p=0.04, Fig.3.5B), while there is no difference between the prediction by saliency in 1-s-frame replay as compared to free exploration (t(14)=0.09, p=0.93). In summary, saliency predicts eye positions to some extent in all conditions, and best for continuous replay.

A possible reason why saliency's prediction is worse in 1-s frame presentation than in

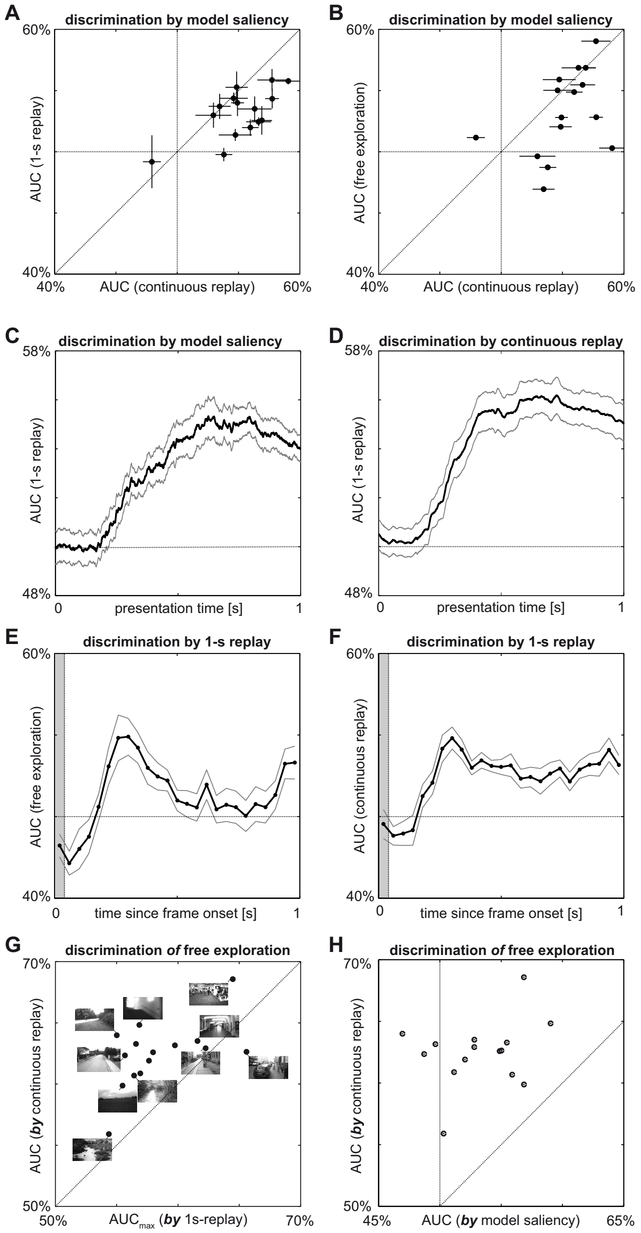

***Figure 3.5***

*(A) Area under curve (AUC) for saliency map discriminating fixated from baseline locations ("prediction" of fixated locations); each data point corresponds to one movie and depicts mean AUC and s.e.m. over the 4 observers; x-axis continuous replay condition, y-axis 1-s-frame replay. (B) As panel A, with results of free exploration on y-axis. (C) Time-course of AUC for saliency predicting fixated locations over the 1-s presentation duration in the 1-s-frame replay condition. Mean and s.e.m. over movies. (Note that the curve does not need to level off to 0.5 at t=1s, as there is a sharp transition of stimuli, and the saliency map of the current stimulus can still be predictive in the subsequent frame). (D) Prediction of 1-s replay by continuous replay, using the empirical map recorded in continuous replay at the time the respective frame is shown. (E) Prediction of free exploration by maps generated from 1-s-frame replay condition, mean and s.e.m. over movies. The frame corresponding to the 1-s-replay frame is shown from 0 to 40ms (shaded area), but predicts best about 240-280ms afterwards. (F) Prediction of continuous replay by maps generated from 1-s-frame replay condition, analogous to panel E. (G) Prediction of free exploration data by 1-s-frame replay and by continuous replay. Each movie corresponds to one data point. For 1-s-frame replay, the maximum AUC over all time-points is used for each movie, nevertheless the continuous replay condition predicts free exploration fixations better in all but one movie. Thumbnails identify selected movies. (H) Comparison of prediction of free exploration eye position between empirical saliency map from continuous replay (y-axis, same data as panel G's y-axis) and saliency (x-axis, same data as panel B's y-axis).*

continuous presentation is the time it takes until saliency gets effective. Subsequent frames are highly correlated in movies (temporal continuity), but independent in 1-s-frame presentation. Saliency can thus only be expected to predict eye position after the visual system has processed the new stimulus and initiated an eye movement. To analyze how long it takes for saliency's prediction to become effective, we compute the AUC separately for each time-point during the 1-s presentation. As expected, the prediction starts at chance level (50%) because the present frame is unrelated to the preceding one and generic effects are accounted for by baseline. After about 200ms the prediction by saliency starts to become effective. The mean across movies starts to differ significantly from chance (at an alpha level of 0.036 corresponding to an expected false discovery rate of 0.05) for the first time after 256ms and reaches its maximum after 624ms (Fig.3.5C). This characterizes the time it takes until a bottom-up signal related to a novel stimulus affects eye position, and accounts in part for the worse average prediction of 1-s replay by saliency.

### *Predicting fixations across conditions*

To test the mutual prediction between different conditions, we use the empirical saliency maps for the laboratory conditions to predict the other conditions. Within the same condition the prediction of the empirical saliency map, thanks to the small number of observers is always near ceiling, with 97.9%±1.0% for continuous replay and 79.4%±3.1% for 1-s replay (maps pooled over the full second). The eye positions on the respective frame in continuous replay condition predicts 1-s-frame replay with a similar time-course as model saliency and reaches an only slightly higher maximum of 56.2% (Fig.5D) compared to the 55.3% of saliency's prediction (Fig.5C). In turn, we test how well empirical saliency maps generated from 1-s-frame replay predict gaze allocation. Prediction is best 280ms after the stimulus corresponding to the shown frame was encountered in the real world (Fig.5E) or the frame was shown in the continuous-replay condition (Fig.3.5F). The peak AUC reaches about the same height in the free exploration (54.8%±5.4%) as in the continuous-replay condition (54.2%±3.2%), values that are not significantly different (t(14)=0.57, p=0.58, paired t-test).

As an upper bound for the prediction of free-exploration by 1-s-frame replay, we compute the best prediction by 1-s-frame replay *for each movie,* reaching a mean AUC of 58.4% ±3.5%. Note that this number is different from the maximum of the mean curve, resulting from the fact that the maximum prediction is reached at different times for different movies. Even in this measure that is beneficial for 1-s-frame replay, the eye positions in continuous presentation predict real-world gaze allocation better (AUC 62.5%±2.7%) than the individual frame (t(14)=5.20, $p=1x10^{-4}$, paired t-test). This advantage for predicting the real-world by continuous replay holds for all but one individual movie (14/15, Fig.3.5G), which constitutes a significant fraction of movies (Fig.3.5G; $p=9.8x10^{-4}$, sign-test). Both predictions are, however, correlated (r=0.60, p=0.03), and exceed chance for all movies. Qualitatively, both laboratory conditions predict eye positions during free exploration well, when the scene contains man-made structures with plenty of isolated objects (residential areas, indoor environments). In contrast, prediction is worse for natural outdoor sceneries (e.g., desert, open field). The continuous replay condition seems to have particular benefits in situations of highly dynamic character, such as passing uneven terrain or climbing a flight of stairs (Fig.3.5G).

Empirical saliency maps from continuous replay predict eye positions during free exploration better than the model saliency maps for all movies (15/15, Fig.3.5H) and significantly better on average (paired t-test, t(14)=9.50,$p=1.8x10^{-7}$). If the peak prediction in each movie is considered (as in Fig.3.5G), this also hold for the empirical maps from 1-s replay (better in 13/15 movies, t(14)=4.8, $p=2.6x10^{-4}$). However, there is no individual time-point (0, 40ms, 80ms, etc.) for which the mean prediction of this empirical map exceeds the prediction of the model map significantly, and for the first 160ms the prediction is even significantly worse. Hence only the data from continuous replay is a better predictor of free-exploration eye movements than model saliency. This shows that the eye position in the laboratory has some predictive value for eye positions in the real-world, but replicating the dynamic aspect of the scene, especially its temporal continuity, is of importance. In all, our results render the development of models based on laboratory data useful, but also stress the importance of spatially and temporally realistic stimuli, and validation in the real world.

## Discussion

We compare laboratory measurements of eye position to free exploration data. Observers' eye positions are more consistent when static frames are presented than for movies, but most of this surplus consistency is explained by spatial biases that are independent of the specific visual stimulus shown. A prototypical bottom-up model of attention, the saliency map, exhibits a weak but significant correlation with eye position inside and outside the laboratory. There is a slight advantage for saliency in continuous input over 1-s-frame replay. This may be explained by the time a bottom-up signal needs to be processed and to trigger an eye movement when a novel stimulus is onset. This is, the benefit is a result of temporal continuity in the real world and continuous replay. Finally, we show that gaze recorded in the laboratory possesses some predictive power for gaze allocation in the real-world, which is improved if the full dynamics of the stimulus is maintained.

Spatial biases on visual attention and gaze direction have received increasing interest. Mannan et al. (1996) stress that most of their features' effect on eye position vanishes when correcting analysis for shared spatial biases in stimuli and eye position. Similarly, Tatler et al. (2005) demonstrate that a varying spatial bias fully explains the changing effect of low-level features over prolonged viewing (Parkhurst et al., 2002). In turn, spatial biases – or prior knowledge on the likely locations of search targets – complement saliency in guiding eye movements in search (Torralba, Oliva, Castelhano, & Henderson, 2006). A probabilistic model that learns a saliency representation from natural scene statistics – combining bottom-up saliency and generic top-down biases - also outperforms image-specific saliency models for free-viewing (Zhang, Tong, Marks, Shan, & Cottrell, 2008). A systematic study on spatial biases in free viewing of natural scenes found that – at least under laboratory conditions – central fixation biases prevail irrespective of known biases in stimulus features (Tatler, 2007). This suggested either a role of the artificially limited setup that makes it more effective to look at the center or a general bias to look straight ahead. Recently, we have used the relation of gaze-centered and head-centered feature statistics to argue against the latter alternative (Schumann et al., 2008). Individual features, as saliency does here, typically show an environment-dependent bias towards the upper half of the head-centered visual field. Gaze centers and refines this bias, which argues against a pure *centering* of eyes in their orbit. By using the camera-control signals relative to the head, we here do find a weak bias in viewing direction for free exploration. This bias is, however, not entirely central in the vertical, although the size of its upward deviation compared to the full oculomotor range still may justify a dubbing as "central". More importantly, however, the bias is strongest in the 1-s-frame replay condition, especially compared to continuous replay. It exhibits a time-course that suggests that new information arriving triggers this reset to the center. This suggests that central bias indeed serves to select an optimal starting location for early scene processing on a limited screen when no other prior is available, the first hypothesis proposed by Tatler (2007). Provided the results on spatial priors in search (Torralba et al., 2006), it is likely that task and presentation conditions influence spatial biases. In free-exploration, the spatial distribution of gaze suggests a bias towards the open path to be walked on, which is weaker in the laboratory conditions. This might be interpreted as prior for the implicit task of actually navigating the terrain. To what extent such motor planning and action contribute to the difference between “free-exploration” and “free-viewing” remains an open issue, for which virtual-reality experiments may provide interesting complementary data (cf. Jovancevic et al., 2006).

Inter-observer consistency is a necessary condition for a purely bottom-up model to make any predictions (causal or correlative) on gaze allocation. If “bottom-up” is restricted to the presently presented stimulus, a successful prediction needs to exceed that of generic biases. In our small set of laboratory observers, inter-observer consistency exceeds the baseline from generic biases alone in all conditions and observer pairs. Remarkably, the stimulus-specific consistency in 1-s frame replay persists longer than the generic component. This suggests a prolonged effect of bottom-up signals for static stimuli and shows that bottom-up models can be useful for gaze prediction. Yet, our results stress that – even in the absence of an explicit task - a good part of inter-observer consistency is determined by setup, presentation conditions and – through spatial biases in head-centered videos – the stimuli.

As example for a bottom-up model, we tested Itti & Koch’s (2000) saliency map, using its original version rather than its more recent developments (Itti, 2005; Itti & Baldi, 2006; Cerf, Harel, Einhäuser, & Koch, 2008). The correlation between model saliency and fixation probability is weak, but exceeding chance. It is, however, remarkable that even this static model achieves better predictions for the continuous replay condition, at least when compared to a baseline that takes generic biases into account. This stresses the importance to faithfully match not only the spatial but also the temporal statistics of stimuli to the real world. The importance of temporal dynamics is supported by the fact that empirical saliency derived from continuously presented stimuli predicts real-world gaze better than 1-s replay does, even when the latter is used at the optimal latency from stimulus presentation to its prediction. These results highlight that temporal continuity is a key principle not only for object recognition (Wallis & Rolls, 1997; Einhäuser, Hipp, Eggert, Körner & König, 2005), but also for attention deployment under natural conditions. It thus might be used to learn attention models from natural stimulus statistics (Zhang et al., 2008).

One reason for the worse “prediction” in and by 1-s replay is the time needed to deploy bottom-up attention after stimulus onset. The minimal time needed for a bottom-up effect of a newly onset image that is unrelated to a previous one is the time to process this stimulus by the visual system and to execute one volitional eye movement. Although fast saccades are possible in response to natural scene categorization (Kirchner & Thorpe, 2006), even the fastest express saccades require at least 100ms (Crouzet, Kirchner, & Thorpe, 2008). Hence, a delay of bottom-up responses till 150ms-200ms after stimulus onset is in line with physiological constraints and – by itself - does not argue against the “pre-attentive” nature of target selection, suggested by the saliency map. Furthermore it is notable that there is no substantial decline of saliency’s prediction over the 1-s period. Our data furthermore confirm earlier findings of a decline of inter-observer consistency after an initial increase when the starting eye position is random or – in our case – determined by an unrelated stimulus (Tatler et al., 2005).

Provided the task’s importance for eye movements (Buswell, 1935; Yarbus, 1967) and saliency map predictions (Henderson et al., 2006), its effect on the comparability of eye tracking in the real-world to laboratory settings remains an interesting issue. It is likely that eye movements in highly-trained, stereotypic motor tasks that are tied to specific settings (driving, sports, etc.) cannot be reproduced by visual display alone. Hence eye movements obtained during real-world tasks provide the opportunity to improve and test computational models with data from natural tasks that are difficult to elicit in the laboratory. It is conceivable that such models may then even excel empirical data from the laboratory with respect to gaze prediction in real-world scenarios.

In any case, the comparison between laboratory and real-world data can help in uncovering the role of a specific modality (e.g., vision) in attention allocation during everyday implicit tasks, such as walking on uneven terrain, stair climbing or navigating. As a first step in this direction, we here followed this approach for the arguably most naïve tasks possible, free-viewing and free exploration.

In conclusion, we here quantified for the first time the differences between laboratory and real-world settings. The potential sources of the observed differences are manifold: First, it is unclear whether "free-viewing" really represents a laboratory version of "free exploration". Second, there is the limited display in the laboratory, third, the effect of the restriction of head movements, and forth the absence of vestibular and other cross-modal information. Future experiments, that systematically modify task, display size and location, head position and input from modalities other than vision, and finally a larger number of observers performing multiple conditions, will allow a detailed investigation of all of these issues. Here, we delivered the proof of concept that our novel recording setup allows addressing these topics and provided first quantitative results.

### *Acknowledgements*

The authors gratefully acknowledge the following grants and institutions: BMtH was supported by the Deutsche Forschungsgemeinschaft (DFG) Research Training Group 885-Neuroact; JV received support by the Bayerische Forschungsstiftung (DOK-88-07); PK by the EU project IST-027268-POP "Perception On Purpose". This work was conducted in part within the DFG excellence initiative research cluster Cognition for Technical Systems – CoTeSys.

## References

Baddeley, R.J., & Tatler, B.W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research, 46*, 2824-2833.

Brainard, D.H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433-436.

Brandt, T., Glasauer, S., & Schneider, E. (2006). A Third Eye for the Surgeon. *Journal of Neurology, Neurosurgery, and Psychiatry, 77,* 278.

Buswell, G.T. (1935). How people look at pictures. A study of the psychology of perception in art. Chicago, IL: The University of Chicago Press.

Carmi, R & Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research, 46* (26), 4333-4345.

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing (NIPS) 20,* 241-248.

Chajka, K., Hayhoe, M., Sullivan, B., Pelz, J., Mennie, N., & Droll, J. (2006). Predictive eye movements in squash [Abstract]. *Journal of Vision, 6*(6):481, 481a.

Crouzet, S., Kirchner, H., & Thorpe, S.J. (2008) Saccading towards faces in 100 ms. What's the secret? *Perception* **37** ECVP Abstract Supplement, p.119.

Cornelissen, F.W., Peters, E.M., & Palmer, J. (2002) The Eyelink Toolbox: eye tracking with MATLAB and the Psychophysics Toolbox. Behav Res Meth Instr Comput 34:613-617.

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience, 17*, 1089-1097.

Einhäuser, W., Hipp, J., Eggert, J., Körner, E., & König P. (2005). Learning viewpoint invariant object representations using a temporal coherence. principle. *Biological Cybernetics, 93,* 79-90.

Einhäuser, W., Schumann, F., Bardins, S., Bartl, K., Böning, G., Schneider, E. & König, P. (2007). Human eye-head co-ordination in natural exploration. *Network: Computation in Neural Systems, 18*(3), 267-297.

Einhäuser, W., Spain M., & Perona, P. (2008a). Objects predict fixations better than early saliency. *Journal of Vision. 8*(14):18, 1-26.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008b). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision, 8*(2):2, 1-19.

Einhäuser, W., Moeller, G.U., Schumann, F., Conradt, J., Vockeroth, J., Bartl, K., Schneider, E., König, P. Eye-head coordination during free exploration in human and cat. *Annals of the New York Academy of Sciences, 1431* (in press).

Einhäuser, W., Schumann, F., Vockeroth, J. Bartl, K., Cerf, M., Harel, J., Schneider, E., & König P. Distinct roles for eye and head movements in selecting salient image parts during natural exploration. *Annals of the New York Academy of Sciences, 1431* (in press).

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision, 8*(3):3, 1-15.

Fairchild, M.D., Johnson, G.M., Babcock, J., & Pelz, J.B. (2001). Is Your Eye on the Ball? Eye Tracking Golfers while Putting. (*http://www.cis.rit.edu/people/faculty/fairchild/*)

Fuller, H.J. (1992). Head movement propensity. *Experimental Brain Research. 92,* 152-164.

Hayhoe, M., Ballard, D., Triesch, J., Shinoda, H., Alvar, P., & Sullivan, B. (2002). Vision in natural and virtual environments. *Proceedings of the symposium on Eye Tracking Research & Applications.* 7–13.

Hayhoe, M., Mannie, N., Sullivan, B., & Gorgos, K. (2005). The role of internal models and prediction in catching balls. *Proceedings of AAAI*, Fall 2005.

Henderson, J.M., Brockmole, J.R., Castelhano, M.S., & Mack, M. (2006) Visual Saliency does not account for Eye-Movements during Visual Search in Real-World Scenes. R. van Gompel, M. Fischer, W. Murray, R. Hill (eds). *Eye Movement Research: Insights into Mind and Brain.* Elsevier.

Itti, L., Koch, C., & Niebur, E. (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20* (11), 1254-1259.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*(10-12), 1489-1506.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition, 12*(6), 1093-1123.

Itti, L., & Baldi, P. (2006) Bayesian Surprise Attracts Human Attention. *Advances in Neural Information Processing Systems (NIPS 2005), 19,* 1-8.

Jovancevic, J., Sullivan, B. & Hayhoe, M. (2006) Control of attention and gaze in complex environments. *Journal of Vision*, 6:1431-1450.

Kayser, C., Nielsen, K.K., Logothetis, N.K. (2006). Fixations in natural scenes: interaction of image structure and image content. *Vision Research, 46*(16), 2535-2545.

Kirchner, H. & Thorpe, S.J. (2006) Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. *Vision Research.* **46**(11):1762-1776.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology, 4*, 219-227.

Krieger, G., Rentschler, I., Hauske, G., Schill, K. & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision, 13*, 201-214.

Land M.F., & Lee, D.N. (1994). Where we look when we steer. *Nature, 369*, 742 – 744.

Land, M.F., Mennie, N., & Rusted J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception, 28*, 1311–1328.

Land., M.F., & McLeod, P. (2000). From eye movements to actions: how batsmen hit the ball. *Nature Neuroscience,* ***3***, 1340-1345.

Land, M.F., Tatler, B.W. (2001). Steering with the head. the visual strategy of a racing driver. *Current Biology, 11,* 1215-1220.

Land, M.F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research, 41,* 3559-3565.

Mannan, S.K., Ruddock, K.H. & Wooding, D.S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision 10*(3), 165-188.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention, *Vision Research, 45*, 205-231.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*, 107-123.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10,* 437-442.

Peters, R.J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research 45*(18), 2397-2416.

Peters, R.J., & Itti, L. (2008). Applying computational tools to predict gaze direction in interactive visual environments. A*CM Transactions on Applied Perception, 5*(2), 8.

Privitera C., & Stark, L. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations, *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(9).

Reinagel, P., & Zador, A.M. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems, 10*, 341-350.

Rothkopf, C.A., Ballard, D. H., & Hayhoe, M.M. (2007) Task and context determine where you look. *Journal of Vision*, 7(14):16, 1-20.

Schneider, E., Bartl, K., Bardins, S., Dera, T., Boning, G., & Brandt, T. (2005). Eye Movement Driven Head-Mounted Camera: It Looks Where the Eyes Look. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, 2437–2442.

Schneider, E., Bartl, K., Dera, T., Boning, G., Wagner, P., & Brandt, T. (2006). Documentation and teaching of surgery with an eye movement driven head-mounted camera: see what the surgeon sees and does. *Studies in Health Technology and Informatics. 119*, 486-490.

Schneider, E., Villgrattner, T., Vockeroth, J., Bartl, K., Kohlbecher, S., Bardins, S., Ulbrich, H., & Brandt, T. EyeSeeCam: An eye movement-driven head camera for the examination of natural visual exploration. *Annals of the New York Academy of Sciences, 1431*, (in press).

Schumann, F., Einhäuser, W., Vockeroth, J., Bartl, K., Schneider, E., & König, P. (2008). Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *Journal of Vision* 8(14),12, 1-17.

Sprague, N., & Ballard, D. (2003). Eye movements for reward maximization. *Advances in Neural Information Processing (NIPS) 16.*

Sprague, N., Ballard, D., & Robinson, A.(2007) Modeling embodied visual behaviors. *ACM Transactions on Applied Perception,* 4(2):11, 1-23.

Stahl, J.S. (1999). Amplitude of human head movements associated with horizontal saccades. *Experimental Brain Research, 126*, 41–54.

Tatler, B.W., Baddeley, R.J., & Gilchrist, I.D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, *45*, 643-659.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14):4, 1-17.

Torralba, A., Oliva, A., Castelhano, M.S., & Henderson, J.M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review, 113*, 766-786.

Tosi, V., Mecacci, L., & Pasquali, E. (1997). Scanning eye movements made when viewing film: preliminary observations. *International Journal of Neuroscience*, *92*(1-2), 47-52.

Vockeroth, J., Bardins, S., Bartl, K., Dera, T., & Schneider, E. (2007). The Combination of a Mobile Gaze-Driven and a Head-Mounted Camera in a Hybrid Perspective Setup. *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*, 2576-2581.

Wallis, G., & Rolls, E.T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51*(2),167-194.

Yarbus, A.L. (1967). *Eye movements and vision*. New York: Plenum Press.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G.W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7):32.1-20.

## *Study III*

*Mind the step: complementary roles for eye-in-head and head-in-world orientation when negotiating a real-life path*

Submitted as:

’t Hart, B.M. and Einhäuser, W. Mind the step: complementary roles for eye-in-head and head-in-world orientation when negotiating a real-life path.

# Mind the step: complementary roles for eye-in-head and head-in-world orientation when negotiating a real-life path

### *Abstract*

Gaze in real-world scenarios is controlled by a huge variety of parameters, such as stimulus' features, instructions, or context, all of which have been studied systematically in laboratory studies. It is, however, unclear how these results transfer to real-world situations, when participants are largely unconstrained in their behavior. Here we measure eye and head orientation and gaze in two conditions, in which environment (i.e., context and features) and instruction are identical, while the effect of implicit task set is varied by terrain regularity. We show that terrain regularity causes specific differences in head orientation and gaze behavior, which are restricted to the vertical direction. Participants direct their head and eyes lower when terrain difficulty increases, but only the eyes compensate partially for this by spreading eye-in-head orientation more in the vertical direction. Our results quantify the importance of task set for gaze allocation in the real world, and imply qualitatively distinct contributions of eyes and head in gaze allocation. This underlines the care that needs to be taken when inferring real-world behavior from constrained laboratory data.

## Introduction

Sensory systems provide an organism with the information needed for skilled behavior, adapted to a changing environment. In the case of human vision, high spatial resolution is limited to the fovea, demanding shifts of gaze for detailed visual sampling of the environment. For several decades research on gaze-shifts for natural stimuli has focused on eye movements (Buswell, 1935; Yarbus, 1967), although in real-world situations humans shift gaze by coordinated movements of head and eyes. To what extent do behavioral constraints influence eye-in-head and head-in-world movements, respectively, to select the relevant information from sensory stimuli?

When viewing natural scenes, allocation of gaze can be partly explained by stimulus-driven factors. The most abundantly used concept is that of a saliency map: originally developed to explain shifts in covert attention using simple stimuli (Koch & Ullman, 1985), it predicts gaze allocation in natural scenes exclusively based on low-level stimulus features (Itti & Koch, 2000). However, the model can be seen as a predictor of characteristic objects in a scene (Carmi & Itti, 2006) and those objects – when known – explain away the effect of low-level features (Einhäuser et al., 2008b). Consequently, the saliency map may not be understood as causal model of gaze allocation and analysis of their performance is often confounded by common factors to gaze and saliency (e.g., central bias, Tatler, 2007). Nevertheless, for free exploration of an outside environment, saliency maps still have some predictive power, albeit lower than for “free-viewing” in typical laboratory conditions (’t Hart et al., 2009). This underlines the need for experiments under truly natural conditions.

Besides features of the current stimulus, contextual or prior knowledge about expected stimulus statistics drives attention to a considerable extent. For example, when looking for a pedestrian in a street scene, fixations are better predicted by a model that weighs lower locations of the stimulus more strongly (Torralba, 2003). Furthermore, pre-attentive knowledge of scene layout biases fixation distributions (Ehinger et al., 2009). In parallel to the usage of such spatial priors on fixation behavior, several models predict attention (and thus gaze) capture by deviations from (learnt) prior expectations on feature (Itti & Baldi, 2006) or scene (Zhang et al., 2008; Bruce & Tsotsos, 2009) statistics. All these approaches model attention allocation based on stimulus statistics, but – in contrast to the basic saliency-map approaches – relate to more stimuli than just the one available at the very moment.

While all the aforementioned approaches are stimulus-driven (either by the current stimulus, its relation to the sequence or the stimulus ensemble), so called “top-down” factors control the deployment of attention and thus shifts of gaze as well. Besides memory (Droll & Eckstein, 2009) and other idiosyncratic factors (Hidalgo-Sotelo & Oliva, 2010), the task is arguably the best studied of those top-down factors. Classically studied in search displays (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977; Treisman & Gelade, 1980; Wolfe et al., 1989), in natural scenes, search overrides stimulus-driven saliency completely (Henderson et al., 2007) and immediately (Einhäuser et al., 2008a). Sophisticated models, such as Guided Search (Wolfe et al., 1989; for the latest version see Wolfe, 2007) explain search in visual displays, incorporating top-down information. Recently, such quantitative analysis has been extended to template search in static natural scenes (Navalpakkam & Itti, 2005; Pomplun, 2006; Zhang et al., 2008). Search is a great starting point for quantitative analysis of task-dependent gaze allocation. However, even if we have to search for misplaced items or well-defined signposts more often than we might like to, in daily living, template search is hardly the most common or “natural” mode of human behavior. In the present paper we therefore aim at considering more natural task sets, which are not given explicitly by instruction, but implicitly by constraints of the environment.

Unlike in most laboratory situations, in real life the organism itself decides in part what the actual ‘task’ is and this task may change from moment to moment. A participant’s intentions indeed affect not only change detection (Triesch et al., 2003) and the processing of specific object-features (Hannus et al., 2005), but also gaze distributions directly (Castelhano et al., 2009; Rothkopf & Ballard, 2009). As predicting the trajectory of a bouncing ball by making eye-movements to upcoming relevant locations exemplifies (Hayhoe et al., 2005), gaze behavior adapts to increasing knowledge of the environment's physical properties. With the recent advent of wearable eye-trackers, studying the allocation of gaze in real-life scenarios has become feasible. However, most research in this area has restricted itself largely to either free exploration (Cristino & Baddeley, 2009; 't Hart et al., 2009; Schumann et al., 2008) or to specific – thus experimentally readily controllable - domains such as driving, sandwich making or sports (Hayhoe & Ballard 2005; Kandil et al., 2009; Land et al., 1999; Land & McLeod, 2000). In free exploration, there is no explicit instruction and the participant implicitly selects their task. In the other scenarios, the task (sandwich making, driving, tea making, cricket) typically specifies the full range of actions to be taken and does not leave - or in the case of driving, should not leave – much room for other actions, such as exploring the visual environment. In contrast, walking outdoors – at least in healthy adults – is a natural task that leaves considerable room for visual exploration (Calow & Lappe 2008; 't Hart et al., 2009). The visual environment, context and instruction can be held constant by constraining the path to be walked on, while at the same time the experimenter can control the implicit part of the task (e.g., not tripping) by varying the terrain regularity. Here we therefore use walking on terrain of varying regularity as a paradigm that constrains parts of the task, but does not fully occupy participants, such that they still may decide on their behavior in a naturalistic way.

In most laboratory settings, gaze direction is changed by movements of the eyes alone, whereas in real-world settings movements of the body and head are available as well. Fixing head and body not only ignores the possibility of orienting head or body to allocate gaze, but also neglects reflexive eye movements that accompany head and body movements in the real world. The orientation of the eye is offset against changes of the orientation of the head via the vestibulooccular reflex (VOR), and similarly the vestibolocollic reflex (VCR) stabilizes head-in-world orientation during larger body movements while VOR is suspended (Guitton & Volle, 1987). The way these movements interact to direct gaze has for example been investigated in the real-life tasks of making tea (Land et al., 1999) and driving (Land, 1992). In both tasks, large gaze changes were accompanied by head movements proportional to the gaze change, but head movements were smaller when body movements were made as well (Land, 2004). Similar patterns of interaction between these three types of movements have been found in a walking task in a laboratory setting on even terrain (Imai et al., 2001). With respect to terrain regularity, the relative importance of the lower visual field for negotiating irregular terrain has been shown by blocking downward viewing (Marigold & Patla 2008) and by tracking gaze while participants walked a short, irregular path in the laboratory (Marigold & Patla 2007; Patla & Vickers 2003). Here we combine the measurement of gaze direction during walking outdoors with the variation of terrain regularity. We quantify two contributions to gaze, eye-in-head and head-in-world orientation, where the latter comprises head and body movements.

Besides the relative contributions of eye, head and body to gaze during largely unconstrained walking, the question as to where to optimally direct gaze to is of crucial importance when walking more complex environments, with the avoidance of collisions being a central issue. When walking on an obstacle course or in traffic, gaze is usually directed at obstacles (Ballard & Hayhoe, 2009), at locations where cars are likely to appear (Geruschat et al., 2003) or at pedestrians that are more likely to collide with the participant than others (Jovansevic-Misic & Hayhoe, 2009). In all these cases, items that have to be dealt with (obstacles, cars, pedestrians) attract gaze. It is conceivable that this generalizes to walking on terrain, where terrain irregularities have to be negotiated and thus may also attract gaze. If true, varying terrain regularity should affect gaze.

Even in the absence of obstacles or other road users, walking is a complex task (Hausdorff et al., 2005), which uses depth cues (Hayhoe et al., 2009), motion parallax or optic flow (Bardy et al., 1996; Callow & Lappe, 2008; Warren et al., 2001), and vestibular information (Fitzpatrick et al., 1999; Jahn et al., 2000). Walking has an effect on gaze allocation, as there is an abundance of downward directed eye movements during walking as compared to the same visual stimulation with fixed body and head ('t Hart et al., 2009). This "T-shaped" distribution of eye-in-head orientation during walking has also been described earlier (Callow & Lappe, 2008) and emphasizes the importance of the lower visual field in real-life walking (Marigold & Patla, 2008; Timmis et al., 2009). Visual information is indeed used during walking: when depriving participants of all visual input during specific segments of each step, foot placement precision drops (Chapman & Hollands, 2006; Hollands & Marple-Horvat, 1996). This furthermore demonstrates a link between the sampling of visual information on the terrain and phases of the step cycle, and suggests there are specific visuomotor routines organizing walking (Imai et al., 2001). The role of these routines becomes evident when they operate under strict constraints or are disturbed, such as in the elderly (Cavanagh & Higginson, 2002; Chapman & Hollands, 2006; Jahn et al., 2010; Startzell et al., 2000) or patients suffering from Parkinsonian syndromes (Pinkhardt et al., 2008). In analogy to restrictions imposed by bodily or sensory impairments, variation of terrain regularity should then affect the sampling of visual information in healthy participants. This again yields the hypothesis that terrain regularity affects gaze.

Laboratory studies have shown that terrain regularity correlates with look-ahead distance on the path (Marigold & Patla, 2007; Patla & Vickers, 2003). In an artificially sparse laboratory environment, however, looking at the surroundings serves little purpose. Therefore gaze behavior may differ in this restricted situation as compared to real-world behavior. One study that investigated gaze on path in a real-world setting (Pelz & Rothkopf, 2007) manipulated path difficulty together with changing the environment, such that it cannot be excluded that this change in surroundings accounts for the effects found. To the best of our knowledge, no study to date has addressed the relation between real-life gaze allocation and terrain difficulty without changes to the visual environment.

In the present study, we use walking to investigate the role of implicit task set on gaze during natural behavior. Participants are asked to walk up and down an inclined street, while their eye-in-head movements and head-in-world orientation are tracked with a novel, wearable eye-tracking device (“EyeSeeCam”; Schneider et al., 2009). In the two experimental conditions, instructions and environment are identical, but terrain regularity is varied by once using irregularly placed steps, once the comparably smooth road running in parallel to the steps. This procedure allows us for the first time to quantitatively assess in a realistic scenario, how different contributions to gaze direction depend on an implicit task set (safely negotiating terrain) with all other parameters (environment, instruction, etc.) held constant.

## Materials and Methods

### *Participants*

Eight volunteers (4 male, 4 female; mean age ± SD: 30.3 ± 7.1 years) with normal or corrected to normal vision and no walking deficits participated in this experiment. All participants gave written informed consent before the experiment. The experiment conformed to institutional and national regulations and the Declaration of Helsinki.

### *Conditions*

The main experiment took place in a local street (‘Hirschberg’) that has a sidewalk with irregularly placed steps on one side (Figure 4.1A). The main street is an inclined cobbled road. The sidewalk and main street are separated by a metal railing. Participants walked on the road as well as on the steps, close to this railing. Since this was repeated for walking up and down, each

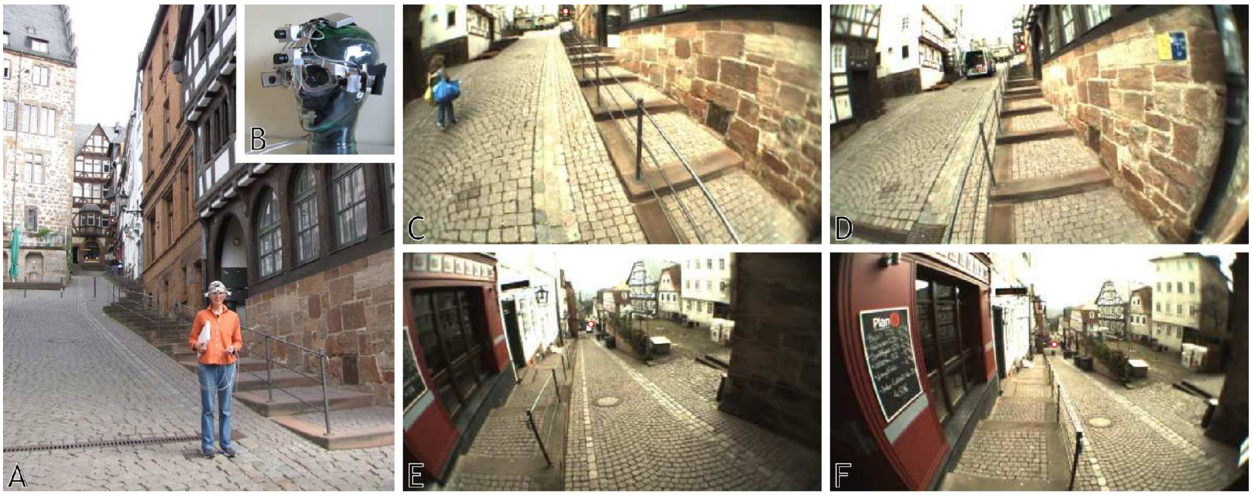

***Figure 4.1: Setup & Conditions.***

*A) Local street, in which experiments were conducted. "steps" are to the right, "road" to the left of the handrail. B) EyeSeeCam device. C-F) First frame of analyzed data of the same participant for different conditions. Red circle indicates vanishing point (see Methods) C) up/road D) up/steps E) down/road F) down/steps.*

participant walked through the street a total of four times. To counter any effects of order on behavior, the order of the four walks was randomized. The path that was to be taken was explained to the participants right before the walk and the instruction for each of the four walks was to 'walk as you normally would'. That is, with the exception of whether to use steps or road, instructions were exactly identical in all conditions. As the environment remains unchanged as well, the only difference between conditions, which we will refer to as "road" (Figure 4.1C,E) and "steps" (Figure 4.1D,F), is the implicit task set of negotiating terrain of distinct difficulty.

To verify that other environments induce qualitatively similar eye movements, six of the eight volunteers participated in two additional conditions, referred to as "stairs" and "alley". In the "stairs" condition, they walked up and down a continuous flight of stairs, which is considerably more regular than the "steps" condition. In the "alley" condition, they walked a path with negligible incline compared to the "road" condition.

Since the alley and stairs were considerably shorter than the inclined road and steps, recording time in the main conditions "steps" and "road" were longer than in the conditions "alley" and "stairs". On average participants took 59.81s ± 3.00s (mean ± SD) for "steps", 51.87s ± 7.76s for "road", 11.56s ± 1.91s for "alley" and 12.67s ± 2.14s for "stairs".

### Setup

In all conditions, eye-movements were recorded during the walk with a mobile, wearable eye-tracker ("EyeSeeCam"; Schneider et al., 2009; Figure 4.1B). The eye tracker recorded the eye-in-head signal at 305 Hz for both eyes, and – if signals from both eyes were available – the average of the eyes was used for further analysis. In addition a camera fixed to the forehead recorded a movie of the environment with a wide angled lens ("head-cam") and a camera moving with the direction of gaze recorded a gaze-centered movie. In the present study, we used this gaze-centered video to verify the eye-in-head measurements and the head-centered video to determine head-in-world orientation.

The EyeSeeCam software defines the origin for eye-in-head orientation as a straight-ahead direction relative to the device, such that there is some variability (up to a few degrees) between individuals. Similarly the head-in-world orientation (see below) is defined relative to device-centered reference frame (the head-cam). Since the camera is not removed from the participant throughout the experiment and there is only little slippage of the goggles to which the device is mounted, the definition of the origin is consistent across all conditions. Hence none of the differential effects between conditions can be confounded by the choice of reference frame. In addition, for the determination of gaze (eye-in-world) the offsets of the two device-centered origins compensate each other, such that the gaze coordinate systems of different individuals are identical.

### Eye-in-head orientation

To analyze eye-in-head orientation, the eye-tracker data was separated in a horizontal and a vertical component. For each component, we determined the mean and the standard deviation of the eye-in-head orientation. We quantified the dependence of these parameters on terrain (road/steps) and walking direction (up/down) by a 2-factor ANOVA.

### *Head-in-world orientation*

As a proxy for head-in-world orientation we determined the position of a point at ground level beyond the end of the walking track in frame coordinates in the head-centered movies. In loose analogy to descriptive geometry, we refer to this point as the vanishing point. The definition of this point depends on environment and necessarily on walking direction (up/down). Since environment is identical for "steps" and "road" and the line of sight is much longer than the width of the used path (steps and road combined), the vanishing point is virtually independent of terrain. Consequently, any effect of terrain cannot be attributed to the vanishing-point definition, while the effects of walking direction may. However, since we are interested in the factor terrain, the effect of walking direction will remain irrelevant, unless we would observe an interaction between walking direction and terrain.

To determine the vanishing point, we used the following manual procedure. First a pixel was selected by clicking on the frame with a mouse pointer in every 30th frame. A square section around the selected pixel was used as a template. In the preceding and succeeding frames this template was used for 2D convolution to determine the location of the vanishing point in all frames of the movie during all walks (Figure 4.1C-F). Using the Camera Calibration Toolbox for Matlab (Bouguet 2010), we characterized the lens of the head-fixed camera and calculated normalized coordinates for all vanishing point locations. Since the extent of the camera was known (120° x 70°) an angular signal for head-in-world orientation could be estimated using the same metric as the eye-in-head signal (degrees of visual angle). By analyzing the distribution of the vanishing point's position in the head-centered movies, we can assess the contribution of head-in-world orientation to terrain induced changes of gaze direction.

### *Gaze-in-world*

For each head-in-world sample, eye-in-head orientation and head-in-world orientation signals were added to obtain a gaze-in-world signal (cf. Land & Tatler, 2001). Both the vertical and horizontal component of this signal were tested for their dependence on walking direction (up/down) and terrain (road/steps) using a 2-factor ANOVA.

## Results

### *Gaze-in-world*

To investigate the effect of implicit task sets on gaze allocation, we asked participants to negotiate different real-world terrains, while eye-in-head and head-in-world orientation was recorded. Before addressing the separate contributions of head-in-world movements and eye-in-head movements to gaze, we will analyze the combined signal; that is, the distribution of gaze. Visually inspecting the raw distributions of gaze one observes two distinct peaks on both types of terrain (road/steps). In the road condition (Figure 4.2A), one peak is centered near the vanishing point and one peak falls about 20° below that, presumably on the path itself. In the steps condition (Figure 4.2B) this second peak is more pronounced and located about 10° lower than in the road condition, while the central peak remains virtually at the same location. Horizontally, all peaks are close to the midline. The difference between both conditions is a first qualitative indication that effects of terrain act mostly along the vertical axis.

To quantify the effects of terrain on gaze, four ANOVAs were performed, all used the factors terrain (road vs. steps) and walking direction (up vs. down). In the horizontal, mean gaze

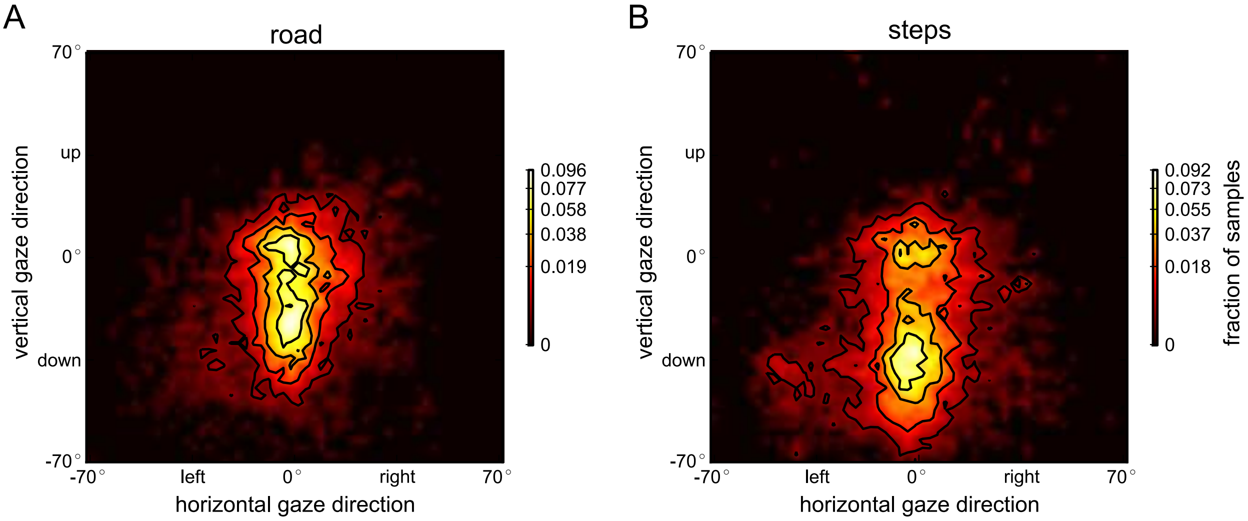

***Figure 4.2: Gaze-in-world histograms.***

*Histograms of gaze-in-world relative to the vanishing point (2.5° x 2.5° bins, interpolated for display) Note the logarithmic scale for individual panels. Vertical and horizontal axes correspond to real space, measurement range is 70° from the center in all directions. Height lines correspond to colorbar ticks. Data are first averaged per participant, such that each individual contributes equal amounts of data independent of walking speed. A) road B) steps.*

direction depends on walking direction (F(1,28) = 117.28, *p* = .005, Figure 4.3A-C), but not on terrain (F(1,28) = 8.81, *p* = .408). There is no interaction between the two factors (F(1,28) = 15.92, *p* = .268). In contrast, the mean vertical gaze direction is affected by terrain only (F(1,28) = 1134.51, *p* < .001), whereas the effect of walking direction (F(1,28) = 3.59, *p* = .794) and the interaction (F(1,28) = 113.04, *p* = .151) are not significant.

In addition to the mean direction of gaze, the aggregate data (Figure 4.2A,B) suggests a wider spread of gaze in the vertical for the steps condition. To quantify this, we compute the standard deviation of the horizontal and vertical gaze component (Figure 4.3D-F). In the horizontal, there is no significant main effect of terrain or walking and no interaction (all *p* > . 180). In contrast, the vertical component shows an effect of terrain (F(1,28) = 112.32, *p* < .001) but no effect of walking direction (F(1,28) = .01, *p* = .967) or interaction between the two factors (F(1,28) = 2.29, *p* = .535).

Taken together, the data on mean and standard deviation show that terrain affects gaze in

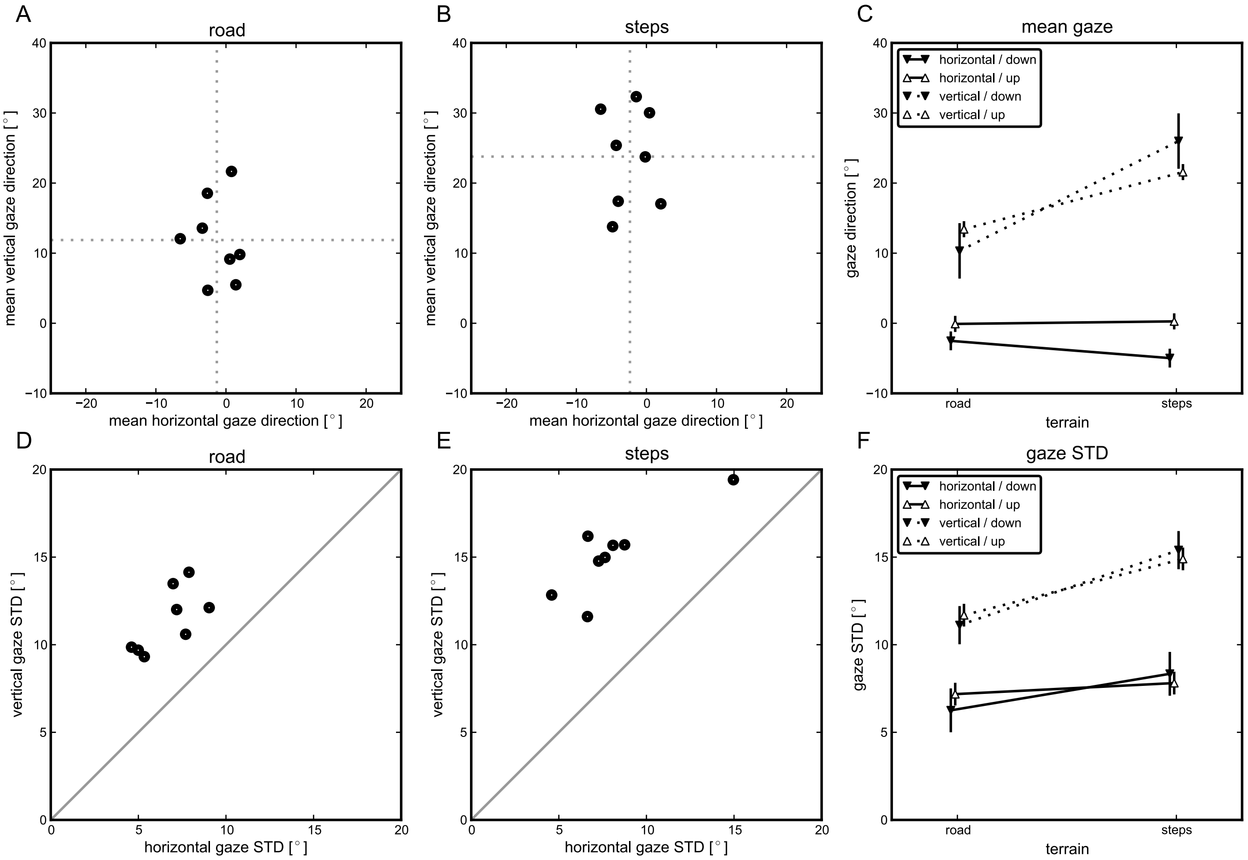

***Figure 4.3: Gaze-in-world mean and standard deviation.***

*A-B) Mean horizontal and vertical gaze-in-world for each individual (circles), and average (dotted lines) A) road B) steps. C) Comparison of the road and step data of panel A and B. Errorbars denote standard errors of the mean. D-E) Standard deviation over horizontal and vertical gaze-in-head orientation. D) road E) steps. Note the consistently larger spread in the vertical. F) Comparison of the data of panels D and E, further split up for walking up and down the path. Errorbars denote standard errors of the mean.*

the vertical direction. The more difficult terrain ("steps") induces lower and more spread gaze than the easier terrain. It is important to note, that visual environment and instructions are identical in both cases, such that all effects result from the interaction of the implicit task set of negotiating terrain with terrain difficulty. The effect of walking direction on horizontal gaze direction, however, might be a consequence of the necessarily different choice of vanishing points for determining head-in-world orientation, which factors into the measure of gaze direction. Both effects raise the question to what extent eye-in-head as compared to head-in-world movements contribute to the observed differences.

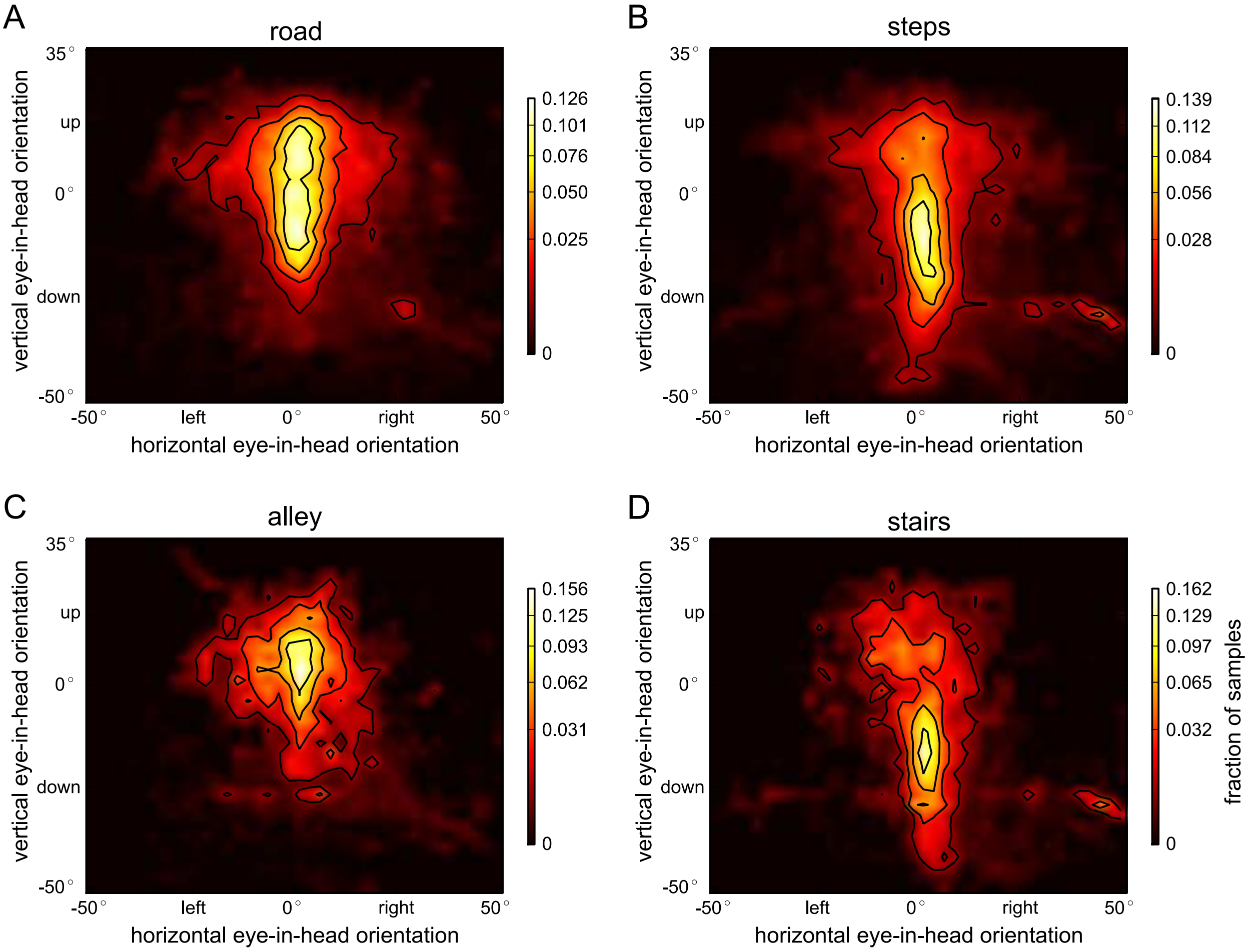

***Figure 4.4: Eye-in-head orientation histograms.***

*Histograms of eye-in-head orientation (2.5° x 2.5° bins, interpolated for display) Note the logarithmic scale for individual panels. Vertical and horizontal axes correspond to real space, measurement range is 35° from the midline towards the top, 50° towards all other directions. Height lines correspond to colorbar ticks. Data are first averaged per participant, such that each individual contributes equal amounts of data independent of walking speed. A) road B) steps C) alley D) stairs.*

### Eye-in-head orientation

As one contribution to gaze, we analyze eye-in-head orientation. Visual inspection of the raw distributions shows that the peak of the eye-in-head orientation distribution is higher in the "road" (figure 4.4A) than in "steps" (figure 4.4B) condition. Quantitative analysis shows that the mean horizontal eye-in-head orientation (Figure 4.5A-C) does not depend on terrain or direction, nor is there an interaction between these factors (all $p > .402$). In contrast, the mean vertical eye-in-head orientation does depend on terrain (F(1,28) = 458.25, $p < .001$, Figure 4.5C) and walking

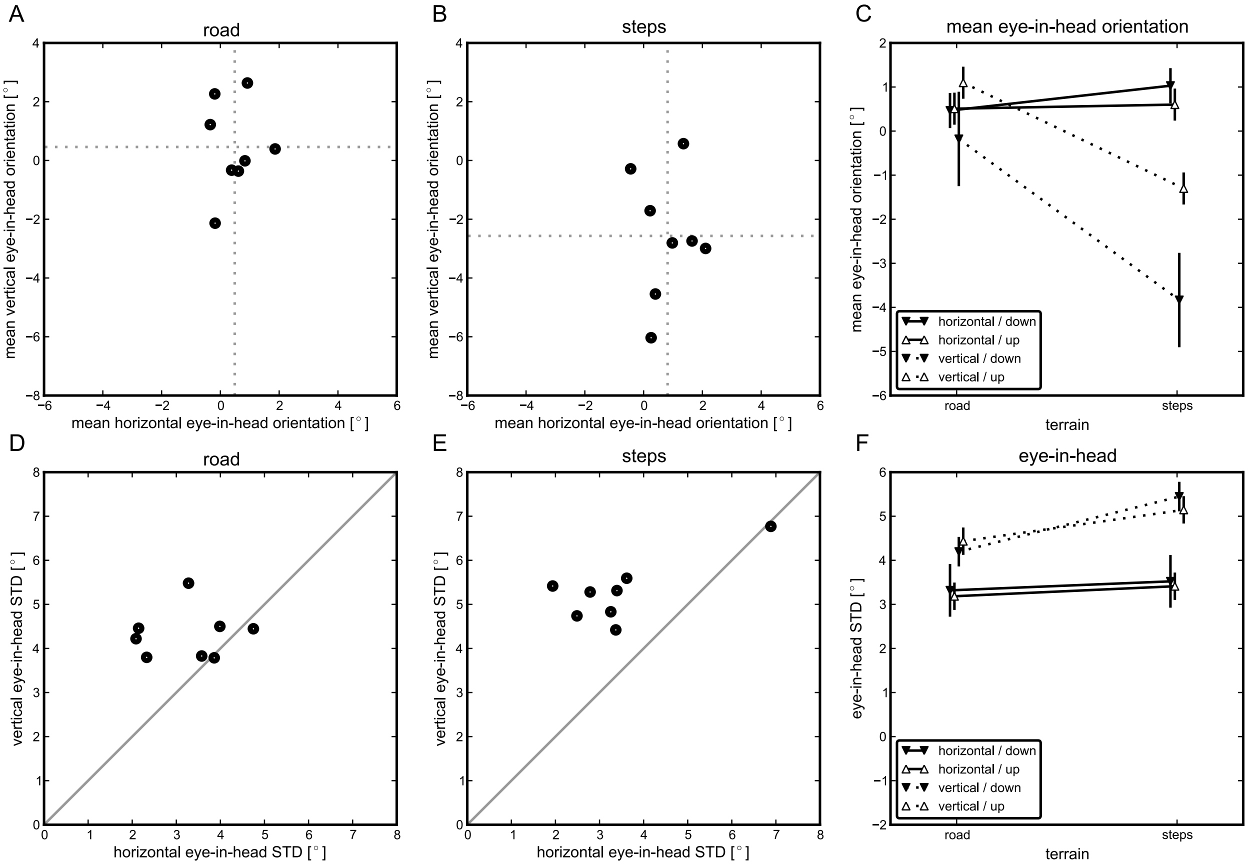

***Figure 4.5: Eye-in-head orientation mean and standard deviation.***

*A-B) Mean horizontal and vertical eye-in-head orientation for each individual (circles), and average (dotted lines) A) road B) steps. C) Comparison of the road and step data of panel A and B. Errorbars denote standard errors of the mean. D-E) Standard deviation over horizontal and vertical eye-in-head orientation. D) road E) steps. Note the consistently larger spread in the vertical. F) Comparison of the data of panels D and E, further split up for walking up and down the path. Errorbars denote standard errors of the mean.*

direction (F(1,28) = 181.16, $p$ = .016). There is no interaction between terrain and walking direction (F(1,28) = 19.69, $p$ = .403). While the mean eye-in-head orientation on average is almost on the midline for the road (Figure 4.5A), it falls clearly below for steps (Figure 4.5B). This is a first indication that terrain difficulty affects vertical eye orientation in that the more irregular terrain (“steps”) demands eye position to be directed more towards the ground.

To quantify the spread of eye-in-head orientation, we calculate the standard deviation over the vertical and horizontal components of eye-in-head orientation. For all conditions, we find that for the majority of participants the standard deviation over the vertical eye-in-head orientation is larger than the horizontal one (road: 6/8 participants, steps: 7/8 participants, Figure 4.5D, E). This shows that eye orientations are more spread in the vertical than in the horizontal direction. The standard deviation over the horizontal eye-in-head orientation does not depend on terrain or walking direction, nor is there an interaction between these factors (all *p* > .663; Figure 4.5F). The standard deviation over the vertical coordinates does depend on terrain (F(1,28) = 48.08, *p* < .001) and is larger for "steps" than for "road". Walking direction does not have an effect on the standard deviation over the vertical eye-in-head orientation (F(1,28) = 0.05, *p* = . 90)2 and there is no interaction between terrain and walking direction (F(1,28) = 3.64, *p* = .301). Hence eye-in-head orientation is more spread vertically if terrain gets more irregular. This may imply that there are more or larger eye movements for the irregular terrain, but may also result from longer fixations in the lower visual field. It should be noted, however, that – unlike in viewing static images in the lab – not only saccades contribute to eye-in-head orientation, but also stabilizing and tracking eye movements, which yields a highly dynamic situation that renders a

precise categorization of eye movements at each point in time difficult. In sum, we find robust effects of terrain irregularity on eye-in-head orientation, which are restricted to the vertical direction. This suggests that with increasingly irregular (i.e., more "difficult") terrain eye movements increasingly direct gaze to the path.

### Effect of environment

Unlike head-in-world orientation and gaze, eye-in-head orientation is independent from the definition of the vanishing point. This allows the comparison to other visual environments. We chose two environments ("alley", "stairs"), which are similar in terrain regularity to the main conditions (road/steps), but present a different visual environment. Visual inspection of raw distributions of eye-in-head orientations (Figure 4.4) indicates that the distribution for steps is more similar to stairs and alley more similar to road than main and control conditions are relative to each other. Since the visual environment and the path inclination change between main conditions and control conditions (and among these), quantitative isolation of terrain's effect is not possible, which is the key rationale of locating the main conditions in the same environment. Qualitatively, however, the observation that the (visual) environments of alley and stairs are more similar to each other than to the road/steps environment (e.g., with respect to openness), make the data suggest that the effect of terrain may at least partially supersede the effect of environment. In any case, the predominant elongation of eye-in-head orientation distributions along the vertical as compared to the horizontal is present for all environments tested.

### *Head-in-world orientation*

To test if head-in-world movements are also affected by demands posed by the terrain, we analyzed the position of the vanishing point within the head-centered movie frames. From the vanishing point's position we determine the head-in-world orientation by inverting the transfer function of the camera. This yields a representation of head orientation in visual angle relative to the vanishing point.

Visual inspection of the raw distributions (Figure 4.6) shows that they are shifted more upward for the "steps" as compared to the "road" condition. Since these data are given from the perspective of the head, not the point in the movie frame, they imply that the head points *down*wards in both conditions, but more downward in the steps condition. Abstracting from the way the data are obtained, we hereafter follow the more intuitive convention for "head-in-world orientation", such that "downward" implies the head pointing lower, etc. As for the eye-in-head orientation, we quantify the distribution by the mean head-in-world orientation (Figure 4.7A-C) and the standard deviation over the vertical and horizontal components (Figure 4.7D-F). In both conditions the standard deviation over the horizontal component is larger than the vertical one in

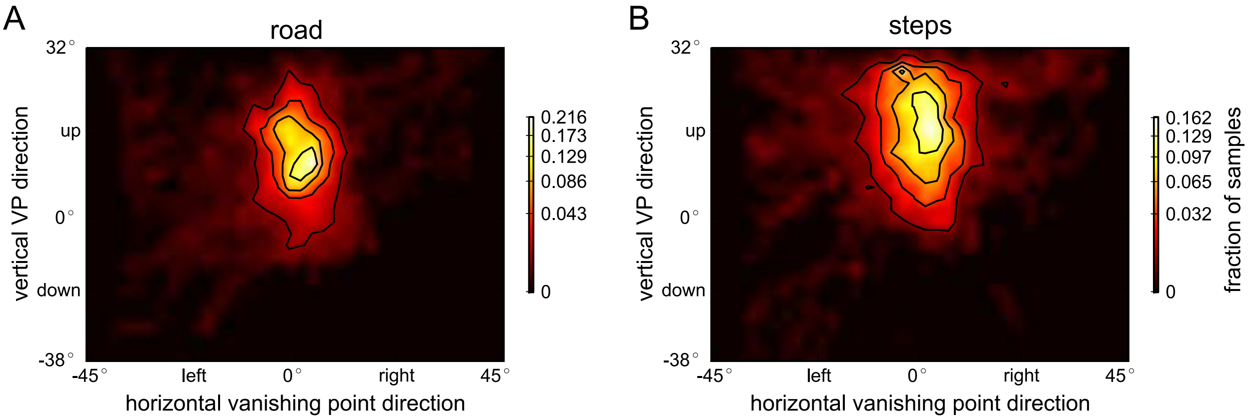

***Figure 4.6: Vanishing point distributions for head-in-world orientation.***

*Histograms of the vanishing point positions from the perspective of the head for A) road and B) steps (2.5° x 2.5° bins, interpolated for display) Data are first averaged per participant, such that each individual contributes equal amounts of data independent of walking speed. Note that in these raw representation of the data a bin more upward means that the vanishing point is higher and thus the head is pointing lower, etc.*

the majority of participants (Figure 4.7D,E), which is the opposite result as compared to the mean eye-in-head orientation (Figure 4.5 D,E). Thus – for the environment tested – eye-in-head movements mainly subserve the vertical spread of gaze, while head-in-world movements mainly subserve its horizontal spread.

For quantification of the effect of terrain on head orientation, four ANOVAs with the factors terrain and walking direction were performed. We find a main effect of walking direction (F(1,28) = 96.05, $p < .001$, Figure 4.7A-C) on the mean horizontal head-in-world orientation. There is no effect of terrain and no interaction (all $p > .17$). For the mean vertical head-in-world orientation there is an effect of terrain (F(1,28) = 242.02, $p = .001$, Figure 4.7A-C) and of walking direction (F(1,28) = 185.22, $p = .004$). There is no interaction between the two factors (F(1,28) = 19.19, $p = .323$). For standard deviations over the vertical and horizontal head-in-world orientation there are neither main effects nor interactions for any factor (horizontal: all $p > .271$; vertical: all $p > .128$; Figure 4.7D-F), showing that the head position is kept equally stable in both conditions. The effects of walking direction on mean horizontal head-in-world orientation can likely be attributed to the necessarily different choice of the vanishing point. In

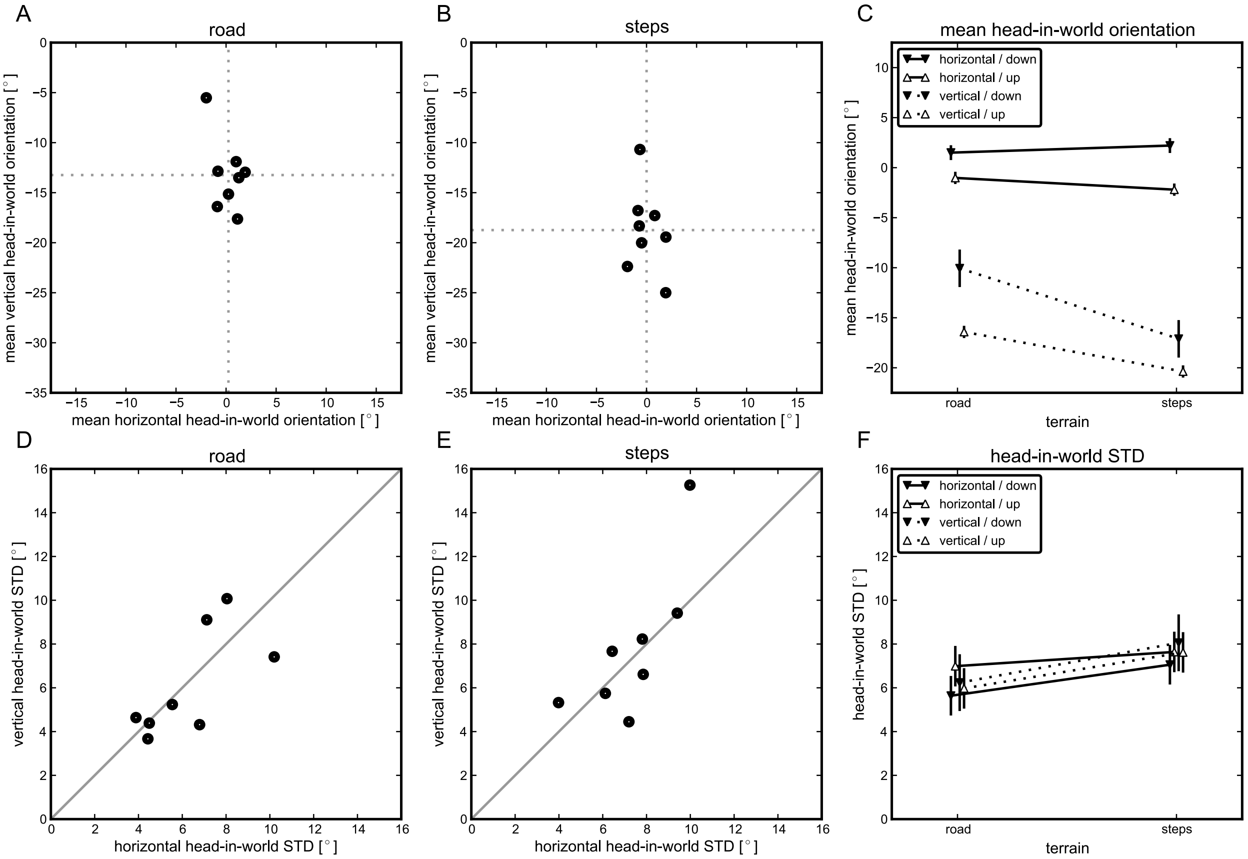

***Figure 4.7: Head-in-world orientation mean and standard deviation.***

*A-B) Mean horizontal and vertical head-in-world orientation for each individual (circles), and average (dotted lines) A) road B) steps. C) Comparison of the road and step data of panel A and B. Errorbars denote standard errors of the mean. D-E) Spread measured as standard deviation over horizontal and vertical head-in-world orientation. D) road E) steps. Note the consistently larger spread in the horizontal, in contrast to figure 3D,E. F) Comparison of the data of panels D and E, further split up for walking up and down the path. Errorbars denote standard errors of the mean. In all panels data refer to the head in the world, that is, downward implies the head pointing lower, etc.*

contrast, since the environments (and thus the vanishing point choice) are identical for both terrains (within a walking direction), the effect on the vertical component is striking. Head-in-world orientation is lower when walking on the irregular steps than when walking on the more regular road.

In sum, eye-in-head orientation and head-in-world orientation are pointed more towards the ground when the terrain is more irregular (and thus more difficult). Only for eye-in-head orientation, the vertical spread is increased for the more difficult terrain. This suggests that both eye and head subserve the adjustment of gaze for terrain negotiation. However, while eyes and head are oriented more downward for more difficult terrain, only the eyes partially compensate for this through more or larger vertical movements. Hence, head orientation presumably generically adjusts gaze according to global task-set, while the eyes still ensure a gaze component for exploration and/or path planning.

## Discussion

Our study shows distinct effects of terrain on real-world eye and head movements, when all other factors (environment, instructions) are kept constant. Most likely as an adjustment to terrain regularity, gaze is distributed differently on the two types of terrain. The contributions of eye-in-head movements as compared to head-in-world movements to gaze appear complementary. Both serve to point gaze lower when terrain gets more irregular (i.e., difficult), while only eye movements are adjusted to maintain some exploratory gaze to the upper part of the visual field.

Interestingly, a strong predominance of head movements to allocate gaze to task-relevant points is observed in driving, when a highly experienced driver is negotiating a familiar track (Land & Tatler, 2001). In this situation, eye movements get decoupled from head movements and only head movements are strongly coupled to a specific task-relevant variable. This decoupling seems to be a consequence of experience as it is not observed in a non-professional driver on a non-overtrained road (Land, 1992). Given that our observers as healthy adults have a life-time of experience with walking (though not on the particular track used), the relative flexibility of the eyes relative to the head is in line with these data. If decoupling between eye and head is a general pattern of experience there are two predictions, first, head position should be predictive of a task relevant variable (e.g., aspects of the terrain) over time, and second, the decoupling should get stronger, when a difficult terrain is negotiated repeatedly. While clearly beyond the scope of the present study, both – relating gaze and the actual walking pattern as well as training the same path – remain issues for further research.

Although eye-movement behavior on natural pictures or photographs has been extensively studied for nearly a century (Buswell, 1935; Yarbus, 1967), observations during real-life behavior are rare. Pioneering work in this direction typically dealt with specified tasks that occupied participants in full (Hayhoe & Ballard, 2005; Jovancevic-Misic & Hayhoe, 2009; Kandil et al., 2009; Land et al., 1999; Land & McLeod, 2000;) or had no control over the task (Schumann et al., 2008). Here we vary a single parameter (terrain regularity) in a common task (walking), thus transferring some of the controllability of laboratory experiments to a natural activity and setting.

One of the few studies that have investigated the relation between bodily orientation and eye-in-head movements on gaze allocation in a naturalistic task found that both the onsets and offsets of whole body movements precede those visual fixations of the task-relevant object, which in turn precede those of manipulation of the object (Land et al., 1999). This contrasts with data obtained in a more artificial, visually reduced setting, where eye movements can precede head- and body movements, although an equally tight link between these movements is observed (Hollands et al., 2004). Highlighting the importance of naturalistic settings, our finding that humans orient themselves in a generic way to the terrain by adjusting their head orientation is in line with Land et al.'s (1999) data. Depending on instantaneous terrain demands, such as the steps in this task, eye-in-head orientation is spread out to gather specific information necessary for immediate action. Hence, the role of head movements is limited to infrequent and coarse reorientations – which were apparently largely absent in this task – whereas eye-movements serve to refine gaze for immediate informational demands. In this respect, our data are a first step to transfer the data obtained under rather constrained conditions (sports, food preparation, laboratory walking tasks) to a (nearly) unconstrained environment with a real-world activity.

As the relationship between step cycle and eye-movements indicates (Hollands & Marple-Horvat, 1996; Chapman & Hollands, 2006) it is likely that eye-movements are an integral part of skilled behavior, probably embedded in visuomotor routines established over many years of experience with walking on streets. By combining our approach of unconstrained, natural behavior with systematically varied terrain difficulty and enforced or instructed eye-movement behaviors (e.g., by dynamically blocking certain parts of the visual field), it is well conceivable that adaptation of task set specific eye-movement behavior to experience can be assessed also under natural conditions and for prolonged periods. Combining our current setup with measurements of footfalls may add a temporal component to our results as shifts of gaze – by eye, head and body – then can be determined relative to the phases of the step cycle.

Besides varying the difficulty of terrain as we do here, the demands walking imposes on gaze direction may also be affected by various neurological conditions. Parkinson's disease and related syndromes, in which walking is severely impaired and performance in eye-movement tasks serves as clinically relevant maker (Corin et al., 1972; Van Koningsbruggen et al., 2009; Pinkhardt et al., 2008), exemplify this relation of concurrent impairment of gaze and gait. Treating oculomotor symptomps, in turn, can lead to improvements in walking during daily living, at least in a Parkinsonian syndrome associated with severe oculomotor impairment (Zampieri & Di Fabio, 2008). A better understanding of the roles of eye, head and body for the allocation of gaze during walking under conditions of varying difficulty may thus also eventually be of relevance for clinical applications.

Task affects eye movements (Buswell, 1935; Yarbus, 1967) and can override stimulus-related signals robustly (Henderson et al., 2007) and immediately (Einhäuser et al., 2008a). Similarly, context and environment influence gaze allocation (Torralba, 2003; Ehinger et al., 2009). In our main conditions ("road", "steps"), we held all these variables constant and only had the implicit task set given by terrain negotiation varied. Our finding of specific differences shows that parameters that are virtually impossible to mimic in the laboratory have a profound influence on gaze behavior. This underlines the importance of experiments in the real world, to quantify the extent to which psychophysical data and models remain applicable outside very constrained laboratory settings.

### *Acknowledgements*

This research was supported by the DFG (German Research Foundation) research training group DFG 885/1 to BMtH and DFG grant EI 852/1 to WE. We wish to thank Josef Stoll for his assistance in setting up the experiment and Ben Tatler for his valuable feedback on an earlier version of the manuscript.

## References

Ballard DH, Hayhoe MM (2009) Modelling the role of task in the control of gaze. *Vis Cogn* 17(6-7):1185-1204

Bardy BG., Warren WH, Kay BA (1996) Motion parallax is used to control postural sway during walking. *Exp Brain Res* 111:271-282.

Bouguet J-Y (2010) Camera Calibration Toolbox for Matlab. URL: http://www.vision.caltech.edu/bouguetj/calib_doc/ ; downloaded Dec, 13th 2010

Bruce NDB, Tsotsos JK (2009) Saliency, attention, and visual search: An information theoretic approach. *J Vis* 9(3):5, 1-24

Buswell GT (1935) *How people look at pictures: A study of the psychology of perception in art.* (Chicago: University of Chicago Press)

Callow D, Lappe M (2008) Efficient encoding of natural optic flow. *Netw Comput Neural Syst* 19(3):183-212

Carmi R, Itti L (2006) Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Res* 46(26):4333-4345

Castelhano MS, Mack ML, Henderson JM (2009) Viewing task influences eye movement control during active scene perception. *J Vis* 9(3):6, 1-15

Cavanagh PR, Higginson JS (2002) "What is the Role of Vision During Stair Descent?" In *Visual perception: the influence of H W Leibowitz*, J Andre, DA Owens, and LO Harvey eds (American Psychological Association: Washington, DC): pp 213-230

Chapman GJ, Hollands MA (2006) Age-related differences in stepping performance during step cycle-related removal of vision. *Exp Brain Res* 174:613-621

Corin MS, Elizan TS, Bender MB (1972) Oculomotor Function in Patients with Parkinson's Disease. *J Neurol Sci* 15:251-265

Cristino F, Baddeley R (2009) The nature of the visual representations involved in eye movements when walking down the street. *Vis Cogn* 17(6/7):880-903

Droll JA, Eckstein MP (2009) Gaze control and memory for objects while walking in a real world environment. *Vis cogn* 17(6/7):1159-1184

Ehinger KA, Hidalgo-Sotelo B, Torralba A and Oliva A (2009) Modelling search for people in 900 scenes: A combined source model of eye guidance. *Vis Cogn* 17(6/7): 945-978

Einhäuser W, Rutishauser U, Koch C (2008a) Task-demands can immediatley reverse the effect of sensory-driven saliency in complex visual stimuli. *J Vis* 8(2):2, 1-19

Einhäuser W, Spain M, Perona P (2008b) Objects predict fixations better than early saliency. *J Vis* 8(14):18, 1-26

Fitzpatrick RC, Wardman DL, Taylor JL (1999) Effects of galvanic vestibular stimulation during human walking. *J Physiol* 517(3):931-939

Geruschat DR, Hassan, SE, Turano, K (2003) Gaze behavior while crossing complex intersections. *Optom Vis Sci* 80(7):515-528

Guitton D, Volle M (1987) Gaze control in humans: Eye-head coordination during orienting movements to targets within and beyond the oculomotor range. *J Neurophysiol* 58(3):427-459

Hannus A, Cornelissen FW, Lindemann O, Bekkering H, (2005) Selection-for-action in visual search. *Acta Psychol (Amst)* 118(1-2):171-191

’t Hart BM, Vockeroth J, Schumann F, Bartl K, Schneider E, König P, Einhäuser W (2009) Gaze allocation in natural stimuli: Comparing free exploration to head-fixed viewing conditions. *Vis Cogn* 17(6/7):1132-1158

Hayhoe M, Ballard D (2005) Eye movements in natural behavior. *Trends Cogn Sci* 9(4):188-194

Hayhoe, M, Gillam B, Chajka K, Vecellio E (2009) The role of binocular vision in walking. *Vis Neurosci* 26:73-80

Hayhoe M, Mennie N, Sullivan B, Gorgos K (2005) The role of internal models and prediction in catching balls *Proc Conf AAAI Artif Intell 2005 Fall Symposium*.

Hausdorff JM, Yogev G, Springer S, Simon ES, Giladi N (2005) Walking is more like catching than tapping: gait in the elderly as a complex cognitive task. *Exp Brain Res* 164:541-548

Henderson JM, Brockmole JR, Castelhano MS, Mack M (2007) Visual saliency does not account for eye-movements during visual search in real-world scenes. In *Eye movement research: Insights into mind and brain*, R van Gompel M Fischer, W Murray, and R Hills, eds. (Oxford: Elsevier): pp 437-562

Hidalgo-Sotelo B, Oliva A. (2010) Person, place, and past influence eye movements during visual search. *Proc Annu Conf Cogn Sci Soc*

Hollands MA, Marple-Horvat DE (1996) Visually guided stepping under conditions of step cycle-related denial of visual information. *Exp Brain Res*, 109:343-356.

Hollands MA, Ziavra NV, Bronstein AM (2004) A new paradigm to investigate the role of head and eye movements in the coordination of whole-body movements. *Exp Brain Res*, 154:261-266

Imai T, Moore ST, Raphan T, Cohen B (2001) Interaction of the body, head, and eyes during walking and turning.

*Exp Brain Res* 136:1–18

Itti L, Baldi P (2006) Bayesian Surprise Attracts Human Attention. *Adv Neural Inf Process Syst* 19: 547-554

Itti L, Koch C (2000) A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res* 40:1489-1506

Jahn K, Strupp M, Schneider E, Dieterich M, Brandt T (2000) Differential effects ofvestibular stimulation on walking and running. *Neuroreport* 11(8):1745-1748

Jahn K, Zwergal A, Schniepp R (2010) Gait Disturbances in Old Age. *Dtsch Arztebl Int* 107(17): 306-316

Jovancevic-Misic J, Hayhoe M, (2009) Adaptive gaze control in natural environments. *J Neurosci* 29(19): 6234-6238

Kandil FI, Rotter A, Lappe M (2009) Driving is smoother and more stable when using the tangent point. *J Vis*, 9(1):11 1-11

Koch C, Ullman S (1985) Shifts in selective visual attention: Towards the underlying neural circuitry. *Hum Neurobiol* 4:219-227

Koningsbruggen MG van, Pender T, Machado L, Rafal RD (2009) Impaired control of the oculomotor reflexes in Parkinson’s disease. *Neuropsychologia* 47:2909-2915

Land MF (2004) The coordination of rotations of the eyes, head and trunk in saccadic turns produced in natural situations. *Exp Brain Res* 159:151-160

Land MF, Tatler BW (2001) Steering with the head: The visual strategy of a racing driver. *Curr Biol* 11:1215-1220

Land MF, McLeod P (2000) From eye movements to actions: how batsmen hit the ball. *Nat Neurosci* 3:1340-1345

Land M, Mennie N, Rusted J (1999) The role of vision and eye movements in the control of activities of daily living. *Perception* 1999, 28:1311-1328

Marigold DS, Patla AE (2007) Gaze fixation patterns for negotiating complex ground terrain. *Neuroscience* 144:302-313

Marigold DS, Patla AE. (2008) Visual information from the lower visual field is important for walking across multi-surface terrain *Exp Brain Res* 188(1):23-31

Navalpakkam V, Itti L (2005) Modeling the influence of task on attention *Vision Res* 45(2):205-231

Patla AE, Vickers JN (2003) How far ahead do we look when required to step on specific locations in the travel during locomotion? *Exp Brain Res* 148:133-138.

Pelz JB, Rothkopf C (2007) Oculomotor behavior in natural and man-made environments. In: Van Gompel, RPG, Fischer, MH, Murray, WS, Hill, RL (Eds.) *Eye Movements: A Window on Mind and Brain.* Elsevier: Amsterdam

Pinkhardt EH, Jürgens R, Becker W, Valdarno F, Ludolph AC, Kassubek J (2008) Differential diagnostic value of eye movement recording in PSP-parkinsonism, Richardsons's syndrome, and idiopathic Parkinson's disease. *J Neurol* 255(12):1432-1459

Pomplun M, (2006) Saccadic selectivity in complex visual search displays. *Vision Res* 46(12): 1886-1900

Rothkopf CA, Ballard DH (2009) Image statistics at the point of gaze during human navigation. *Vis Neurosci* 26:81-92

Schneider W, Shiffrin RM (1977) Controlled and Automatic Human Information Processing: I. Detection, Search and Attention. *Psychol Rev* 84(1):1-66

Schneider, E, Villgrattner T, Vockeroth J, Bartl K, Kohlbecher S, Bardins S, Ulbrich H, Brandt T (2009) EyeSeeCam: an eye movement-driven head camera for the examination of natural visual exploration. *Ann N Y Acad Sci* 1164:461-467

Schumann F, Einhäuser W, Vockeroth J, Bartl K, Schneider E, König P (2008) Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *J Vis* 8(14):12, 1-17

Shiffrin RM, Schneider W (1977) Controlled and Automatic Human Information Processing: II. Perceptual Learning Automatic Attending and a General Theory *Psychol Rev* 84(2):127-190

Startzell JK, Owens DA, Mulfinger LM, Cavanagh PR (2000) Stair Negotiating in Older People: A Review. *J Am Geriatr Soc* 48:567-580

Tatler BW, (2007) The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions *J Vis* 7(14):4 1-17

Timmis, MA, Bennett SJ, Buckley JG (2009) Visuomotor control of step descent: evidence of specialised role of the lower visual field. *Exp Brain Res* 195:219-227

Torralba A (2003) Contextual Priming for Object Detection. *Int J Comput Vis* 53(2) 169-191

Triesch J, Ballard DH, Hayhoe MM, Sullivan BT (2003) What you see is what you need. *J Vis* 3(1):9, 86-94

Treisman A & Gelade G (1980) A feature integration theory of attention *Cogn Psychol* 12:97–136

Warren WH, Jr, Kay BA, Zosh WD, Duchon AP, Sahuc S (2001) Optic flow is used to control human walking. *Nat Neurosci* 4(2): 213-216

Wolfe JM, Cave KR, Franzel SL (1989) Guided search: An alternative to thefeature integration model for visual search. *J Exp Psychol Hum Percept Perform* 15(3):419-433

Wolfe, JM (2007) Guided Search 4.0: Current Progress with a model of visual search. In *Integrated Models of Cognitive Systems* W Gray ed (New York: Oxford): pp 99-119

Yarbus AL (1967) *Eye movements and vision* (New York: Plenum Press.)

Zampieri C, Di Fabio RP (2008) Balance and Eye Movement training to Improve Gait in People With Progressive Supranuclear Palsy: Quasi-Randomized Clinical Trial *Phys Ther* 88(12):1460-1473

Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW (2008) SUN: A Bayesian framework for saliency using natural statistics. *J Vis* 8(7):32, 1-20

## Study IV

*Online action-to-perception transfer: only percept-dependent action affects perception*

Published as:

Beets, I.A.M., ’t Hart, B.M., Rösler, F., Henriques, D.Y.P., Einhäuser, W., & Fiehler, K. (2010). Online action-to-perception transfer: only percept-dependent action affects perception. Vis Res, 50(24), 2633-2641.

# Online action-to-perception transfer: only percept-dependent action affects perception

## Abstract

Perception self-evidently affects action, but under which conditions does action in turn influence perception? To answer this question we ask observers to view an ambiguous stimulus that is alternatingly perceived as rotating clockwise or counterclockwise. When observers report the perceived direction by rotating a manipulandum, opposing directions between report and percept ('incongruent') destabilize the percept, whereas equal directions ('congruent') stabilize it. In contrast, when observers report their percept by key presses while performing a predefined movement, we find no effect of congruency. Consequently, our findings suggest that only percept-dependent action directly influences perceptual experience.

## Introduction

The integration between action and perception makes up one of the most important facets of everyday life. The common coding theory (Prinz, 1997) and the theory of event coding (Hommel, Müsseler, Ascherleben and Prinz, 2001) posit that the final stages of perception and the initial stages of motor control share common representations, in which planned actions are represented in the same format as perceived events. Many studies support the idea that perception affects action (Hecht, Vogt and Prinz, 2001; McCullagh, Weiss and Ross, 1989). In addition, visual stimuli tend to dominate over perception in other modalities, even when the visual modality has no task-relevant information (e.g., Colavita, 1974; Posner, 1980; Posner, Nissen and Klein, 1976; Sinnett, Spence and Soto-Faraco, 2007). On the other hand, if perception and action share the same representation, changes due to action should lead to corresponding changes in perception (Hecht et al., 2001; Prinz, 1997; Schütz-Bosbach & Prinz, 2007 for review).

Some studies demonstrated an influence of action on perception. Previously learned movements improve visual discrimination of the same movement (Beets, Rösler and Fiehler, 2010; Casile & Giese, 2006; Hecht et al., 2001) and lead to increased cortical activity of the motor-related brain areas when observing that movement (Calvo-Merino, Glaser, Grèzes, Passingham and Haggard, 2005; Engel, Burke, Fiehler, Bien and Rösler, 2008; Reithler, van Mier, Peters and Goebel, 2007). This is not restricted to motor learning, but also applies to online interactions between the motor system and visual perception (for review, Müsseler, 1999; Schütz-Bosbach & Prinz, 2007). For example, when reaching to grasp a bar with a certain orientation, the mere motor preparation suffices to facilitate responses to a congruent visual stimulus (Craighero, Fadiga, Rizzolatti and Umiltà, 1999). Hence on various time scales – learning or online – an action can facilitate perception of a related visual stimulus.

Direct and online influence of action on the corresponding perceptual representations so far has mainly been investigated in the oculomotor system. For example, smooth pursuit eye movements can induce a distorted perception of image velocity (e.g., Freeman, Champion and Warren, 2010; Souman, Hooge and Wertheim, 2006). Moreover, eye movements necessarily change a visual stimulus in either retino- or craniocentric coordinates. Limb movements, in contrast, allow a visual stimulus to be stationary in both reference frames. It is less well understood how limb movements directly incluence motion perception.

Here we use a dynamic ambiguous stimulus, so called "perceptual rivalry", to test action-to-motion perception transfer without changing the visual input. Rivalry refers to a situation in which a constant stimulus evokes multiple perceptual interpretations that alternate over time (e.g., Leopold & Logothetis, 1999). Frequently, rivalry is induced by presenting distinct stimuli to either eye (“binocular rivalry”, Blake & Logothetis, 2002 for review). Alternatively, an ambiguous figure, such as the Necker Cube (Necker, 1832) or Rubin’s vase-faces image (Rubin, 1915), can be applied (“perceptual rivalry”). Besides in vision, rivalry has been observed in other modalities such as touch (Carter, Konkle, Wang, Hayward and Moore, 2008), audition (van Noorden, 1975), and olfaction (Zhou & Chen, 2009). Thus, rivalry seems to be a ubiquitous phenomenon covering many modalities. Rivalry is also subject to cross-modal interactions: for instance, the direction of tactile stimulation biases the perceived direction of an ambiguous visual stimulus (Blake, Sobel and James, 2004). Yet, research on how the motor system affects the perception of visual ambiguity is sparse. Since in rivalry the stimulus remains unchanged, action planning and execution cannot operate on the stimulus itself but can affect its perceptual representation. Hence, such ambiguous stimuli are ideal to test the direct effects of action on perceptual representations of motion.

In binocular rivalry, movement has indeed been found to relate to perceptual changes, in particular in the realm of oculomotor effects. On the one hand reflexive eye movements, like optokinetic nystagmus (OKN), have been used to monitor dominance in binocular rivalry (Logothetis & Schall, 1990; Sun, Tong, Yang, Tian and Hung, 2002) and are modulated by the perception of ambiguous motion (Laubrock, Engbert and Kliegl, 2008). Whether or not eye movements in turn have an influence on perceptual dominance has been a subject of debate for over a century (Necker, 1832; Einhäuser, Martin & König, 2004; Wheatstone, 1838). While the coupling between oculomotor behavior and rivalry has been studied extensively, little is known about the role of other effectors in rivalry. In one of the few studies on the effect of other effector movements on rivalry perception, Maruya, Yang and Blake (2007) used a binocular rivalry paradigm. Observers were trained to make sinusoidal hand movements when the percept of either a rotating sphere or an unrelated stimulus was dominant. The self-produced movements (which determined the speed of the stimulus motion) led to prolonged durations in the perception of the same movement and shorter stimulus suppression rates. It is possible that this visuo-motor coupling as well as intensive training may have affected these results. Furthermore, it is unknown how these findings generalize to perceptual rivalry, which shares most but not all the characteristics of binocular rivalry (van Ee, 2009).

Wohlschläger (2000) investigated the effect of manual action on perceptual rivalry presenting dots which could be perceived to be rotating clockwise or counterclockwise. In different task conditions, observers either rotated a knob by hand, or pressed a button, or planned to press a button. The frequency of the perceived movement direction was determined for each condition. Observers were more likely to perceive the stimulus move in the same direction as their planned or executed movement than in any other direction or plane. Importantly, observers' hand movements started and ended presentation of the visual stimulus, causing a confounding effect of action on perception. This pioneering study leaves the question open as to how action needs to be coupled to perception in order to exert an effect on perception.

The present study addresses this question by asking whether concurrent action influences the visual perception of a constant (ambiguous) stimulus and to what degree the motor output needs to be related to the perception in order to trigger action-to-perception transfer. Specifically, we ask whether a mere generation of actions in a predefined direction will shape perception, or whether the action needs to be functionally coupled with the current percept. Therefore, a structure-from-motion cylinder which may be perceived as rotating either clockwise (CW) or counterclockwise (CCW), is presented. We carefully distinguish between conditions in which action, the rotation of a manipulandum, is used to report the current perceptual experience from conditions in which observers perform the same movements, but unrelated to their current perceptual state. In other words, in contrast to previous studies, we present observers with a visual stimulus whose motion is independent of the observers' actions. That is, our findings are not confounded by a direct influence of action on the stimulus. While viewing this stimulus, the observers either perform predefined actions which are independent of the current percept, or actions which depend upon the current perceptual state. To investigate the effects of action, we determine the duration that one percept dominates, i.e., percept stability.

The main experimental conditions in the present experiment follow a 2*2 design with the factors movement type (percept-dependent movements vs. percept-independent, predefined movements), and congruency (percept-congruent movements vs. percept-incongruent movements). We measure how long observers stay in one perceptual state ("dominance durations"). If movements per se affect the perceptual state, we hypothesize changes in dominance durations for predefined movements, perception-dependent movements and even for unrelated vertical movements. If, however, action must depend upon perception to trigger action-to-perception transfer, there should be no or little effect of congruency on dominance durations during predefined movements. Dominance durations during during movements depending on the perceptual state should then be the only ones affected by congruency.

## Materials and Methods

### *Observers*

Seventeen naïve observers participated in the study. Data from three observers was excluded due to technical reasons: one observer aborted the experiment; in another, the movement data were not usable due to a technical problem; and another failed to comply with task instructions. Before analysing the data, we tested observers' ability to perform the task using congruent and incongruent tracking of an unambiguous stimulus in the 'catch blocks' (see Procedure). Out of the fourteen observers that provided a usable dataset, three were excluded due to low performance in these catch blocks (see Results). Data from the remaining eleven observers between the ages of 20 and 27 years (mean age: 23.5 ± 2.5 years; 4 male / 7 female) was used for analysis. These observers had normal or corrected-to-normal vision, were right-handed as assessed by a German translation of the Edinburgh Handedness Inventory (mean ± standard deviation: 89.1 ± 12.5) (Oldfield, 1971), and had no history of psychiatric or neurological disorders. All observers were recruited from the Philipps-University Marburg, and were compensated with course-credits or money (€6 per hour) for their participation. Written informed consent was obtained, and the procedure was in accordance with the ethical standard laid down in the Declaration of Helsinki (2000) as well as with departmental guidelines.

### Stimuli

Four-hundred white dots of ~0.065° * ~0.065° were presented within an aperture of ~2.86° * ~6.53° on a 1024*768 pixel, 16” black screen (refresh rate 75Hz) to perceptually induce the shape of a rotating cylinder (structure-from-motion) (fig. 5.1a). The cylinder made one full revolution every 3.6s. Dot life-time was set at 0.3s. This ambiguous structure-from-motion stimulus produced a percept of a cylinder, switching between CW and CCW rotation.

For some conditions, we created an unambiguous version of the stimulus. A red bar of ~0.16° * ~8.16° was drawn over and rotated along with the cylinder. When moving along the ‘back’ of the cylinder, the bar was partially occluded. To enhance disambiguation, the dots at the back were fully occluded.

### Apparatus

Stimuli were viewed through a black cardboard tunnel with a length of 110 cm to prevent interference from other visual input (fig. 5.1b). Observers’ distance to the monitor was ~ 110 cm. A black cloth covered the back of the head and part of the tunnel to prevent observers from watching their own movements. Observers were instructed to direct their gaze toward the centre of the stimulus and to try seeing the stimulus as a whole. A manipulandum with a turntable on the horizontal plane was used to perform actions during perception of the ambiguous cylinder (fig. 5.1b). Observers rotated the turntable using the attached vertical handle with an effective radius of 5 cm. In the motor conditions (see procedure), observers sat facing the screen and grasped the vertical handle of the manipulandum with a precision grip using their thumb, index and middle finger of the right hand (fig. 5.1b). The perception of the direction of motion of the visual stimulus was indicated by either moving the manipulandum or by pressing one of two arrow keys (left arrow key for CW; right arrow key for CCW) with the left hand (see procedure). For the unrelated movement condition (see Procedure), a freely movable stylus was used to execute straight vertical trajectories. The stylus was 78 mm long and had a diameter of 15 mm and was held between the thumb and fingers with the same precision grip as used for the manipulandum handle and was moved between an upper and a lower stopper mounted on the right side of the tunnel. A chinrest was used to keep a stable head position throughout the experiment. The chair and chinrest were adjusted individually to assure a comfortable position.

Movement trajectories were recorded with an ultrasound motion recording device (ZEBRIS CMS20, Zebris Medical GmbH, Isny im Allgäu, Germany). To measure hand movements, a sensor was attached to the top of the vertical handle of the turntable or to the top of the stylus. The movement data was sampled with 100 Hz and analyzed offline.

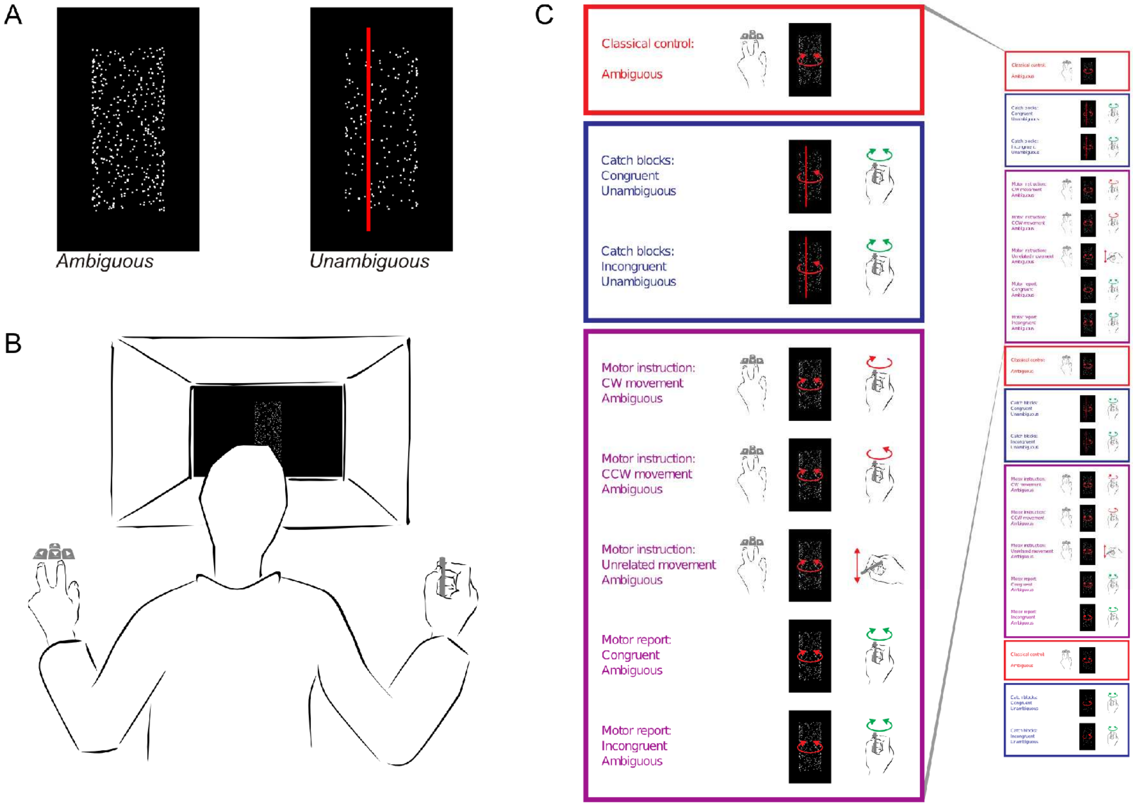

***Figure 5.1. Stimuli, setup and conditions.***

*(A) Visual structure-from-motion stimuli which observers viewed through the tunnel. Left: The ambiguous stimulus could be interpreted as a cylinder rotating CCW or CW. Right: The unambiguous stimulus over which a red bar was drawn. (B) Setup. Observers sat in front of a tunnel through which the visual stimuli were presented by which the self-produced movements were occluded. Observers pressed one of the arrow keys with the index and ring finger of the left hand. The right hand was used for rotating the turntable, or to make movements along the vertical plane of the right side of the tunnel (not shown). (C) Conditions. Within each colored frame, the blocks were randomized. A green arrow above the right hand indicates that the manipulandum was used to indicate perceptual state. The block order is illustrated on the right.*

### Procedure

The unambiguous stimulus (fig. 5.1a, right) was used only for a control condition ('catch blocks', fig. 5.1c, blue frame) to investigate motor behavior, whereas the ambiguous stimulus (fig. 5.1a, left) served to investigate the durations of the dominating percept (CW or CCW rotation) in all other conditions. There were two kinds of report modes: a key press and the rotation of the manipulandum. In the case of key presses, observers held the key corresponding to the percept, until it switched. In all conditions that involved moving the manipulandum (fig. 5.1c, blue and purple frames), observers were asked to match their velocity with that of the cylinder. When observers were not sure about the rotational direction of the ambiguous stimulus, they were asked to press no key in case of keyboard report, and not to move in case of manipulandum report.

The experiment consisted of eight conditions (fig. 5.1c). The main experimental conditions of interest used the ambiguous stimulus (fig. 5.1a) and were organized into a 2*2 design. In these conditions the effects of movement type (instructed vs. percept dependent movements) and congruency (actions and perceived motion in equal vs. opposite direction) were investigated. The first two conditions of interest were the ‘motor instruction’ blocks in which observers rotated the manipulandum either CW or CCW throughout the block regardless of percept, resulting in ‘motor instruction CW’ and ‘motor instruction CCW’ blocks (fig. 5.1c, purple frame, first two conditions). The action performed was thus independent of the perceptual interpretation of the visual stimulus. Concurrently, observers indicated using the keyboard with the left hand, which percept was currently dominating. The effect of congruency was later investigated by splitting dominance durations into percepts that were congruent with the instructed movement and percepts that were incongruent with the instructed movement (when active movements were CCW but cylinder perception was CW, or vice versa). The other two conditions of interest were the ‘motor report’ blocks in which the manipulandum was rotated either CW or CCW, depending upon the current perceptual interpretation of the visual stimulus (fig. 5.1c, last two conditions in the purple frame). Instead of using the keys to report the percept, the percept was reported by rotating the manipulandum in the same direction as the visually perceived rotation in the ‘congruent motor report’ condition (fig. 5.1c, purple frame, 4th condition), or in the opposite direction from the visually perceived rotation in the ‘incongruent motor report’ condition (fig. 5.1c, purple frame, 5th condition). The performed action was thus dependent upon the perceptual interpretation of the visual stimulus. In a fifth experimental condition (fig. 5.1c, purple frame, middle condition), the effect of movement per se was investigated by executing movements unrelated to the stimulus (‘motor instruction unrelated’). Here, ongoing vertical movements (i.e., unrelated to the rotational axis of the visual stimulus) were made along the vertical axis of the tunnel using the stylus. Simultaneously, key presses were used to indicate rotation direction of the ambiguous stimulus.

The other conditions served as control conditions to obtain a baseline measurement of perceptual dominance durations (‘classical control’) and to test if observers were able to perform the task equally well when reporting a percept by using congruent or incongruent rotation of the manipulandum (‘catch blocks’). In the classical control condition (fig. 5.1c, red frame), the ambiguous cylinder stimulus was viewed while the observer indicated by key presses in which direction the ambiguous stimulus rotated. During the catch blocks (fig. 5.1c, blue frame), observers viewed an unambiguous cylinder stimulus and were instructed to rotate the manipulandum; either along with the stimulus in the congruent catch blocks or to rotate in the opposite direction from the stimulus in the incongruent catch blocks. The rotational direction of the red bar and the cylinder changed repeatedly within each block. To make the task, and the experience of switches in the cylinder comparable to the ‘motor report’ blocks, the durations per rotation direction were determined by the observers’ own shuffled dominance durations from the preceding ‘classical control’ block (with all dominance durations shorter than 500ms removed). No key presses were made. Since the timing of ‘switches’ was known in these blocks, they were suitable as a baseline measure of the ability to report switches of percept equally well for congruent and incongruent blocks.

Before starting the experiment, observers were familiarized with the procedure and the stimulus by performing each of the eight different conditions for one minute. The experiment consisted of 19 blocks lasting 5 minutes each. In between blocks, there was an opportunity to take a break. The order of the blocks was as follows (see fig. 5.1c): the experiment started with the classical control after which the unambiguous catch blocks were performed. The order of congruent and incongruent catch blocks (fig. 5.1c, blue frame) was counterbalanced over observers. Then, all five experimental conditions (fig 5.1c, purple frame) were performed in a randomized order. Finally, this sequence was repeated and a second repetition of classical control and unambiguous catch blocks was performed at the end of the experiment. Thus, the experiment consisted of two sets of experimental blocks surrounded by three sets of control blocks at the beginning, in the middle, and at the end of the procedure. The three sets of control blocks allowed the effect of time-on-task on dominance durations to be quantified. Within each colored frame in figure 5.1c, the order was randomized (the order in the figure serves as an example) but held constant for repetitions within observer.

### *Movement data pre-processing*

Since observers' movement trajectories were constrained by the manipulandum to a circular movement with a constant radius, we have a one-dimensional movement given by the angle as a function of time. The direction of this movement (counterclockwise or clockwise) corresponds to the reported percept in the catch blocks and the two motor report conditions. In order to extract motion direction and velocity from the raw manipulandum position data, the data was pre-processed in Python (Version 2.6.5) using Numpy (Oliphant, 2007) and SciPy (Jones et al., 2001). Due to measurement noise, some samples fell out of the radius, which could be misinterpreted as a perceptual switch. Therefore, we discarded samples whose Euclidian distance deviated more than 3 standard deviations from the mean with respect to the previous sample. Cubic splines on the remaining data were used to interpolate the discarded samples. A circle was fitted to the samples which allowed converting the position data to angles. This signal was smoothed using a 5-sample median filter before conversion to an angular velocity signal and extracting the perceptual states indicated by the observers.

### *Data analyses*

Dominance durations for CW and CCW percepts were extracted from the keyboard data in the classical control and motor instruction blocks. The dominance duration was the period of time that exactly one key was held down. Periods in which no key or two keys were simultaneously pressed were discarded. When one percept was interrupted by a short period in which both keys were pressed, the percept was separated and thus resulted in two dominance durations (plus the short period of discarded data). In 2.4% of the time across blocks in which the task was executed, either no key or two keys were pressed. These intervals were discarded from analysis as the dominant percept could not be determined. Dominance durations were extracted from the manipulandum movement data for the unambiguous catch blocks and the motor report blocks. Velocities below a threshold of 1°/s were counted as no movement. From the classical control condition, we defined for each observer a threshold as the first half percentile of dominance durations; we discarded values below this threshold to remove jitter in the motor report conditions. Due to these differences in extracting dominance durations from key press and manipulandum data, any direct comparisons between key-press report and manipulandum-report data should be interpreted with caution.

Besides dominance durations in the ambiguous stimulus blocks, movement characteristics were investigated in the catch blocks and the motor report blocks. In the catch blocks, we determined root mean-squared error (RMSE) from the required speed to check whether congruent and incongruent reports were comparable. In the motor report blocks, the acceleration was compared at the moment of direction change between congruent and incongruent movements.

### Statistical tests

Since dominance durations in rivalry typically follow leptokurtic (heavy-tailed) distributions (e.g., Logothetis, 1998), we use medians (rather than means) to characterize the distribution of dominance durations per observer and block. Across observers, however, the median dominance durations can safely be assumed to follow a Gaussian distribution such that statistics could be performed with standard parametric tests. First, we conduct a 2*2 ANOVA to investigate the effect of movement type (motor instruction vs. motor report) and action-perception congruency (congruent vs. incongruent). For testing effects directly between conditions, pairwise t-tests and repeated measures ANOVA (for testing effects over multiple blocks in the classical control condition) were conducted. All statistics were computed using R (Version 2.10.1; R Development Core Team, 2009) maintaining a critical alpha level of 0.05. Results
The question addressed in our study was to what extent action needs to be coupled to perception to cause perceptual changes. More specifically, we investigated how concurrent actions, congruent or incongruent with perception, influence processes underlying perceptual rivalry in ambiguous structure-from-motion stimuli.

## Results

### *Catch blocks*

To test whether observers could veridically report their percepts by rotating the manipulandum, we used a disambiguated version of the rotating cylinder. To obtain an accuracy measure, we calculated the mean response time (RT) to a switch of rotation direction. This was done by dividing the total time observers rotate opposite from the required direction by the number of direction switches given by the stimulus. Most observers' average RT's were in the range 0.28s – 0.51s for congruent catch blocks, although one subject had an average RT of 2.34s. In the incongruent catch blocks most observers had an average RT in the range of 0.17s – 0.91s, whereas two had an average RT of 2.99s and 8.72s. The three observers with very high RT's were excluded from all further analyses as the reliability of their reports in the motor report condition cannot be guaranteed (see Table 1 in Appendix A for their median dominance durations). For the remaining 11 observers the RTs were 0.40s ± 0.15s (mean ± SD over observers) for incongruent catch blocks and 0.36s ± 0.05s for congruent catch blocks. These did not differ significantly ($t(10) = 1.170$, $p = 0.269$) and represent a typical response time. Speed accuracy as measured by RMSE from the goal angular velocity was 62.4 °/s ± 26.1°/s in the congruent and 64.7°/s ± 19.4°/s in the incongruent catch blocks, which did not differ significantly ($t(10) = 0.489$, $p = 0.636$). Both RT and RMSE show that the 11 remaining observers performed the task correctly and reported movement directions with the manipulandum equally well for congruent and incongruent movements in the catch blocks. This strongly suggests that observers also performed equally well in the congruent and incongruent motor report conditions.

### *Experimental conditions*

To investigate the effect of movement type and congruency, a 2*2 ANOVA was conducted. The main effects of movement type and congruency were not significant ($F(1,10) = 0.161$, $p = 0.697$; $F(1,10) = 4.247$, $p = 0.066$, respectively), suggesting that dominance durations did not differ between motor instruction and motor report conditions nor between congruent and incongruent movements in general. The borderline significant main effect of congruency is probably due to the effect of congruency on motor report dominance durations. Indeed, the two factors interacted significantly ($F(1,10) = 7.801$, $p = 0.019$), showing a differential effect of congruency between both movement types (fig. 5.2, right). To explore this interaction more closely, we examined the results of both the motor instruction and motor report conditions in more detail.

In the motor report conditions, observers were asked to report their percept with the movement of the manipulandum. In one condition observers were instructed to move the manipulandum in the same direction as their percept (“congruent motor report”), and in the opposite direction in the other condition (“incongruent motor report”). In these conditions (fig. 5.2, dashed line) percept durations were significantly shorter for incongruent movements than for congruent movements ($t(10) = -2.522$, $p = 0.030$). This shows that percept-related action affects the perceived direction of ambiguous stimuli.

When observers rotated the manipulandum irrespective of the perceived motion, they reported their percept by key presses. We separated the data according to times when

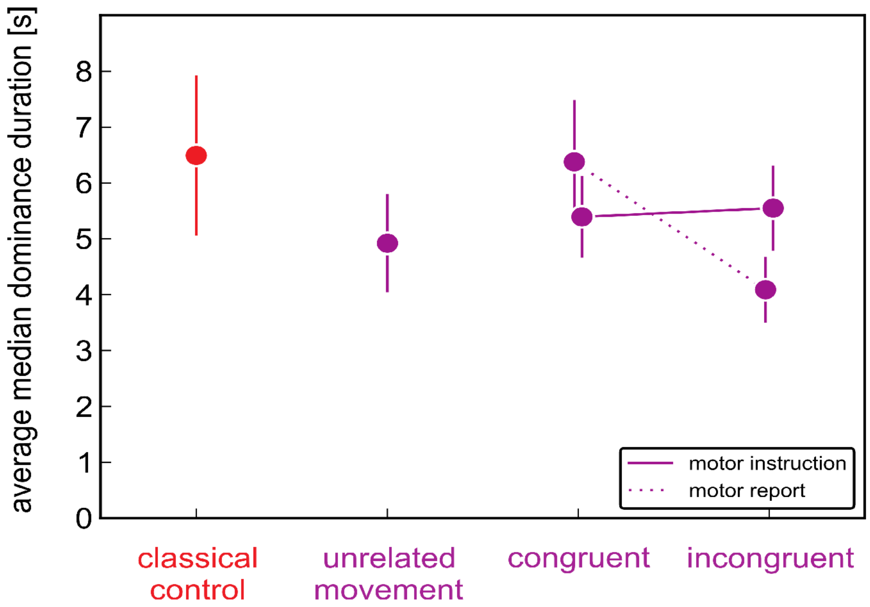

***Figure 5.2. Dominance durations per condition in seconds.***

*Median dominance durations for classical control, motor instruction (CW and CCW are split into congruent / incongruent and are connected by a solid line) and motor report conditions (congruent and incongruent are connected by a dashed line). The error bars represent standard errors of the mean.*

manipulandum movement and perceived motion were in the same ("congruent motor instruction") or in the opposite ("incongruent motor instruction") direction (fig 5.2, solid line). Dominance durations did not differ significantly between incongruent and congruent movements in these conditions ($t(10) = 0.509$, $p = 0.621$; table 1 in Appendix A). These dominance durations also did not differ from a condition in which observers performed an unrelated movement perpendicular to the table (comparison to congruent movements: $t(10) = -1.023$, $p = 0.331$; comparison to incongruent movements: $t(10) = -1.189$, $p = 0.262$). Nor did the motor instruction conditions differ from a condition in which no manipulandum movement was required (congruent vs. classical control: $t(10) = 1.295$, $p = 0.224$; incongruent vs. classical control: $t(10) = 0.927$, $p = 0.376$; unrelated vs. classical control: $t(10) = -1.684$, $p = 0.123$). In summary, none of the movements that were conducted irrespective of the current perceptual state exerted an influence on the percept duration.

These results show that the dominance durations are not affected by congruency in the motor instruction condition, that is, when predefined movements are executed independent of the perceptual experience. In the motor report condition, however, dominance durations are affected by congruency suggesting that only actions which are dependent on the current percept can influence visual perception.

### *Classical control condition*

The median dominance duration in the 'classical control' blocks (where no movements except for key presses are executed) was 6.49s ± 4.99s. In line with earlier findings (Nawrot & Blake, 1991; Blake et al., 2004), none of the observers showed a significant bias toward CW (48.5% ± 5.7%) or CCW (51.4% ± 5.7%) percepts. When the longer median dominance durations of all observers were taken and tested against all shorter median dominance durations, no significant difference was found (t(10) = 1.476, p = 0.170). Furthermore, dominance durations were stable across repetitions (F(2,10) = 2.271, p = 0.129). This verifies that pooling dominance durations from both percepts and across repetitions for all other analyses is justified.Direction transitions in motor report conditions

To verify whether transitions were similar for reporting percept by congruent and by incongruent movements using the manipulandum, we investigated the change in direction of the movement data in the motor-report conditions. To this end, we aligned all movement traces to the time of transition between the two rotation directions (fig. 5.3). Visual inspection of the velocity traces (fig. 5.3a, b) suggest that the velocity profile is smooth and is comparable between conditions.

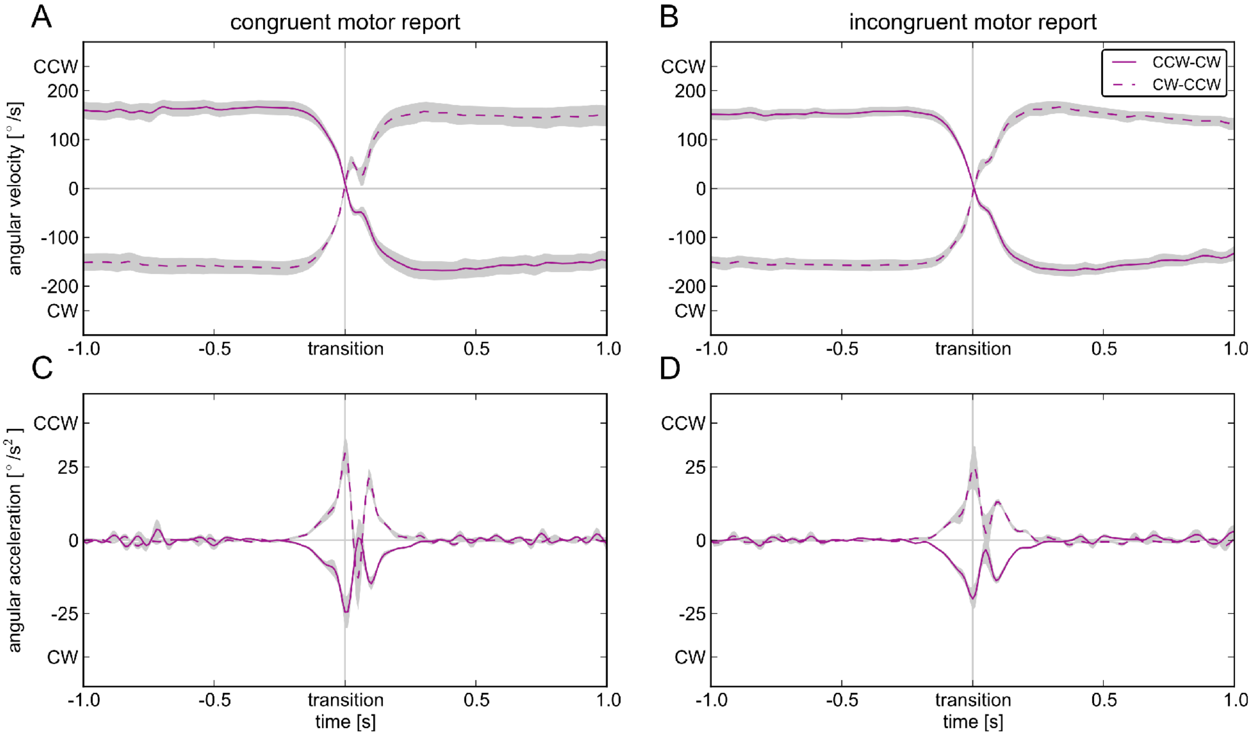

***Figure 5.3. Movement transitions***

*Movement trajectories were aligned to time of perceptual transitions (defined as zero-crossings of the angular velocity) in motor-report conditions; positive values denote CCW movement, negative CW movement; solid lines denote mean velocities across observers for switches from CCW to CW, dashed lines from CW to CCW; shaded areas represent standard error of the mean. (A) Speed in the motor-report condition in which percept was indicated by congruent movement. (B) Speed in the motor-report condition in which percept was indicated by incongruent movement. (C) Acceleration in the motor-report condition in which percept was indicated by congruent movement. (D) Acceleration in the motor-report condition in which percept was indicated by incongruent movement.*

To quantify this, we investigated the acceleration (i.e., the derivative of speed) on the moment of the transition, and compared this between conditions (fig. 5.3c, d). We found that acceleration did not differ between congruent and incongruent motor report conditions (*F*(1,36) = 1.316, *p* = 0.259), nor between transition types (i.e., from CW to CCW and from CCW to CW) (*F*(1,36) = 0.658, *p* = 0.422), nor was there an interaction between transition type and condition (*F*(1,36) = 0.070, *p* = 0.792). Hence, our findings that dominance durations were shorter in the incongruent motor report condition than in the congruent one cannot be explained by a difference in motor performance in the two conditions.

## Discussion

Our results show that action shapes perception, but only when the action is dependent on the current percept. When observers use rotational movements to indicate their percept of an ambiguous stimulus, percept durations change significantly. In contrast, rotating in a predefined direction does not lead to changes in percept durations in the same visual stimuli.

In previous studies (Maruya et al., 2007; Wohlschläger, 2000), it has been shown that predefined movements influence the visual interpretation of ambiguous stimuli. In these experiments, however, observers' movements initiated and terminated the movement of the stimulus. Furthermore, in Maruya et al. (2007), observers were trained to make movements in order to drive the speed of the visual stimulus. Thus, in these studies action had a direct effect on the perceptual form of the stimulus which may have led to a tight interplay of action and perception through stimulus manipulation, rather than a direct effect of action on perceptual representations. Here, in contrast, stimulus presentation was always independent of observers' actions, allowing us to compare task conditions in which the executed movements were independent of or dependent on percept. Our results clearly show that a direct effect of action on perception requires the action to be percept-related. The stability of percept is affected by congruency only in percept-related actions, in which congruent movements stabilize the percept and incongruent movements destabilize the percept.

Recent studies have demonstrated that rivalry elicited in one sensory modality can be altered by other sensory modalities. In these cases the perception of the ambiguous stimulus is biased towards the percept consistent with the non-ambiguous modality (Blake et al., 2004; van Ee, van Boxtel, Parker and Alais, 2009). Here we confirm that not only other modalities but also action influences rivalry (Maruya et al., 2007, Wohlschläger, 2000). Beyond these earlier studies, our findings demonstrate that motor effects on rivalry are specific to movements that relate to the percept. The similarity between the effect of other modalities and action may provide a link between two seemingly distinct fields: common coding theory (Prinz, 1997) or the theory of event coding (Hommel et al., 2001) on the one hand and multisensory processing (e.g., Alais & Burr, 2004; Ichikawa & Masakura, 2006; Repp & Knoblich, 2007; Sekuler, Sekuler and Lau, 1997; Shimojo & Shams, 2001; Witten & Knudsen, 2005) on the other hand. In cross-modal rivalry, it seems that if the unambiguous modality provides a signal converging with the ambiguous modality this stabilizes the interpretation of the visual input, whereas two diverging signals destabilize it. One of the signals accompanying movement execution is somatosensory (re)afferences, which may have the same function. For example, passive motor training, which in large part relies on reafferent information, can lead to the acquisition of new motor skills (Beets et al., 2010). The sensory information accompanying active movement execution could thus have contributed to the effects on visual perception. To what extent efferent vs. afferent information contributes to action-to-perception transfer remains an interesting topic for future research.

While there has been little research on the effect of hand movements on rivalry, many studies have addressed the relationship between *eye* movements and rivalry. Over 175 years after

Necker's (Necker, 1832) original proposal that perceptual switches of his eponymous cube were a consequence of "the adjustment of the eye for obtaining distinct vision" (Necker, 1832, p 336-337), a wide consensus on a coupling between eye movements and perceptual dominance seems to exist (e.g., Brouwer & van Ee, 2006; Laubrock et al., 2008; Toppino, 2003; van Dam & van Ee, 2005), although the direction of causality is still in debate (Ellis & Stark, 1978; Eure, Hamilton and Pheiffer, 1956; Kawabata, Yamagami and Noaki, 1978; Zimmer, 1913) and is likely to be bi-directional (Einhäuser et al., 2004). In the context of (visual) rivalry, oculomotor behavior brings two additional challenges: first, any eye movement has a direct impact on the retinal stimulus; second, eye movements are coupled to shifts in focal attention, which itself influences switch rates (Paffen, Alais and Verstraten, 2006). Despite all the advantages of the oculomotor system acting as the interface between input and output (i.e., between perception and action) to test how action influences perceptual representations while minimizing other factors (stimulus, focal attention), manual movements, as used here, circumvent these potential confounds.

Since attention speeds up rivalry (Paffen et al., 2006) and this increase in speed is not restricted to one modality (Alais, van Boxtel, Parker and van Ee, 2010), we have to ask whether our results can be explained by attention alone. One may argue that reporting by incongruent tracking is more difficult and thus requires more attentional resources which would consequently speed up switching between percepts. We consider this explanation unlikely for several reasons. First, one can also argue for the opposite with equal justification: incongruent action requires more attention, thus less attention is available for perception and thus rivalry should slow down, contrary to our findings. Second, we failed to find any differences in dominance durations between classical control and unrelated movements on the one hand, and between dominance durations in predefined incongruent or congruent movements (i.e., percept unrelated) on the other hand. This implies that movement *per se* is not an attentionally challenging task. Third, for unambiguous stimuli, movement characteristics and errors between congruent and incongruent tracking were very similar, again arguing against a different attentional effect on both. However, it is undisputable that attention plays a key role in rivalry. We argue, however, that there is no differential effect of attention on incongruent and congruent movements, and consequently, our main finding cannot be explained solely by differences in attentional demand. As binding diverse representations is a main function of attention in the sensory domain (Wolfe & Bennett, 1997), it seems conceivable that attention is a key ingredient to bind sensory and motor representations. This implies that in certain cases, the common coding framework only applies when additional attention is given to corresponding movements of an effector. Beyond a potential impact of attentional processes, our findings provide support for the common coding concept and refine this model by demonstrating that action-to-perception transfer requires the action to be directly coupled to motion perception.

The common coding theory (Prinz, 1997) and the theory of event coding (Hommel et al., 2001) state that action and perception share common representational domains. Therefore action and perception reciprocally influence each other. Although this theory has been supported by empirical data that demonstrate a bidirectional link between action and perception (Hecht et al., 2001) and direct effects of action on perception (e.g., Beets et al., 2010; Casile & Giese, 2006; Craighero et al., 1999; Wohlschläger, 2000), it is unknown to what extent action-to-perception transfer is dependent on percept-related action. Our results show that action can only influence perception when it acts on the perceptual representations, i.e., a mere generation of an action is insufficient to trigger a transfer from action to perception. Action planning in relation to the stimulus thus seems to be crucial to induce binding between action and perception (Hommel, 2004). When an action does not need to be integrated with a visual stimulus in order to perform the task, this effect is absent. In summary, common coding of a stimulus and an action seems to occur only when they are directly relevant to each other and the predicted effects of action on perception can only occur when this is the case. This fits with the prediction that perception and action planning can only interact when they refer to the same feature of the motor system (Hommel et al., 2001).

Future research will determine to what extent action-to-perception transfer can still occur when for example, the axes involving action and perception are at odds (e.g., diagonal vs. vertical). In summary, this study demonstrates for the first time that action and perception need to be functionally coupled in order to affect each other. Given that people make movements within a continuously changing and moving environment, the notion that only actions that are relevant for the perceived events can influence the perception of these events, is likely the most efficient strategy for human behavior.

### *Acknowledgments*

This research was supported by the DFG (German Research Foundation) research training group DFG 885/1 “Neuronal representation and action control” to IAMB and BMtH; DFG grant EI 852/1 to WE; DFG grant FI 1567 to KF and FR; by the research unit FOR 560 “Perception and Action” to KF and FR; by the TransCoop-grant of the Alexander-von-Humboldt Foundation to KF and DYPH. DYPH is an Alfred P Sloan fellow. We also wish to thank Johanna Kues for help in data-acquisition.

## Appendix A

To illustrate the large inter-observer differences in dominance duration and to provide a condensed version of the data to the interested reader, all median dominance durations and their standard deviations in the experimental conditions and the classical control conditon are listed in Table 1.

**Table 1. Dominance durations per observer.**

| Observer | Classical control | Motor report | | Motor instruction | | |
|---|---|---|---|---|---|---|
| | | congruent | incongruent | congruent | incongruent | unrelated |
| 1 | 4.92 ± 8.07 | 3.58 ± 6.03 | 6.26 ± 6.22 | 5.98 ± 8.07 | 6.09 ± 6.81 | 6.97 ± 5.41 |
| 2 | 4.22 ± 7.08 | 7.34 ± 13.97 | 3.00 ± 4.64 | 4.36 ± 7.64 | 5.08 ± 5.86 | 4.56 ± 5.53 |
| 3 | 14.31 ± 33.61 | 14.90 ± 16.16 | 7.23 ± 12.21 | 8.04 ± 26.63 | 6.12 ± 12.87 | 5.20 ± 11.99 |
| 4 | 7.46 ± 7.19 | 7.41 ± 8.06 | 4.79 ± 6.20 | 6.38 ± 8.63 | 6.93 ± 9.20 | 7.14 ± 3.76 |
| 5 | 0.92 ± 16.38 | 2.61 ± 10.84 | 1.53 ± 7.87 | 2.26 ± 18.22 | 3.85 ± 19.42 | 1.91 ± 15.72 |
| 6 | 10.08 ± 12.40 | 7.97 ± 18.81 | 6.27 ± 18.02 | 8.40 ± 15.07 | 10.28 ± 11.44 | 6.25 ± 7.10 |
| 7 | 4.22 ± 15.92 | 2.72 ± 3.67 | 2.83 ± 5.26 | 6.35 ± 9.55 | 5.15 ± 9.65 | 5.25 ± 21.20 |
| 8 | 2.01 ± 2.76 | 2.01 ± 5.68 | 2.04 ± 1.92 | 1.67 ± 4.54 | 1.55 ± 2.07 | 1.76 ± 2.71 |
| 9 | 16.93 ± 18.74 | 9.06 ± 10.36 | 2.54 ± 5.24 | 9.88 ± 16.36 | 9.91 ± 14.77 | 12.61 ± 16.52 |
| 10 | 7.85 ± 31.71 | 6.31 ± 10.25 | 2.82 ± 6.17 | 8.17 ± 35.23 | 21.74 ± 18.30 | 9.30 ± 20.40 |
| 11 | 4.01 ± 5.02 | 4.50 ± 5.39 | 2.92 ± 3.45 | 4.53 ± 4.31 | 4.50 ± 3.86 | 3.83 ± 3.73 |
| 12 | 3.97 ± 6.23 | 5.48 ± 5.78 | 3.49 ± 3.60 | 4.59 ± 4.26 | 4.03 ± 4.68 | 4.01 ± 4.67 |
| 13 | 5.18 ± 8.94 | 4.62 ± 4.91 | 4.70 ± 4.24 | 3.84 ± 4.43 | 3.31 ± 3.46 | 2.57 ± 2.59 |
| 14 | 5.52 ± 4.86 | 5.68 ± 5.79 | 4.04 ± 4.17 | 5.71 ± 8.77 | 6.42 ± 4.10 | 4.32 ± 4.06 |

*Values are median dominance duration in seconds. ± SD gives the standard deviation within each observer. Observers marked in gray did not perform well in the catch blocks and their data were left out of the analyses, but are included here for the interested reader.*

## References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*, 257-262.

Alais, D., van Boxtel, J. J., Parker, A., & van Ee, R. (2010). Attending to auditory signals slows visual alternations in binocular rivalry. *Vision Research, 50*, 929-935.

Beets, I. A. M., Rösler, F., & Fiehler, K. (2010). Non-visual motor training affects visual motion perception: Evidence from violating the two-thirds power law. *Journal of Neurophysiology*. doi:10.1152/jn.00974.2009.

Blake, R., & Logothetis, N. K. (2002). Visual competition. *Nature Reviews Neuroscience, 3*, 13-21.

Blake, R., Sobel, K. V., & James, T. W. (2004). Neural synergy between kinetic vision and touch. *Psychological Science, 15*, 397-402.

Brouwer, G. J., & van Ee, R. (2006). Endogenous influences on perceptual bistability depend on exogenous stimulus characteristics. *Vision Research, 46*, 3393-3402.

Calvo-Merino, B., Glaser, D. E., Grezes, J., Passingham, R. E., & Haggard, P. (2005). Action observation and acquired motor skills: an FMRI study with expert dancers. *Cerebral Cortex, 15*, 1243-1249.

Carter, O., Konkle, T., Wang, Q., Hayward, V., & Moore, C. (2008). Tactile rivalry demonstrated with an ambiguous apparent-motion quartet. *Current Biology, 18*, 1050-1054.

Casile, A., & Giese, M. A. (2006). Nonvisual motor training influences biological motion perception. *Current Biology, 16*, 69-74.

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics, 16*, 409-412.

Craighero, L., Fadiga, L., Rizzolatti, G., & Umiltà, C. (1999). Action for perception: a motor-visual attentional effect. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 1673-1692.

Einhäuser, W., Martin, K. A., & König, P. (2004). Are switches in perception of the Necker cube related to eye position? *European Journal of Neuroscience, 20*, 2811-2818.

Ellis, S. R., & Stark, L. (1978). Eye movements during the viewing of Necker cubes. *Perception, 7*, 575-581.

Engel, A., Burke, M., Fiehler, K., Bien, S., & Rösler, F. (2008). Motor learning affects visual movement perception. *European Journal of Neuroscience, 27*, 2294-2302.

Eure, S. P., Hamilton, C. B., & Pheiffer, C. H. (1956). Reversible figures and eye-movements. *American Journal of*

*Psychology, 69*, 452-5.

Freeman, T.C.A., Champion, R.A., & Warren, P.A. (2010). A Bayesian model of perceived head-centered velocity during smooth pursuit eye movement. *Current Biology, 20*, 757-762.

Hecht, H., Vogt, S., & Prinz, W. (2001). Motor learning enhances perceptual judgment: a case for action-perception transfer. *Psychological Research, 65*, 3-14.

Hommel, B. (2004). Event files: feature binding in and across perception and action. *Trends in Cognitive Sciences, 8*, 494-500.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The Theory of Event Coding (TEC): a framework for perception and action planning. *Behavioral and Brain Sciences, 24*, 849-878; discussion 878-937.

Ichikawa, M., & Masakura, Y. (2006). Auditory stimulation affects apparent motion. *Japanese Psychological Research, 48*, 91-101.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: open source scientific tools for python. Available at: http://www.scipy.org

Kawabata, N., Yamagami, K., & Noaki, M. (1978). Visual fixation points and depth perception. *Vision Research, 18*, 853-854.

Laubrock, J., Engbert, R., & Kliegl, R. (2008). Fixational eye movements predict the perceived direction of ambiguous apparent motion. *Journal of Vision, 8*, 13.1-13.17

Leopold, D. A., & Logothetis, N. K. (1999). Multistable phenomena: changing views in perception. *Trends in Cognitive Sciences, 3*, 254-264.

Logothetis, N. K. (1998). Single units and conscious vision. *Philosophical Transactions of the Royal Society B, 353*, 1801-1818.

Logothetis, N.K., & Schall, J.D. (1990). Binocular motion rivalry in macaque monkeys: eye dominance and tracking eye movements. *Vision Research, 30,* 1409-1419.

Maruya, K., Yang, E., & Blake, R. (2007). Voluntary action influences visual competition. *Psychological Science, 18*, 1090-1098.

McCullagh, P., Weiss, M. R., & Ross, D. (1989). Modeling considerations in motor skill acquisition and performance: an integrated approach. *Exercise and Sport Sciences Reviews, 17*, 475-513.

Müsseler, J. (1999). How independent from action control is perception? An event-coding account for more equally-ranked crosstalks. In G. Ascherleben, T.Bachman & J. Müsseler (Eds.), *Cognitive contributions to the* *perception of spatial and temporal events* (pp. 121-147). Amsterdam: Elsevier.

Nawrot, M., & Blake, R. (1991). The interplay between stereopsis and structure from motion. *Perception and Psychophysics, 49*, 230-244.

Necker, L. (1832). Observations on some remarkable optical phenomena seen in Switzerland, and on an optical phenomenon which occurs on viewing a figure of a crystal or geometrical solid. *London Edinburgh Philosophical Magazine and Journal of Science, 1*, 329-337.

Oldfield, R. C. (1971). The assessment & analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9*, 97-113.

Oliphant, T. E. (2007). Python for Scientific Computing. *Computing in Science and Engineering, 9*, 10-20.

Paffen, C. L., Alais, D., & Verstraten, F. A. (2006). Attention speeds binocular rivalry. *Psychological Science, 17*, 752-756.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*, 3-25.

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: an information-processing account of its origins and significance. *Psychological Review, 83*, 157-171.

Prinz W. (1997). Perception & action planning. *European journal of cognitive psychology, 9*, 129-154.

R Development Core Team (2009). R: A language and environment for statistical computing. Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Reithler, J., van Mier, H. I., Peters, J. C., & Goebel, R. (2007). Nonvisual motor learning influences abstract action observation. *Current Biology, 17*, 1201-1207.

Repp, B. H., & Knoblich, G. (2007). Action can affect auditory perception. *Psychological Science, 18*, 6-7.

Rubin, E. (1915). *Visuell wahrgenommene figuren,* Copenhagen: Glyndendalske.

Schütz-Bosbach, S., & Prinz, W. (2007). Perceptual resonance: action-induced modulation of perception. *Trends in Cognitive Sciences, 11*, 349-355.

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385*, 308.

Shimojo, S., & Shams, L. (2001). Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinions in Neurobiology, 11*, 505-509.

Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: the Colavita effect revisited. *Perception & Psychophysics, 69*, 673-686.

Souman, J.L., Hooge, I.T.C., & Wertheim, A.H. (2006). Frame of reference transformations in motion perception during smooth pursuit eye movements. *Journal of Computational Neuroscience, 20*, 61-76.

Sun, F., Tong, J., Yang, Q., Tian, J., & Hung, G.K. (2002). Multi-directional shifts of optokinetic responses to binocular-rivalrous motion stimuli. *Brain Research, 944,* 56-64.

Toppino, T. C. (2003). Reversible-figure perception: mechanisms of intentional control. *Perception and Psychophysics, 65*, 1285-1295.

van Dam, L. C., & van Ee, R. (2005). The role of (micro)saccades and blinks in perceptual bi-stability from slant rivalry. *Vision Research, 45*, 2417-2435.

van Ee, R. (2009). Stochastic variations in sensory awareness are driven by noisy neuronal adaptation: evidence from serial correlations in perceptual bistability. *Journal of the Optical Society of America A, 26*, 2612-2622.

van Ee, R., van Boxtel, J. J., Parker, A. L., & Alais, D. (2009). Multisensory congruency as a mechanism for attentional control over perceptual selection. *Journal of Neuroscience, 29*, 11641-11649.

van Noorden, L. (1975). *Temporal coherence in the perception of tone sequences*. Eindhoven, The Netherlands: Unpublished doctoral dissertation, Eindhoven University of Technology.

Wheatstone, C. (1838). On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions Royal Society London, 128,* 371-394.

Witten, I. B., & Knudsen, E. I. (2005). Why seeing is believing: merging auditory and visual worlds. *Neuron, 48*, 489-496.

Wohlschläger, A. (2000). Visual motion priming by invisible actions. *Vision Research, 40*, 925-930.

Wolfe, J. M., & Bennett, S. C. (1997). Preattentive object files: shapeless bundles of basic features. *Vision Research, 37*, 25-43.

Zhou, W., & Chen, D. (2009). Binaral rivalry between the nostrils and in the cortex. *Current Biology, 19*, 1561-1565.

Zimmer, A. (1913). Die Ursachen der Inversionen mehrdeutiger stereomatrischer Konturenzeichnungen. *Zeitschrift für Sinnespysiologie*, 47, 106–158.

## Zusammenfassung

Menschen richten üblicherweise ihren Blick und ihre Aufmerksamkeit auf Stellen, die wichtig sind für die Tätigkeit, mit der sie sich momentan befassen. Mittels Blickrichtungsmessungen kann man die relative Wichtigkeit jeder Stelle abschätzen. Dies lässt Rückschlüsse auf die grundlegenden kognitiven Prozesse zu, denen die Auswahl der Blickrichtung unterliegt. Seit Jahrzehnten wird dies unter Laborbedingungen gemacht, mit dem großen Vorteil, gut kontrollierbar zu sein. In dieser Arbeit wird visuelle Aufmerksamkeit in natürlicheren Umgebungen untersucht, damit sowohl Laborergebnisse auf Realitätsnähe getestet , als auch Experimente unter realeren Bedingungen durchgeführt werden können, die im Labor nur schwierig nachzuahmen sind. Alle vier Studien in dieser Arbeit tragen zum Verständnis von visueller Aufmerksamkeit und Wahrnehmung unter komplizierteren Umständen, als man sie in herkömmlichen Laborexperimenten auffindet, bei.

Bottom-up-Modelle für Aufmerksamkeit verwenden lediglich die optischen Reize zur Vorhersage von Aufmerksamkeit oder sogar der Blickrichtung. Solche Modelle verarbeiten ein Bild zuerst getrennt nach den unterschiedlichen Merkmalen. Das klassische Saliency Map Model benützt die Merkmale Farbkontrast, Luminanzkontrast und Orientierungskontrast. Pro Merkmal wird die "Interessantheit" aller Stellen im Bild in einer sogenannten 'conspicuity map' ('Auffälligkeitskarte') wiedergegeben. Diese Karten werden linear addiert zu einer Salienzkarte und diese Additivität wurde in letzter Zeit in Frage gestellt. Eine Alternative wäre, jeweils für jede Stelle das Maximum aller Karten zu verwenden. In der ersten Studie wurden die Merkmale Farbkontrast und Luminanzkontrast an Bildern von natürlichen Szenen bearbeitet, um zu testen, welcher der beiden Mechanismen das menschliche Verhalten besser vorhersagt. Es konnte

gezeigt werden, dass lineare Additivität, so wie im ursprünglichen Modell, am besten mit dem menschlichen Verhalten übereinstimmt. Weil alle Annahmen vom Modell der Salienzkarte bis zur Addition der Karten auf Ergebnissen physiologischer Experimente beruhen, ist dieser Befund eine Einschränkung für zukünftige Modelle.

Wenn Modelle für visuelle Aufmerksamkeit realitätsnah sein sollten, ist ein Vergleich zwischen natürlichen Bedingungen und Laborbedingungen erforderlich, und dies wurde in der zweiten Studie gemacht. In der ersten Bedingung wurden kopfzentrierte Filme aus der Eigenperspektive aufgenommen und simultan die Augenbewegungen gemessen, während die Teilnehmer 15 natürliche Umgebungen erkundeten ("free exploration"). Abschnitte aus diesen Filmen wurden Teilnehmern in zwei Laborversuchen gezeigt. Im ersten wurden die Abschnitte, so wie sie aufgenommen wurden, ("video replay") und im zweiten wurden daraus ausgewählte Einzelbilder jeweils für eine Sekunde in willkürlicher Reihenfolge ("1s frame replay") gezeigt. Dabei gemessene Augenspuren weisen vor, dass im Vergleich zur 1s-frame-replay-Bedingung die Blickwinkelverteilung der video-replay-Bedingung qualitativ ähnlicher zur free-exploration-Bedingung ist und dass die Modellsalienz die Blickwinkel während der free-exploration-Bedingung quantitativ am besten vorhersagt. Ausserdem ruft das Zeigen eines neuen Einzelbilds bei der 1s-frame-replay-Bedingung eine Neuorientierung der Blickwinkel zur Mitte hervor. Das heisst, die Darstellung von einem Reiz unter Laborbedingungen beeinflusst Aufmerksamkeit auf eine Weise, die im echten Leben nur sehr unwahrscheinlich vorkommen wird. Die video-replay-Bedingung modelliert schlußfolglich natürliche visuelle Reize am besten.

Die Hypothese, ob Laufen auf unregelmäßigem Terrain verlangt, dass die Aufmerksamkeit mehr auf den Weg gerichtet wird, wurde auf einer örtlichen Straße in Marburg ("Hirschberg") in der dritten Studie geprüft. Die Teilnehmer haben dabei Strecken auf beiden Seiten dieser geneigten Straße zurückgelegt; die gepflasterte Straße und der anliegende, unregelmäßig gestufte Fußweg. Die Umgebung und die Anweisungen an die Teilnehmer wurden gleich gehalten. Der Blick wurde häufiger auf den Weg gerichtet, wenn die Teilnehmer auf dem Fußweg gelaufen sind, als auf der Straße. Dabei waren sowohl der Kopf als auch die Augen auf dem gestuften Fußweg mehr nach unten orientiert als auf der Straße, während die Orientierung von Auge im Kopf auf dem gestuften Fußweg vertikal weiter verteilt war, was auf häufigere oder größere Augenbewegungen deutet. Diese Ergebnisse untermauern frühere Befunde, dass Auge und Kopf bei der Blickausrichtung in der realen Welt unterschiedliche Rollen spielen. Darüber hinaus zeigen sie, dass eine implizite Aufgabe (nämlich nicht zu stürzen, in diesem Fall,) visuelle Aufmerksamkeit ebenso bestimmt, wie eine explizite Aufgabe.

In der letzten Studie wurde die Frage untersucht, ob Wahrnehmung durch Handlung beeinflusst wird. Dazu wurde ein zweideutiger Reiz benutzt, der entweder als im, oder als gegen den Uhrzeigersinn drehend (das ‘Perzept’) wahrgenommen wird. In Bedingungen, wo die Teilnehmer fortlaufend ein Manipulandum in eine vorgegebene Richtung drehen mussten – entweder im oder gegen den Uhrzeigersinn – und gleichzeitig das Perzept über eine Tastatur angegeben haben, wurden die Perzepte nicht von den Handlungen beeinflusst. Wenn die Teilnehmer das Manipulandum benutzt haben, um das Perzept anzugeben – entweder durch Rotieren in die gleiche oder in die entgegengesetzte Richtung, als das Perzept – wurde Wahrnehmung von der Handlung beeinflusst. Das Ergebnis zeigt, dass die Auflösung von Ambiguität in visuellen Reizen auf Motorsignalen beruht, aber nur wenn diese relevant für die momentane Aufgabe sind.

Sowohl durch die Verwendung von natürlichen Stimuli, durch den Vergleich von Verhalten im Labor mit dem Verhalten in der realen Welt, durch die Durchführung von einem Experiment auf der Straße als auch durch das Studieren der Integration zweier verschiedenartiger aber

alltäglicher Informationsquellen wurde das Sehfähigkeit in realitätsnahen Umständen untersucht. Die Stichhaltigkeit einiger Laborergebnisse wurde überprüft und bestätigt und einige erste Schritte zur Durchführung von Experimenten unter realitätsnahen Umständen wurden getan. Beide Ansätze scheinen vielversprechend zu sein für zukünftige Forschungen.

## Summary

Humans typically direct their gaze and attention at locations important for the tasks they are engaged in. By measuring the direction of gaze, the relative importance of each location can be estimated which can reveal how cognitive processes choose where gaze is to be directed. For decades, this has been done in laboratory setups, which have the advantage of being well-controlled. Here, visual attention is studied in more life-like situations, which allows testing ecological validity of laboratory results and allows the use of real-life setups that are hard to mimic in a laboratory. All four studies in this thesis contribute to our understanding of visual attention and perception in more complex situations than are found in the traditional laboratory experiments.

Bottom-up models of attention use the visual input to predict attention or even the direction of gaze. In such models the input image is analyzed for each of several features first. In the classic Saliency Map model, these features are color contrast, luminance contrast and orientation contrast. The “interestingness” of each location in the image is represented in a ‘conspicuity maps’, one for each feature. The Saliency Map model then combines these conspicuity maps by linear addition, and this additivity has recently been challenged. The alternative is to use the maxima across all conspicuity maps. In the first study, the features color contrast and luminance contrast were manipulated in photographs of natural scenes to test which of these mechanisms is the best predictor of human behavior. It was shown that a linear addition, as in the original

model, matches human behavior best. As all the assumptions of the Saliency Map model on the processes preceding the linear addition of the conspicuity maps are based on physiological research, this result constrains future models in their mechanistic assumption.

If models of visual attention are to have ecological validity, comparing visual attention in laboratory and real-world conditions is necessary, and this is done in the second study. In the first condition, eye movements and head-centered, first-person perspective movies were recorded while participants explored 15 real-world environments ("free exploration"). Clips from these movies were shown to participants in two laboratory tasks. First, the movies were replayed as they were recorded ("video replay"), and second, a shuffled selection of frames was shown for 1 second each ("1s frame replay"). Eye-movement recordings from all three conditions revealed that in comparison to 1s frame replay, the video replay condition was qualitatively more alike to the free exploration condition with respect to the distribution of gaze and the relationship between gaze and model saliency and was quantitatively better able to predict free exploration gaze. Furthermore, the onset of a new frame in 1s frame replay evoked a reorientation of gaze towards the center. That is, the event of presenting a stimulus in a laboratory setup affects attention in a way unlikely to occur in real life. In conclusion, video replay is a better model for real-world visual input.

The hypothesis that walking on more irregular terrain requires visual attention to be directed at the path more was tested on a local street ("Hirschberg") in the third study. Participants walked on both sides of this inclined street; a cobbled road and the immediately adjacent, irregular steps. The environment and instructions were kept constant. Gaze was directed at the path more when participants walked on the steps as compared to the road. This was accomplished by pointing both the head and the eyes lower on the steps than on the road, while only eye-in-head orientation was spread out along the vertical more on the steps, indicating more or large eye movements on the more irregular steps. These results confirm earlier findings that eye and head movements play distinct roles in directing gaze in real-world situations. Furthermore, they show that implicit tasks (not falling, in this case) affect visual attention as much as explicit tasks do.

In the last study it is asked if actions affect perception. An ambiguous stimulus that is alternatively perceived as rotating clockwise or counterclockwise (the 'percept') was used. When participants had to rotate a manipulandum continuously in a pre-defined direction – either clockwise or counterclockwise – and reported their concurrent percept with a keyboard, percepts weren't affected by movements. If participants had to use the manipulandum to indicate their percept – by rotating either congruently or incongruently with the percept – the movements did affect perception. This shows that ambiguity in visual input is resolved by relying on motor signals, but only when they are relevant for the task at hand.

Either by using natural stimuli, by comparing behavior in the laboratory with behavior in the real world, by performing an experiment on the street, or by testing how two diverse but everyday sources of information are integrated, the faculty of vision was studied in more life like situations. The validity of some laboratory work has been examined and confirmed and some first steps in doing experiments in real-world situations have been made. Both seem to be promising approaches for future research.

## Samenvatting

Mensen richting hun blik en aandacht gewoonlijk op locaties die belangrijk zijn voor hetgeen waar ze zich mee bezig houden. Door blikrichting te meten kan het relatieve belang van elke locatie worden bepaald, waardoor bloot gelegd kan worden hoe cognitieve processen bepalen waar de blik op gericht wordt. Tientallen jaren lang is dit gedaan in laboratoria, met goede beheersbaarheid als voordeel. Hier wordt visuele aandacht bestudeerd in meer levensechte situaties, wat het testen van ecologische validiteit van laboratorium resultaten toestaat en het gebruik van echte omgevingen die moeilijk na te bootsen zijn in het laboratorium. Alle vier de studies in deze dissertatie dragen bij aan ons begrip van visuele aandacht en perceptie in complexere situaties dan wat men in een traditioneel laboratorium experiment tegenkomt.

Bottom-up modellen van aandacht gebruiken het visuele signaal om aandacht of zelfs blikrichting te voorspellen. Dergelijke modellen analyseren een beeld eerst op verschillende eigenschappen. In het klassieke Saliency Map model zijn deze eigenschappen color contrast, luminance contrast en orientation contrast. De "bemerkenswaardigheid" van elke locatie in het beeld wordt weergegeven in een zgn. 'conspicuity map', één voor elke eigenschap. Het Saliency Map model combineert deze conspicuity maps additief door een lineaire sommatie, en deze additiviteit is recent ter discussie gesteld. Het alternatief is om maxima over alle conspicuity maps te gebruiken. In de eerste studie, werden de eigenschappen color contrast en luminance contrast bewerkt in fotos van natuurlijke scenes om te testen welke van de mechanismen de beste voorspelling van menselijk gedrag levert. Het werd aangetoond dat additiviteit, zoals in het oorspronkelijke model de beste voorspelling voor menselijk gedrag oplevert. Aangezien alle aannames van het Saliency Map model die voorafgaan aan de sommatie van de conspicuity maps zijn gebaseerd op fysiologisch onderzoek, is dit resultaat een restrictie voor de mechanistische aannames van toekomstige modellen.

Wanneer modellen voor visuele aandacht ecologisch valide moeten zijn, dan is een vergelijking tussen visuele aandacht in een laboratorium omgeving en echte omgeving noodzakelijk, en dit werd in de tweede studie gedaan. In de eerste conditie werden oogbewegingen en een film vanuit het eerste-persoons perspectief opgenomen terwijl de deelnemers 15 natuurlijke omgevingen verkenden (“free exploration”). Fragmenten uit deze films werden vertoond aan proefpersonen in twee laboratorium taken. Ten eerste werden de fragmenten zo vertoond als ze opgenomen waren (“video replay”), en ten tweede werd een selectie van frames in willekeurige volgorde elk 1 seconde lang getoond (“1s frame replay”). Opnames van oogbewegingen uit alle drie de condities toonden dat in vergelijking met 1s frame replay, de video replay conditie kwalitatief meer lijkt op free exploration wanneer gekeken wordt naar de verdeling van blikrichtingen en de samenhang tussen blikrichting en gemodelleerde saliency, en de blikrichting in free exploration kwalitatief beter voorspelt. Tevens gaf het tonen van het volgende frame in de 1s frame replay conditie aanleiding tot een heroriëntatie van blikrichting naar het midden. Oftewel, het feit dat er een stimulus getoond wordt in een laboratorium omgeving beïnvloedt aandacht op een manier die in het echte leven onwaarschijnlijk is. Kortom, video replay is een beter model voor natuurlijke visuele stimulatie.

De hypothese dat lopen op een meer onregelmatige ondergrond vereist dat visuele aandacht in toenemende mate op het traject gericht moet worden, werd op de proef gesteld op een plaatselijke weg (“Hirschberg”) in de derde studie. De deelnemers liepen aan beide zijden van deze oplopende weg; een klinkerstraat (“straat”) en de direct aangrenzende stoep met onregelmatig geplaatste treden (“treden”). De omgeving en de instructies werden gelijk gehouden. In vergelijking met straat richten de deelnemers hun blik meer op het traject wanneer ze op de treden liepen. Dit werd bereikt door zowel het hoofd als de ogen lager te richten op de treden dan op de straat, terwijl slechts de oriëntatie van het oog in het hoofd verticaal een grotere spreiding had op de treden, wat er op duidt dat mensen meer of grotere oogbewegingen maakten op de meer onregelmatige treden. Deze resultaten bevestigen eerdere bevindingen die uitwezen dat oog- en hoofdbewegingen verschillende rollen spelen in het bepalen van blikrichting in natuurlijke situaties. Tevens laten de resultaten zien dat impliciete taken (niet struikelen, in dit geval) visuele aandacht net zo goed beïnvloeden als expliciete taken.

In de laatste studie wordt getest of handelen de waarneming beïnvloedt. Een ambigue stimulus die afwisselend waargenomen wordt als met de klok mee of tegen de klok in roterend (het ‘percept’) werd hiervoor gebruikt. Wanneer de deelnemers een manipulandum continu in een voorgeschreven richting moesten roteren – ofwel met de klok mee ofwel tegen de klok in – en tegelijkertijd het percept rapporteerden via een toetsenbord, werden de percepten niet beïnvloed door handbewegingen. Wanneer de deelnemers de manipulandum moesten gebruiken om het percept te rapporteren – door met het percept mee of tegen het percept in te draaien – beïnvloeden de bewegingen de waarneming. Dit toont aan dat ambiguïteit in visuele stimuli opgelost wordt met behulp van motorische informatie, maar slechts dan wanneer die relevant zijn voor de taak die uitgevoerd wordt.

Door het gebruik van natuurlijke stimuli, het vergelijken van gedrag in het laboratorium met gedrag in de echte wereld, het uitvoeren van een experiment op straat, of door te testen hoe twee verschillende maar alledaagse bronnen van informatie worden geïntegreerd, werd het gezichtsvermogen onder meer natuurlijke omstandigheden bestudeerd. De validiteit van enkele laboratorium resultaten is onderzocht en bevestigd en enige eerste stappen in het doen van experimenten in een natuurlijke omgeving zijn gemaakt. Beide benaderingswijzen lijken veelbelovend voor toekomstig onderzoek.

# Acknowledgements

I would like to express my gratitude to everybody who supported me in writing this dissertation and during my work here in Marburg. Two people in particular made this thesis possible. I would like to thank Prof. Dr. Frank Rösler for accepting me as his PhD student. I would also like to thank Wolfgang Einhäuser-Treyer, who supervised me throughout the years. I have learned so much from you about the study of vision and you've shown me how to enjoy being a scientist. I appreciate your guidance and your no-nonsense approach to almost everything, and I hope that we can continue working together in the future.

The other members of the examination board, thank you for committing your time and expertise to examining my dissertation and defense.

I would like to thank the numerous participants in all studies for their patience and hard work.

My colleagues Josef Stoll and Marnix Naber, you have done much for me, but above all I would like to thank you for providing such a positive atmosphere in our team and shared office. Collaborating with you in the real world and in the laboratory has been a joy. Frank Bremmer, thank you for welcoming a dutch guy in your group. I look forward to working with you in the future. I would like to thank Sigrid Thomas and Alexander Platzner for their invaluable support in all projects. All the students I had the privilege to work with; Jan-Hendrik Alsmeier, Anton Riske, Adrien Pfeuffer, Tilman Abresch, Christine Roth, Ole Krüger, Hannah Schmidt, Ingo Klein-Harmeyer, Rabea Rueßwig and Svenja Marx, I would like to thank you all for your contributions, your refreshing views on many topics and for putting up with my continuous demands. The same applies to everybody in the AG Neurophysik, your varied specializations have made my stay in Physics very interesting and educational and your willingness to

participate in (or tolerate) some of my weirder experiments has been great. This lab has become a place to call home. I will stay here a while longer, and hope to keep in touch, wherever I may end up.

For providing me with an opportunity to look beyond the walls of our own laboratory and see how people in other groups and disciplines work and think, I would like to thank all the members of NeuroAct, especially the speakers Frank Bremmer and Karl Gegenfurtner. This has been a very enriching experience, for which I am deeply grateful.

My gratitude also extends to those I collaborated with on a number of manuscripts. Among others, these are Peter König and Sonja Engmann from Osnabrück and Erich Schneider, Günter Kugler, Stefan Kohlbecher and Johannes Vockeroth from Munich. Your feedback and exemplary work has guided me in doing real-world oriented research.

For an excellent collaboration, and for their friendship, I would particularly like to thank Iseult Beets and Katja Fiehler. Your lab at Psychology has become a second home for me at the university, not in the least because of the company of Johanna Mühe, Jasmin Kızılırmak and Anna Seemüller and all the other people there. A special ‘thank you’ goes to another of my collaborators, Denise Henriques, for pushing me to finish this thesis. Your energy and thoughts continue to be inspirational.

Anke, Arnold, Hubert, Elida, Marleen, Marcel en sinds kort Tygo en alle andere familie en vrienden, jullie wil ik bedanken voor de onvoorwaardelijke steun tijdens dit avontuur in het buitenland, voor de logeeradressen en voor de regelmatige bezoekjes aan Marburg.

## Curriculum Vitae

| | |
|---|---|
| **Name:** | Bernard Marius ’t Hart |
| **Born:** | April 28,1977, Stadskanaal |
| **Nationality:** | Dutch |
| **Work address:** | Philipps-University Marburg<br>Faculty of Physics / Dept. of Neurophysics<br>Karl-von-Frisch-Straße 8a<br>35032 Marburg |
| **Phone:** | +49 6421 28 24176 |
| **E-mail:** | thart@staff.uni-marburg.de |

**Education and career:**

1996 – 2004, Psychology at RuG university in Groningen, specialisation: Cognitive Psychology, minor in Technical Cognition Sciences.

1999 – 2000, designed computer practicum part for a course in observing behavior, and taught the whole course to one class of students.

2008 – 2010, associate member of the DFG Research Training Program “Neural representation and Action Control –NeuroAct”.

2008 – 2009 (Winter Semester), Designed and taught: F1 Praktikum Psychophysics.

2009, “Computational Vision”, Training Workshop organized by Prof. Dr. G. Deco and Prof. Dr. K. Gegenfurtner at Schloss Rauischholzhausen.

2008 – 2011, supervised 3 Bachelor projects, 1 Diplom project and several smaller student projects.

2008 – 2011, PhD student at Neurophysics, Philipps University Marburg, Germany.

**Publications:**

’t Hart BM, Einhäuser W (under review). Mind the step: complementary roles for eye-in-head and head-in-world orientation when negotiating a real-life path.

’t Hart BM, Abresch T, Einhäuser W (under review). Faces in places: Humans and machines make similar face-detection errors.

Preuschoff K, ’t Hart BM, Einhäuser W (under review). Pupil dilation signals surprise: evidence for noradrenaline’s role in decision making.

SareyKhanie M, ’t Hart BM, Stoll, J, Andersen M, Einhäuser W (2011). Integration of eye-tracking methods in visual comfort assessments. CISBAT conference 2011.

Beets IAM*, ’t Hart BM*, Rösler F, Henriques DYP, Einhäuser W, Fiehler K (2010). Online action-to-perception transfer: only percept-dependent action affects perception. Vision Research. (*: IAMB & BMtH contributed equally)

’t Hart BM, Vockeroth J, Schumann F, Bartl K, Schneider E, König P, Einhäuser W (2009). Gaze allocation in natural stimuli: comparing free exploration to head-fixed viewing conditions. Visual Cognition.

Engmann S*, ’t Hart BM*, Sieren T, Onat S, König P, Einhäuser W (2009). Saliency on a natural-scene background: Effects of color- and luminance-contrast add linearly. Attention, Perception & Psychophysics. (*: ES & BMtH contributed equally)

Gladwin TE, ’t Hart BM, De Jong R (2008). Dissociations between motor-related EEG measures in a cued movement sequence task. Cortex.

De Jong R, Gladwin TE, ’t Hart BM (2006). Movement-related EEG indices of preparation in task switching and motor control. Journal of Brain Research.

**Invited talks:**

2011, June 30: Prof. Dr. M. Niemeier, University of Toronto at Scarborough, Department of Psychology. “Features, objects, task: what drives real-world visual attention?”

2011, June 28: Prof. Dr. D.Y.P. Henriques, York University, Centre for Vision Research. “Features, objects, task: what drives real-world visual attention?”

2011, May 13: Prof. Dr. S. Swinnen, University Leuven, Research Centre for Movement Control and Neuroplasticity. “Visual Attention: Gazing at the screen vs. walking on the street.”

**Posters:**

’t Hart BM, Kugler GA, Bartl K, Kohlbecher S, Schumann F, König P, Einhäuser W, Brandt T, Schneider E (2011). Real-world search strategies with normal and deficient color-vision. ECVP abstract.

’t Hart BM, Schmidt H, Klein-Harmeyer I, Einhäuser W (2011). Objects in natural scenes: Do rapid detection and gaze-control utilize the same features? ECEM abstract.

’t Hart BM, Einhäuser W (2011). Rapid adaptation of gaze to meet demands in negotiating terrain. CVR abstract.

Abresch TGJ, ’t Hart BM, Einhäuser W (2011). Similar errors in human and computational face-detection. NWG abstract.

Einhäuser W, ’t Hart BM, Preuschoff K (2011). Pupil dilation reflects unexpected uncertainty: a role for noradrenalin in decision-making. NWG abstract.

Kugler G, ’t Hart BM, Bartl K, Kohlbecher S, Schumann F, Einhäuser W, Brandt T, Schneider E (2011). Looking For Candy: Real-world, feature based search. NWG abstract.

Fiehler K, ’t Hart BM, Beets IAM, Henriques DYP, Einhäuser W (2010) The impact of action on perception: Evidence from ambiguous visual stimuli. ECVP abstract.

Preuschoff K, ’t Hart BM, Einhäuser W (2010). Pupil dilation reflects judgement of uncertainty. ECVP abstract.

’t Hart BM, Abresch T, Einhäuser W (2010). Humans make similar errors as computational algorithms when detecting faces. ECVP abstract.

’t Hart BM, Einhäuser W (2009). The effect of terrain on eye movements while walking in the real world. ECVP abstract.

’t Hart BM, Vockeroth J, Schumann F, Bartl K, Schneider E, König P, Einhäuser W (2009). Gaze allocation during natural behavior in the real world. NWG.

’t Hart BM, Gladwin TE, De Jong R (2007). Measuring Partial Motor Programs. EWOMS abstract.

# Erklärung

Ich versichere, dass ich meine Dissertation

*Visual attention in the real world*

selbständig, ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen und Hilfen bedient habe.

Die Dissertation wurde in der jetzigen oder einer ähnlichen Form noch bei keiner anderen Hochschule eingereicht und hat noch keinen sonstigen Prüfungszwecken gedient.

Marburg

(Ort / Datum)

(Unterschrift mit Vor- und Zuname)