“Factors regulating the bottom-up guidance of overt visual attention under natural conditions”

Dissertation

zur Erlangung des Grades

"Dr. rer. nat."

im Fachbereich Humanwissenschaften der

Universität Osnabrück

vorgelegt von

Alper Açık

Osnabrück, September 2014

Thesis Summary

1. Introduction

The main goal of the present thesis is to contribute to the study of overt visual attention under natural conditions. All publications included in the thesis employ eye-tracking methodology and local image feature analysis. The particular scientific question posed, if all parts of the thesis are considered together, can be formulated as: What experimental (i.e. experimenter controlled) and pseudo-experimental (e.g. demographic characteristics such as age) variables define, quantify and set limits for bottom-up determinants of fixation selection? In this summary, I will bring together the answers to this question obtained during the course of the thesis and try to elucidate their significance in relation to the puzzles and riddles of eye movement research. After presenting an overview of the field, I will show how and to what extent fixation selection is altered as a function of age and as certain stimulus components are modified. Finally, I will argue that visual attention involves varying degrees of reliance on bottom-up cues according to the attending agent's knowledge that is relevant in a given context.

2. The role of local image features in the guidance of attention

Eye movements are an excellent example of how agents interact with their environment establishing action-perception loops. With each fixation we perceive a local portion of the environment, and while the eyes are at this location, peripheral input is used to select the end-point of the next eye movement, i.e. saccade. Accordingly, the properties characterizing fixated locations are important both in terms of visual and motor processing.

So what determines where we look? At a coarse level of interpretation, the research conducted in the last two decades makes clear that top-down and bottom-up processes both play important roles (Corbetta & Shulman, 2002; Henderson & Hollingworth, 1999; Schütz et al., 2011). In the psychology literature, top-down perception refers to the role of prior knowledge in a given perceptual experience (Goldstein, 2011). This knowledge encompasses previous perceptual experiences, i.e. memory, the goals of the agent, together with the motivational and emotional states (Kaspar & König, 2012). For instance, since the early work of Yarbus (1967), it is well known that people inspect different regions of the same scene, if they are given different tasks before stimulus presentation. Moreover, global statistical properties of different types of scenes such as urban environments or landscapes that are learned during ontogeny provide another interesting source of top-down knowledge. The stimulus-driven aspects of visual selection, on the other hand, correspond to the bottom-up part of the story (Itti & Koch, 2001). Bottom-up processing addresses which statistical properties of local image regions are registered early in visual processing hierarchy, and how these properties –labeled features– get involved in the computation of eye movement targets. Deviation of luminance, features derived from color, local edge and corner content have been shown to discriminate between fixated and not fixated regions, even though the exact level of discrimination success is still debated (Einhäuser et al., 2008b; Koehler et al., 2014; Rothkopf et al., 2007). Thus, both bottom-up and top-down factors are known to guide fixations during inspection of the environment.

The saliency map model that was originally proposed in order to explain behavioral differences between the two modes of visual search –pop-out and conjunction search– has proved itself an immensely popular conception among the research tradition that emphasizes the bottom-up aspects of fixation selection (Itti & Koch, 2001; Parkhurst et al., 2002). The idea at the core of the application of saliency maps posits that fixation probability is a monotonically increasing function of several local image features, which are derived from local deviations such as luminance contrast (LC). That is, if the luminance in a region displays higher variance than another region, with all other feature values being equal, the former will be a more probable target of fixations. The saliency map model first splits the scene into its feature channels independently in order to create several feature maps. Later, these maps are combined point-wise according to a rule that determines the relative weights of each feature channel. A saliency map is the output of this procedure and reveals the most visually attracting regions in the input scene. Both studies measuring relatively high feature values at fixations (Reinagel & Zador, 1999), and better-than-chance matches between saliency maps and fixation probability densities obtained from human eye movement data seem to support the basic assumptions of these models (Wilming et al., 2011). Accordingly, such saliency map models that compute fixation probability from separate feature channels have become an important analysis tool for the study of fixation selection during inspection of natural scenes.

The fact that relatively high feature values characterize fixated regions does not prove that we look at those regions because of their high feature values. Any elementary statistics textbook underlines that a correlation between the variables *A* and *B* could mean *A* causing *B*, *B* causing *A*, a bidirectional causation, a third variable causing both *A* and *B*, or pure coincidence (for a historical overview see Aldrich, 1995). Similarly, demonstrating a positive correlation between fixation probability and local feature values is not enough to conclude that local features guide eye movements. In order to address this issue one has to carefully modify local LC, and measure the influence of that modification on fixation probability. Einhäuser and König (2003) introduced a novel method with which they were able to decrease or increase local LC with negligible alterations in other features. They have demonstrated that both such decreases and increases have increased the fixation probability at modified regions. This finding is at odds with the monotonic increase assumption, since otherwise the observers should look at LC-decreased regions less often than without modification. The authors suggested that, in natural scenes, LC might happen to be correlated with some other attention attracting higher-order property of such scenes. LC modifications extinguish or change the nature of the correlation between the higher-order properties and LC. Consequently, after the modifications, the relationship between LC and fixation probability does not correspond to a positive correlation. In summary, methods that modify local LC suggest that the relationship between features and fixation probability is not always causal.

3. Present contributions to the causation-correlation debate

Four of the works that are part of this thesis investigate whether (and if, when) the relation between local features and fixation probability is causal or correlational. The first study (Açık et al., 2009), a re-analysis of the data gathered for the purposes of my Master's thesis (Açık, 2006/2010), addresses how local LC modifications alter fixation behavior during inspections of scenes that belong to different image categories. I show that the effects of local LC modifications are qualitatively distinct for different image categories and that effects of modifications can be partially predicted looking at the higher-level properties of the scenes. For landscapes I replicate the findings of Einhäuser and König (2003), but in the case of fractal images there is a complete absence of LC modification effect. The second work (Açık et al., 2008), presented in poster format, employs the LC modification method on fractal images that are subject to the addition of different levels of phase noise in the spatial frequency domain. Here I show that the absence of effect in the case of fractals can be explained with the presence of higher-order correlations in this image category. If these correlations are removed with the addition of phase noise, both LC increases and decreases start to attract attention, which was observed in the previous study only for landscapes. In the third study (Açık et al., 2014), we present natural movies and frames taken from those movies in order to assess the role of movement features in fixation selection. Since the static stimuli are taken from movies, we have access to their spatial movement feature distributions. We demonstrate that local movement features correlate with fixation probability even if there is no motion in the stimulus. Rather than arguing that movement features are not at all causally related to eye movements, we interpret our results with reference to implied motion and motion onset salience literature (Freyd, 1987). In the fourth study, we directly quantify the contributions of stimulus statistics and behavioral relevancy separately. For that purpose, we allow a group of participants viewing a large set of images denote those image regions that they find most interesting with mouse clicks. Another group of subjects view the same images in the absence of a task while their eye movements were recorded. The design enables us to correlate fixation probability both with local stimulus statistics and the interestingness rating, the latter serving as a measure of behavioral relevancy. Moreover, we employ a novel Bayesian framework and measured the saliency of both individual features and their pairwise combinations. We reveal that interestingness ratings correlate with fixation probability much more than individual and combined feature saliencies. Below we revisit these four studies and interpret their findings in relation to each other in order to present an overview of how they contribute to causation/correlation debate.

3.1 Effects of local contrast modifications on fixation behavior

In the first study (Açık et al., 2009) I have re-analyzed the data collected for my thesis in order to address whether local LC modification effects on fixation behavior are image-category specific. We have chosen four image categories: Landscapes that were also used by Einhäuser and König (2003), urban scenes, faces and fractal images. In the scene-viewing literature, urban scenes and landscapes are commonly used since they depict two types of natural environment inhabited by humans, yet differ from each other in some important statistical aspects. This is particularly due to the abundance of vertical and horizontal lines encountered in city environments. Faces were selected since they might provide a test bed for the causation/correlation debate: Do we fixate on the eyes and mouth because of their contrast, or do we search for communicative parts of a face, which happen to contain high contrast? Finally, fractals have similar second-order statistical properties as natural environments due to their spatially self-repeating properties; nevertheless they look unfamiliar and unnatural. Each image was shown under 9 LC modification conditions: No contrast modification, 3 levels of contrast decrease, and 5 levels of contrast increase. For each modified stimulus, the regions that were modified were unique. The comparison of modification distributions at fixated and not fixated points revealed image category differences. In landscapes, both LC increases and decreases are salient, replicating Einhäuser and König's (2003) observations. In faces and urban scenes only LC increases are salient, and the effect size is not as high. Most interestingly, the saliency of LC modified regions in fractal images did not change at all. We have first shown that the changes in fixation behavior for the former categories cannot be explained by unintended texture increases as suggested by Parkhurst and Niebur (2004). Second, we have demonstrated that the absence of effect in the case of fractals is not due to the failure of the modification method in increasing the LC. Finally, we have tested the idea that the amount of "structure" in a given scene determines whether the modifications will alter local saliency. With structure we refer to the global properties of the image (Torralba et al., 2006), such as the presence of the horizon line spanning the whole image, recursive patterns, symmetries, and stereotypical object location configurations such as those contained in faces. In the present study, the mean phase congruency and phase symmetry of a given image were chosen as indirect estimates of structure amount. These contrast-invariant metrics reveal lines, corners and blob-like regions in a scene. We have revealed a statistically significant negative correlation between these metrics and the modification effect. Our landscape images featuring a lot of undergrowth and bushes lacked global image structure, and the LC modifications on these images changed the viewing behavior of the participants. Fractals, on the other hand, contain recursive patterns and large-scale symmetries, have high structure scores, and their local saliency is not influenced by the modifications. Thus, the LC modifications altered local saliency in all categories but fractals, and one reason for this difference is the high amount of global scene structure in that category.

How can we test the proposed relationship between scene structure and local feature modification effects directly? Our second study (Açık et al., 2008) combines local LC modifications with global phase manipulations in the spatial frequency domain. The addition of noise to its phase spectrum –i.e. phase scrambling– changes the image's higher-order statistics, which make up the global scene structure. At moderate levels of phase scrambling, the content of the image is still partially visible, nevertheless the image looks distorted or as though seen through a fog. If all the phase information in the image is randomized, the image appears as pink noise with the absence of any original image content. For this study we have used only fractal images, since, as argued above, they have high amounts of scene structure.

Our hypothesis states that the LC modifications will attract gaze only if phase scrambling removes enough of the structure in the image. After the introduction of global phase modulations, the relationship between LC and fixation probability has two facets. First, one can look at whether LC modified regions are fixated more frequently than before. Second, independent of modifications one can address the difference between the LC values encountered at fixated and control points. Thus, we can talk of modification saliency and the intrinsic LC saliency separately. Our results reveal a very interesting double dissociation for the two types of saliencies as a function of phase modulation. Whereas modification saliency increases with the addition of phase noise, the intrinsic saliency of LC, that is high in fractal images, drops to zero for pink noise images. This result is completely in line with Einhäuser and König's (2003) proposal that in natural scenes LC is correlated with salient higher-order properties. Due to this relation, during viewing of fractals rich in higher-order content, the intrinsic saliency of LC turns out to be high. Nevertheless, LC contrast modifications spatially independent of higher-order properties fail to influence viewing behavior. However, phase scrambling removes higher-order correlations and as a result, intrinsic saliency approaches zero. Importantly, the increase in the modification saliency after the addition of phase noise can be explained with strong modifications assuming a salient object-like quality due to the absence of image structure in pink noise images as previously proposed by Einhäuser and colleagues (2006). Moreover, the mean phase symmetry and congruency measures of phase-scrambled images were analyzed. The relationship between phase congruency and local modification effect was similar to that observed in the image category study (Açık et al., 2009). These results conclusively show that whether contrast modifications will attract attention depends strongly on the amount of scene structure, since such modifications are only salient if the image lacks such structure.

Both studies covered so far demonstrate that local luminance contrast does not always cause higher fixation probability. Even though contrast elevations are salient in landscape images, the same holds for contrast decreases. It is possible that contrast modifications in both directions have the potential of creating object-like regions, which are salient due to the object detection tendency of the visual system. This argument is strengthened by the observation that in urban scenes and faces contrast modifications are less salient. Unlike the spatially more homogeneous landscape scenes used in our study due to the abundance of regions containing bushes and undergrowth, urban scenes and faces depict many individual objects such as cars, streetlamps and face parts. The absence of effect in fractals, and the observed relationship between higher-order statistics and local modification saliency lead to the proposal that the presence of objects can be formalized if the amount of structure in the scene is quantified. Thus, rather than displaying an independent monotonic relation to fixation probability as assumed in purely bottom-up saliency map models, the saliency of local features depends strongly on global scene properties.

3.2 Is movement salient even in the absence of real motion?

Together with the local image feature modifications discussed above, the complete removal of a given feature channel is a second strategy employed in studies of causation and correlation. Frey and colleagues (Frey et al., 2011) presented their subjects rainforest images that were either intact in terms of their color or with the red-green or blue-yellow channel removed. For the fixations performed on color channel removed images, they could still analyze the correlations between fixation probability and color features, since these features could be computed from the base images. In the case of red-green-channel removal, the fixated locations were still characterized by higher feature values of red-green contrast. In a further experiment in the same study (Frey et al., 2011), red-green blind—deuteranope—subjects' data displayed red-green contrast saliencies comparable to healthy controls after the very first fixations. Thus, even though the red-green channel information was not present at the system level –due to color removal or deuteranopia– the subjects in those studies fixated regions which had relatively high color feature values in the underlying stimuli. This suggests that the saliency of color features is at least partly due to the correlations of these features with other attention attracting properties such as higher-order statistics or edge and corner content. That is, studies relying on the selective removal of a feature channel agree with local modification experiments, since both strategies reveal that the relationship between fixation probability and local features is not always causal.

In the third study presented here (Açık et al., 2014), we carry the feature removal idea to the domain of temporally varying natural stimuli, i.e. movies. Even though there is an increase in the number of eye movement studies employing movies rather than static pictures (Böhme et al., 2006; Carmi & Itti, 2006; Dorr et al., 2010; Le Meur et al., 2007; Machner et al., 2012; Marat et al., 2009; Mital et al., 2011; 't Hart et al., 2009), only few of those directly compare eye movements recorded during viewing of comparable dynamic and static stimuli (Dorr et al., 2010; Machner et al., 2012; 't Hart et al., 2009). Our stimuli consisted of 216 short movie segments, and single frames taken from each of them. The latter stimulus condition is crucial, since it allows us to analyze the fixation and movement feature relationship on statically viewed stimuli. Using the movies, we computed for each successive frame pair the spatial distribution of movement energy and movement contrast. Accordingly, we had access to local movement properties for the statically shown stimuli as well. This methodology is analogous to Frey et al.'s (2011); similar to their analysis on the role of color features during viewing of color-removed stimuli, we addressed whether movement features are related to fixation selection while inspecting static scenes.

We have uncovered that movement feature values at fixations were significantly higher than at not fixated locations during viewing of both static and dynamic stimuli (Açık, 2014). Nevertheless, the fixation discrimination ability of movement features was as good as the best static feature (intrinsic dimensionality analysis) only in the case of movie stimuli. At first glance, this suggests that the fixation discriminatory role of movement features observed during static stimulus viewing is due to the correlations between movement features and other attention attracting properties, and that the causal component is added on top of that if the stimuli contain real motion. Our additional analysis suggests otherwise. First, movement and static features display very low levels of statistical dependence among them. Second, movement features discriminate the first few fixations during viewing equally well for both types of stimuli. Moreover, the discrimination ability starts to decline for later fixations on both movies and frames, but less slowly for the former. Accordingly, if the argument regarding separate correlational and causal contributions is correct, during the early part of viewing, both contributions must be present for static stimuli. This can be explained with the notion of implied motion, i.e., motion deduced from static cues (Hubbard, 1995). When viewing photographs, usually it is not difficult to notice which regions would have been moving when the photograph was taken. As such, motion –real or implied– might be a strong cue for the guidance of attention. This argument successfully explains why feature saliencies are comparable for the first few fixations, but the divergence observed for the later phase of viewing remains to be explained. Studies employing artificial stimuli reveal that motion onset is more salient than motion *per se*. Accordingly, when our stimuli first appear, all implied and real motion sources are novel and attract attention. However, after the appearance of the stimulus, regions containing novel motion can enter the frame only if the stimuli are movies. To sum up, this study reveals that onset of motion, which can be real motion or deduced from static cues, is a predictor of attentional allocation during viewing of both movies and static pictures. Whether this relationship between movement cues and fixated locations requires the involvement of a bottom-up mechanism will be discussed below.

3.3. Understanding fixated regions: High feature saliency versus interestingness

If local stimulus statistics, and thus bottom-up information, don't have an exclusively causal role during eye guidance, what are the top-down aspects of scene-viewing that characterize the distribution of fixation locations better than those statistics? In the final study presented in this section (Onat et al., 2014) we have presented one group of participants colored images belonging to three different categories: Natural scenes, urban environments, and fractals. There was no eye-tracking and during image presentation the task of the participant was to click on five points on the image that they found subjectively most interesting. These mouse clicks were employed in order to create topographical interestingness maps akin to local feature maps. These maps generated with the conscious decisions of the subjects reflect the behaviorally relevant portions of the scene in the context of the current task. A different group of eye-tracked subjects viewed these images together with colored phase-scrambled images in the absence of any task. Employing a novel Bayesian framework, we have compared the fixation discrimination performance of three classes of properties: Single image saliencies, combined saliencies of feature pairs and interestingness. Thus, what we do is compare feature saliencies with the behaviorally relevant measure of image region interestingness determined by human observers.

Even though the magnitude of effects once again depended on image category, the qualitative pattern was straightforward: Single feature saliencies were low, paired feature saliencies were slightly higher, and neither of those was a match for interestingness. Moreover, for different image categories the features correlated with fixation probability were not always the same, underlining the importance of scene gist in determining the features of interest. For instance, whereas second-order statistics were useful in predicting fixation locations on urban scenes, natural scene fixations were characterized by high first-order statistics. Finally, we have addressed a typically neglected aspect of scene viewing, namely fixation duration distributions, in relation to saliency and interestingness. Linear analyses revealed that even though a feature saliency that appeared to be relevant for a given image category is able to explain some statistically significant variance in fixation durations, the impact of interestingness was about 3.5 times stronger than the effect of feature saliency. In summary, the results from this study demonstrate that in visual scene exploration both low- and high-level information sources have a say, nevertheless the latter is much more effective than stimulus features.

As mentioned before, all saliency map models require a) a decision on which features to consider and b) a mathematical routine to combine the variations in the selected feature channels so that a single-dimensional saliency value can be ascribed to each image region. As we discussed in the motion saliency paper (Açık et al., 2014), independent of the mechanical details of the selection and integration procedures, in order to maximize the success and efficiency of the model, there are two conditions to be met. First, individual feature saliencies must be high in the sense that high feature values are found at fixated locations. Second, the features to be included should have relative statistical independence among them in order to explain different aspects of the fixation distribution. In the motion study (Açık et al., 2014), both the static intrinsic dimensionality feature and the dynamic motion contrast had high saliencies, and compared to within static or dynamic feature dependencies, they were relatively independent from each other. In the present interestingness study (Onat et al., 2014) we have addressed combinations of feature values at fixated and not fixated locations in order to compute empirical, that is, observer data derived, saliency functions. We show that feature pair saliencies that are computed from fixation data are higher than individual feature saliencies. Moreover, especially for the high saliency second-order features, a linear combination of individual feature saliencies approximate observed joint feature saliency distributions reasonably well. These results legitimize the linear integration of features in the absence of complex interactions that is a common practice in saliency map modeling. In a similar vein, we have computed a joint saliency function for interestingness and the best stimulus feature saliency. In this case, the saliency was to a large degree modulated by interestingness with a modest contribution by the best stimulus feature, which rules out the proposition that low-level features alone can explain the interestingness and fixation probability relationship. Thus, this study confirmed that whereas a linear integration of stimulus features is appropriate for saliency modeling, the regions that are handpicked by observers as most interesting correspond to regions that are fixated most frequently with an inability of combined low-level features to predict which regions in the scene are rated as interesting by human observers.

Before we move on to a general discussion of the above summarized findings, we'd like to compare our interestingness study (Onat et al., 2014) to two recent studies employing similar methodologies (Koehler et al., 2014; Masciocchi et al, 2009). In the Koehler et al. (2014) study, a group of subjects viewing natural images was told to select regions or objects that are most salient with mouse clicks. The experimenters described salient as something that stands out and gave the example of a red follower among a field of common daisies. They found that saliency map models predict such explicit judgments of saliency, measured by the regions that are clicked, significantly better than fixated regions during free viewing and object search. Thus, saliency map models are better at detecting what people subjectively find salient, compared to the prediction of fixation locations. Please note that explicit saliency and interestingness, even if they share similar decision making processes, are qualitatively different, since saliency ratings prepare the subject for the detection of locally varying features as can be seen from the example used in the experimental instructions of Koehler and colleagues (2014). The task of we find in the earlier study (Masciocchi et al., 2009), on the other hand, is identical to our task. The researchers have gathered interestingness ratings from a large group of internet users and evaluated the computational saliency (Itti & Koch, 2000) of regions that are evaluated as interesting. Moreover, they collected fixation data from a different group of participants viewing the same images. Their results are comparable to our findings: Whereas saliency modeling predicts fixated and clicked regions with some success, the overlap between fixations and interesting regions is much stronger (Masciocchi et al., 2009). The researchers draw a causal link between saliency and interestingness and claim that their results show that salient locations that should be fixated are interesting. Nevertheless, as we have shown in our analysis (Onat et al., 2014) employing joint distributions of interestingness and low-level saliencies, there is much more top-down to interestingness as claimed by Massciocchi and colleagues (2014).

3.4. Discussion

The four studies (Açık et al, 2008; 2009; 2014, Onat et al., 2014) covered so far make clear that there is no single rule stating or refuting that high feature values cause higher fixation probability. Whereas our local contrast modification studies uncover the non-causal roles of features in attention guidance, the motion study suggests that both real and implied motion, and especially their onset, act as strong cues drawing attention. So what determines whether and when a local feature is going to attract attention?

We have demonstrated (Açık et al, 2008; 2009) that the salience of local contrast modifications depends on global and higher-order statistics of the scene. Recent approaches to fixation selection combine such holistic properties, object identities, and low-level saliency (Ehinger et al., 2009; Kanan et al., 2009; Wischnewski et al., 2010) in order to model eye-movements. If we consider that eye movements are goal-directed actions, which let the agent get coupled with the environment (even in the absence of an explicit task), being tuned to contextual information defined by higher-order statistics (Torralba et al., 2006) appears extremely relevant. Torralba and colleagues (Torralba et al., 2006) derive scene-based context information from global statistics, in order to limit the search to certain portions of the scene (e.g., no search for animals in the upper part of the image, if the scene includes the sky). Complementary to the inclusion of prior knowledge of location, other models employ information concerning how a target might appear in a given scene (Kanan et al., 2009). These models are shown to outperform classical approaches considering only local salience. Other examples of how knowledge is incorporated into eye movement planning come from studies investigating natural task performance such as driving a car, preparing breakfast, playing cricket, and walking in a crowded environment (Cristino & Baddeley, 2006; Land & Hayhoe, 2001; Land & McLeod, 2001). Together with teaching how our skills can be read out from fixation selection tendencies, these studies also characterize how targets and timings of eye movements change while task performance improves (Land & McLeod, 2001). Our contrast modification studies (Açık et al, 2008; 2009) reveal that if there is structural information about the global properties of the scene, local luminance contrast modifications are not salient. Einhäuser et al. (2006) suggest that strong contrast modifications on landscapes attract attention since they assume an object-like quality. We argue that the same holds for phase scrambled images that are devoid of meaningful structure. Nevertheless, if the image under consideration is characterized by a high amount of structure, which could equal an intrinsic richness of object-like properties in the scene, the modifications are not treated as objects by the visual system.

Our motion study (Açık et al, 2014) revealed that fixation probability is high for locations featuring movement components, even if the stimulus lacks real motion. The latter observation could suggest that the relationship between movement features and fixations is correlational. Alternatively, one could argue that static stimuli are treated as if dynamic, since the default mode of perception assumes a dynamic world (Gibson, 1972). This would require a mechanism that would extract motion from static cues. Psychophysical (Freyd, 1987) and neuroimaging (Kourtzi & Kanwisher, 2003; Krekelberg et al., 2003) studies of implied motion processing demonstrate that motion cues inferred from static stimuli lead to illusory percepts, distorted memories, and cortical motion center (MT and MST) activations. Accordingly, our static stimuli, albeit devoid of real motion, still contained implied motion information. Studies employing simple artificial stimuli suggest that motion is an especially salient feature, having the potential to cause higher fixation probability. Motion areas of the cortex, MT and MST, are part of the “where” pathway (Ungerleider & Mishkin, 1982), and provide direct input to posterior parietal cortex (Baizer et al., 1991), known for its role in attentional and visuomotor control (Behrmann et al., 2004). The observation that static motion cues activate MT and MST supports fixation selection based on movement, but that is the kind of movement that is inferred from static cues such as object identity. That this inference is based on the previous experiences of an observer about potentially moving and stable objects reveals the involvement of a top-down process. Even if we accept the notion that motion is causally related to eye movements, this doesn’t exclude interactions between this feature and global scene statistics, which could be addressed in future studies.

Finally, the interestingness study (Onat et al., 2014) reveals that whereas individual (marginal) or joint distributions of feature values can be employed in order to predict fixated locations above chance levels, the relationship between these locations and image regions that are evaluated as interesting is much stronger. A close inspection of co-occurrence of levels of interestingness and feature values fails to deduce behavioral relevance from pure stimulus statistics. Given that the study employs a large feature bank that includes most investigated features that are computed for saliency map modeling, we argue that even in the absence of a behavioral task, eye-guidance can’t be reduced to local image statistic computation and involves top-down contextual information sensitive to behavioral relevancy.

4. Age-related changes in eye movement behavior

We have argued that knowledge about global scene structure regulates the relationship between fixation probability and local features, and that this is a top-down mechanism (Açık et al, 2008; 2009). Typical definitions of top-down processing in perception and attention (Goldstein, 2011) involve the goals, expectations and other prior knowledge of the perceiving agent, which already exist before the encounter with the stimulus, and hence relate to the past states of the observer. Bottom-up processing, on the other hand, is exclusively stimulus-driven and corresponds to the current moment in time (Corbetta & Shulman, 2002). Thus, in order to prove that global scene statistics guide attention in a top-down manner, we have to demonstrate that their employment is related to the previous experiences of the observer. Otherwise, one can argue that global statistics are properties that are derived purely from the stimulus, perhaps thanks to a parallel spatial analysis system, enabled by large receptive fields of neurons found in areas high in the visual hierarchy of the cortex (Alonso & Chen, 2009).

One way to address the role of experience in fixation selection is contrasting the viewing behavior of observers with different amount of visual experiences. In the final study we compare feature-related viewing tendencies of different age groups in order to address whether knowledge derived from experiences influences eye movements in a top-down manner.

We have recorded the eye movements of 7- to 9-year-old children, 19- to 27-year-old adults, and older adults above 72 years of age during viewing of natural and complex images. The images came from different categories, namely landscapes, urban scenes, fractal images, and colorful phase scrambled images. A post-viewing patch-recognition task ensured attentive viewing without confounding eye movements with active visual search. Twelve different local features addressing different aspects of local statistics were employed. Even though performance, content-independent viewing characteristics, and fixation probability as a function of feature values were analyzed and discussed in the paper, here we limit ourselves to age-related changes in feature-related viewing.

Does the magnitude of the correlation between fixation probability and local feature values change as a function of age? In order to address this question, we have limited our analysis to the image categories where the features were reliably salient, i.e. urban scenes and fractals. Low feature saliencies in landscapes and phase-scrambled stimuli are not surprising and replicate the findings in our other reports (Açık et al., 2009; Onat et al., 2014). In urban scenes and fractal images there was a clear age-related decline in feature-related viewing: Children displayed the highest feature values at fixated points, older adult viewing behavior was the least feature-dependent, and young adult data fell in between. At first sight these observations might suggest an age-related decrease in the involvement of bottom-up mechanisms, an increase in top-down guided viewing, or both. However, fixation probability and local feature relationship informs us only about the correlation between the two, and is not always reflective of the causal contributions. For instance, in our phase scrambling experiments (Açık et al., 2008) we have observed an increase in the saliency of contrast modifications with the addition of phase noise, but the saliency of contrast itself was reduced. Similarly, even though local contrast at fixations on fractal images is quite high, local contrast increases do not draw attention. In order to understand the true nature of age related changes in viewing behavior, we have analyzed the so-called inter-observer saliency that quantifies to what degree participants’ fixation locations agree with each other. We revealed that inter-observer saliency drops with age. The locations fixated by different children are similar to each other, the similarity is lower in the case of young adults, and idiosyncratic viewing is at its maximum in the oldest population tested. This suggests that with the build-up of individual visual experiences different observers end up inspecting different portions of the same image. Thus, our results convincingly demonstrate that with more life-experience, individual top-down strategies overshadow bottom-up aspects of fixation selection, which is paralleled by a decrease in feature-related viewing.

5. Summary

Natural vision is characterized by successive sampling of the environment via fast eye-movements, which bring regions of a scene to fovea for high-resolution processing during fixations. During this spatially central processing, a decision concerning the next eye movement is made, and it is natural to assume that the decision relies on the available information at the periphery of the visual field. The works presented here (Açık et al, 2008, 2009, 2010, 2014; Onat et al., 2014) reveal that the decision concerning fixation selection involves additional knowledge beyond local statistics. Our studies demonstrate that local increases and decreases in certain features don’t influence fixation probability monotonically, and that global statistics of the scene modulate the degree of feature-related viewing. Moreover, both the magnitude of the correlation between fixation probability and local features, and the agreement in fixation locations of several observers decrease with age, suggesting a more stimulus-driven fixation pattern in the case of children, which is gradually replaced with idiosyncratic viewing patterns as a result of individual visual experiences. Whether top-down mechanisms modulate reliance on motion cues, which appear here as good discriminators of fixations even in the absence of real motion (Açık, 2014), remains to be addressed in future studies. Nevertheless, our demonstration that saliency of behaviorally relevant interestingness is superior to feature saliencies, and the fact that its contribution to fixation discrimination can’t be explained away statistically by stimulus characteristics (Onat et al., 2014) shows the adequacy of considering top-down mechanisms. Thus, our studies make a substantial contribution to eye movement literature showing that even during free-viewing, the relationship between local features and fixation probability cannot be characterized independent of global scene statistics and other behaviorally relevant contextual knowledge of the perceiver.

6. Limitations & Future Directions

We have presented a series of eye-movement studies performed in a laboratory that reveal the interaction between bottom-up and top-down processes during guidance of overt visual attention. Even though the presentation of natural and complex stimuli increases the ecological validity of our findings, our lab conditions are definitely no analog of the real world. During daily interactions with the visual world, people rarely view an environment without binocular cues and through a rectangular aperture while sitting passively for an extended period of time. Accordingly, the generalizability of the findings reported here will depend on future studies comparing eye-movements recorded in the lab with those performed during interactions with the real world.

Even though our studies as a whole describe many limitations on bottom-up allocation of attention, there are still unanswered questions that await future studies. In our phase modification study (Açık et al., 2008) we have argued that in fractals, higher-order scene properties go hand-in-hand with high contrast saliency. After phase randomization the intrinsic contrast at fixated and non-fixated regions was equal, but contrast modification saliency was very high. Fractal data from the developmental study (Açık et al., 2010) demonstrated that children had the highest low-level feature related viewing in the case of fractal images. Accordingly, one needs to address whether children are more attuned to higher-order properties, or to contrast itself. In that respect, a local contrast modification study with different age cohorts would be very informative. In the motion study (Açık et al., 2014) it was not possible to reveal whether saccades are predictive of or reactive to local movement saliency. Studies employing complex but local motion onset controlled stimuli might provide clear-cut answers. Moreover, temporal properties of saccade generation during viewing dynamic stimuli might depend on the task. It could be that in the absence of a clear task, the saccades are reactive to salient changes in the environment, but if the viewing agent has certain goals, it might be better to allocate attention to scene regions where changes are expected to happen. Thus, the studies presented here have the potential to inspire future studies in order to elucidate top-down and bottom-up interaction in gaze allocation further.

At a higher level of description, I argue that saliency map models and similar approaches are products of cognitivist and connectionist traditions in cognitive science (Thompson, 2007; Varela et al., 1992). Despite their particular differences, both traditions view cognition as a form of computation in a modular system that operates upon observer's mental representations of the outside world (Thompson, 2007). If viewed like that, the goal of perception becomes information extraction from the outside world in order to build mental representations that will guide future behavior in conjunction with already existing representations in memory (Gregory, 1974). It is easy to describe overt visual attention with this type of discourse: The spatial distribution of light falling on the retina will generate neural activity that encodes the outside world. With the hierarchical and parallel organization of visual information processing, the mental representations will take the form of several feature computations in different parts of the cerebral cortex. Neurally embedded rules will combine the several representations in order to generate the final topographical saliency representation of the scene. The saliency map is then fed to the motor components of the system in order to perform an eye-movement. As a result of the eye-movement, the spatial distribution of photochemical activity on the retina will change and the mental representations are going to be updated accordingly. Computational saliency map models provide us, that is, the researchers, with real-world approximations of elusive mental representations that we see and use in research. Albeit extremely fruitful with respect to research outcome and inspiring for practical applications such as artificial systems capable of detecting particular objects in complex scenes, purely bottom-up models of attention with their reliance on computationalist representations are slowly being left behind with the so-called pragmatic turn (Engel et al., 2013) in cognitive science.

The current alternative to the practice of explaining cognition as a set of computations performed on representations can be found in situated and embodied approaches (Gibson, 1979/1986; Thompson, 2007; Varela et al., 1992). Such approaches characterize cognition as inseparable from and intertwined with action, in which purposeful behavior of the agent is treated in the context of the body that enables the behavior and the situation in which the behavior comes about. In regard to vision, Gibson (1979/1986), well ahead of his time, wrote that "one sees the world not with eyes but with eyes-in-the-head-on-the-body-on-the-ground" (p. 205). Gibson always maintained that the information carried in the light that reaches the retina is all one needs for visual perception, and that this perception is completely *direct* in the sense that there are no internal representations, rendering his theory a radical bottom-up theory. Nevertheless, his emphases on the agent with a body and the task-informed interaction with the environment qualify him as the first embodied vision theorist. From the late nineties onward, eye movement researchers interested in agent-environment coupling began to study overt attentional processes while participants were performing real world tasks such as making tea (Land et al., 1999), playing cricket (Land & McLoad, 2000), walking (Patla & Vickers, 2003), and washing hands (Pelz & Canoza, 2001). Rather than focusing on feature values at fixated locations, this family of studies enlightens how fixations and saccades serve particular functions during task execution, such as finding an empty spot on a kitchen table in order to place a tea kettle (Land et al., 1999). Please note that an empty table surface is going to have very low local feature values in saliency computations due to the uniform surface, but such locations are fixated immediately since task demands override low-level saliency (Einhäuser et al., 2008a). These examples are illustrative of situated and embodied approaches to cognition, since they address how actions - including eye movements - accomplish the kind of coupling between an agent and its environment during the performance of a task with discernible goals. Thus, embodied and situated approaches to overt attention are responsible for an increase in the number of eye movement studies addressing how agents solve particular problems in real world environments (Land & Tatler, 2009). I believe that the future of research on eye movements under natural conditions lies in combining saliency modeling with embodied and 'pragmatic' (Engel et al., 2013) approaches to cognition.

How can we integrate ideas coming from saliency research with embodied and situated approaches to eye-movements? Schumann and colleagues (2008) used the mobile eye-tracking system "EyeSeeCam" in order to address feature saliencies while subjects were walking in different indoor or outdoor real-world environments such as museums, train stations and parks. Even though walking, as a whole, is a natural task, according to the momentary goal of the agent such as approaching an object, searching for a place to sit, or following a person, the viewing strategy might change and reveal varying degrees of reliance on low-level features. Thus, saliency and eye movement analysis can be employed according to the details of the agent-environment interaction (cf. Land & Tatler, 2009, especially the chapters on domestic activities, driving and locomotion). We have observed that both real and implied motion is able to predict fixated location at above-chance levels (Açık et al., 2014). Nevertheless, the relation of optical flow patterns during real locomotion to low- and high-level scene properties remains to be addressed and could be informative for understanding how we understand action opportunities in an environment in terms of what Gibson has called affordances (Gibson, 1979/1986). Such hybrid approaches using local or even global saliency computations in order to characterize goal-directed fixation behavior could serve the transition from cognitivist to enactive traditions in cognitive science (Engel et al., 2013; Varela et al., 1991).

References

Açık, A. (2006/2010). *Effects of luminance contrast and its modifications during viewing of images from different categories.* Osnabrück, Germany: Publications of the Institute of Cognitive Science.

Açık, A., Bartel, A., and König, P. (2014). Real and implied motion at the center of gaze. *Journal of Vision*, *14*(1):2, 1-19.

Açık, A., Onat, S., Schumann, F., Einhäuser, W., and König, P. (2009). Effects of luminance contrast and its modifications on fixation behavior during free viewing of images from different categories. *Vision Research*, *49*, 1541–1553.

Açık, A., Sarwary A., Schultze-Kraft, R., Onat, S., and König, P. (2010) Developmental changes in natural viewing behavior: bottom-up and top-down differences between children, young adults and older adults. *Frontiers in Psychology*, *1*:207. doi: 10.3389/fpsyg.2010.00207.

Açık, A., Steger, J., and König, P. (2008). Effects of phase scrambling and contrast modifications on fixation behavior during viewing of fractal images *FENS Abstr., vol.4, 056.1, 2008*.

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science*, *10*, 364 – 376.

Alonso, J-M, and Chen, Y. (2009). Receptive field. *Scholarpedia*, *4*(1):5393

Baizer, J. S., Ungerleider, L. G., and Desimone, R. Organization of visual input to inferior temporal and posterior parietal cortex in macaques. *The Journal of Neuroscience*, *11*, 168 – 190.

Behrmann, M., Geng, J. J., Shomstein, S. (2004). Parietal cortex and attention. *Current opinion in neurobiology*, *14*, 212 – 217.

Böhme M., Dorr M., Krause C., Martinetz T., and Bartz E. (2006). Eye movement predictions on natural videos. *Neurocomputing*, *69*, 1996–2004.

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

Corbetta, M., and Shulman, G. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*(3), 201–215.

Cristino, F., and Baddeley, C. (2009). The nature of the visual reperesentations involved in eye movements when walking down the street. *Visual Cognition*, *17*(6/7), 880–903.

Dorr M., Martinetz T., Gegenfurtner K. R., and Barth E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, *10* (10): 28, 1–17, http://www.journalofvision.org/content/10/10/28, doi:10.1167/10.10.28.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 Scenes: A combined source model of eye guidance. *Visual Cognition*, *17*, 945-978.

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

Einhäuser, W., Rutishauser, U., & Koch, C. (2008a). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision, 8*(2):2, 1–19, http://journalofvision.org/8/2/2/, doi:10.1167/8.2.2.

Einhäuser, W., Spain, M., and Perona, P. (2008b). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14):18, 1–26, http://journalofvision.org/8/14/18/, doi:10.1167/8.14.18.

Engel, A. K., Maye, A., Kurthen, M., and König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Sciences*, *17*, 202-209.

Frey, H.-P., Wirz, K., Willenbockel, V., Betz, T., Schreiber, C., Troscianko, T., and König. P. (2011). Beyond correlation: do color features influence attention in rainforest? *Frontiers in Human Neuroscience*, *5*:36. doi: 10.3389/fnhum.2011.00036.

Freyd, J. J. (1987). Dynamic representations. *Psychological Review*, *94*, 427-438.

Gibson, J.J. (1972). A Theory of Direct Visual Perception. In J. Royce, W. Rozenboom (Eds.). *The Psychology of Knowing*. New York: Gordon & Breach.

Gibson, J.J. (1979/1986). *The ecological appraoch to visual perception*. East Sussex: Psychology Press.

Goldstein, E. B. (2011). *Cognitive Psychology* (3rd (International) ed.). Belmont, CA: Wadsworth.

Gregory, R. (1974). *Concepts and Mechanisms of Perception*. London: Duckworth.

Henderson, J. M., and Hollingworth, A. (1999). High-level scene perception. *Annual Reviews of psychology*, *50*, 243 – 271.

Hubbard, T. L. (1995). Environmental invariants in the representation of motion: Implied dynamics and representational momentum, gravity, friction, and centripetal force. *Psychonomic Bulletin & Review*, *2*, 322–338.

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

Kanan, C., Tong, M. H., Zhang, L., and Cottrell, G. W. (2009). SUN: Top-down saliency using natural statistics/ *Visual Cognition*, *17*, 979 – 1003.

Kaspar, K., and König, P. (2012). Emotions and personality traits as high-level factors in visual attention: a review. *Frontiers in Human Neuroscience*, *6*:321. doi: 10.3389/fnhum.2012.00321

Koehler, K., Guo, F., Zhang, S., and Eckstein, M. P. (2014). What do saliency models predict? *Journal of Vision*, *14*(3):14, 1–27, http://www.journalofvision.org/content/14/3/14, doi:10.1167/14.3.14.

Kourtzi, Z., and Kanwisher, N. (2000) Activation in human MT/MST by static images with implied motion. *Journal of Cognitive Neuroscience*, *12*, 48–55.

Krekelberg, B., Dannenberg, S., Hoffmann, K.-P., Bremmer, F., and Ross, J. (2003). Neural correlates of implied motion. *Nature*, *424*, 674-677.

Land, M., Mennie, N., and Rusted, J. (1999). The Roles of Vision and Eye Movements in the Control of Activities of Daily Living. *Perception*, *28*(11), 1311-328.

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

Land, M. F., and McLeod, P. (2001). From eye movements to actions: how batsmen hit the ball. *Nature Neuroscience*, *3*, 1340–1345.

Land, M. F. & Tatler, B. W. (2009) *Looking and acting: vision and eye movements in natural behaviour*. Oxford, UK: Oxford University Press

Le Meur, O., Le Callet, P., and Barba, D. (2007). Predicting visual fixations on video based on low-level visual features. *Vision Research*, *47*, 2483–2498.

Machner, B., Dorr, M., Sprenger, A., von der Gablentz, J., Heide, W., Barth, E., and Helmchen, C. (2012). Impact of dynamic bottom-up features and top-down control on the visual exploration of moving real-world scenes in hemispatial neglect. *Neuropsychologia*, *50*, 2415–2425.

Marat, S., Phuoc, T. H., Granjon,L., Guyader, N., Pellerin, D., and Guérin-Dugué, A. (2009). Modelling spatio-temporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision*, *82*, 231–243.

Onat, S., Açık, A., Schumann, F., and König, P. (2014). The contributions of image content and behavioral relevancy to overt attention. *PLoS ONE*, *9*(4): e93254. doi:10.1371/journal.pone.0093254.

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

Parkhurst, D. J., and Niebur, E. (2004). Texture contrast attracts overt attention in natural scenes. *European Journal of Neuroscience*, *19*, 783–789.

Patla, A. E., and Vickers, J. N. (2003). How far ahead do we look when required to step on specific locations in the travel path during locomotion? *Experimental Brain Research*, *148*, 133–138.

Pelz, J.B., and Canosa, R. (2001). Oculomotor Behavior and Perceptual Strategies in Complex Tasks. *Vision Research*, *41*, 3587-3596.

Reinagel, P., and Zador, A.M. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, *10*, 1–10.

Rothkopf, C. A., Ballard, D. H., and Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision*, *7*(14):16, 1–20, http://journalofvision.org/7/14/16/, doi:10.1167/7.14.16.

Schumann F., Einhäuser-Treyer W., Vockeroth J., Bartl K., Schneider E., and 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.

Schütz, A. C., Braun, D. I., & Gegenfurtner, K. R. (2011). Eye movements and perception: A selective review. *Journal of Vision*, *11*(5):9, 1–30, http://www.journalofvision.org/content/11/5/9, doi:10.1167/11.5.9.

't Hart, B. M., Vockeroth, J., Schumann, F., Bartl, K., Schneider, E., and König, P. (2009). Gaze allocation in natural stimuli: Comparing free exploration to head-fixed viewing conditions. *Visual Cognition*, *17*, 1132–1158.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Belknap.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of attention in natural scenes. *Psychological Review*, *113*, 766-786.

Ungerleider, L. G. and Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield (Eds.). "*Analysis of Visual Behavior*", pp. 549–586. Cambridge, MA: MIT Press

Wilming, N., Betz, T., Kietzmann, T. C., and König, P. (2011). Measures and limits of models of fixation selection. *PLoS ONE 6*(9): e24038. doi:10.1371/journal.pone.0024038.

Wischnewski, M., Belardinelli, A., Schneider, W. X., and Steil, J. J. (2010). Where to look next? Combining static and dynamic proto-objects in a TVA-based model of visual attention . *Cognitive Computation*, *2*, 326–343.

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, Mass: MIT Press

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

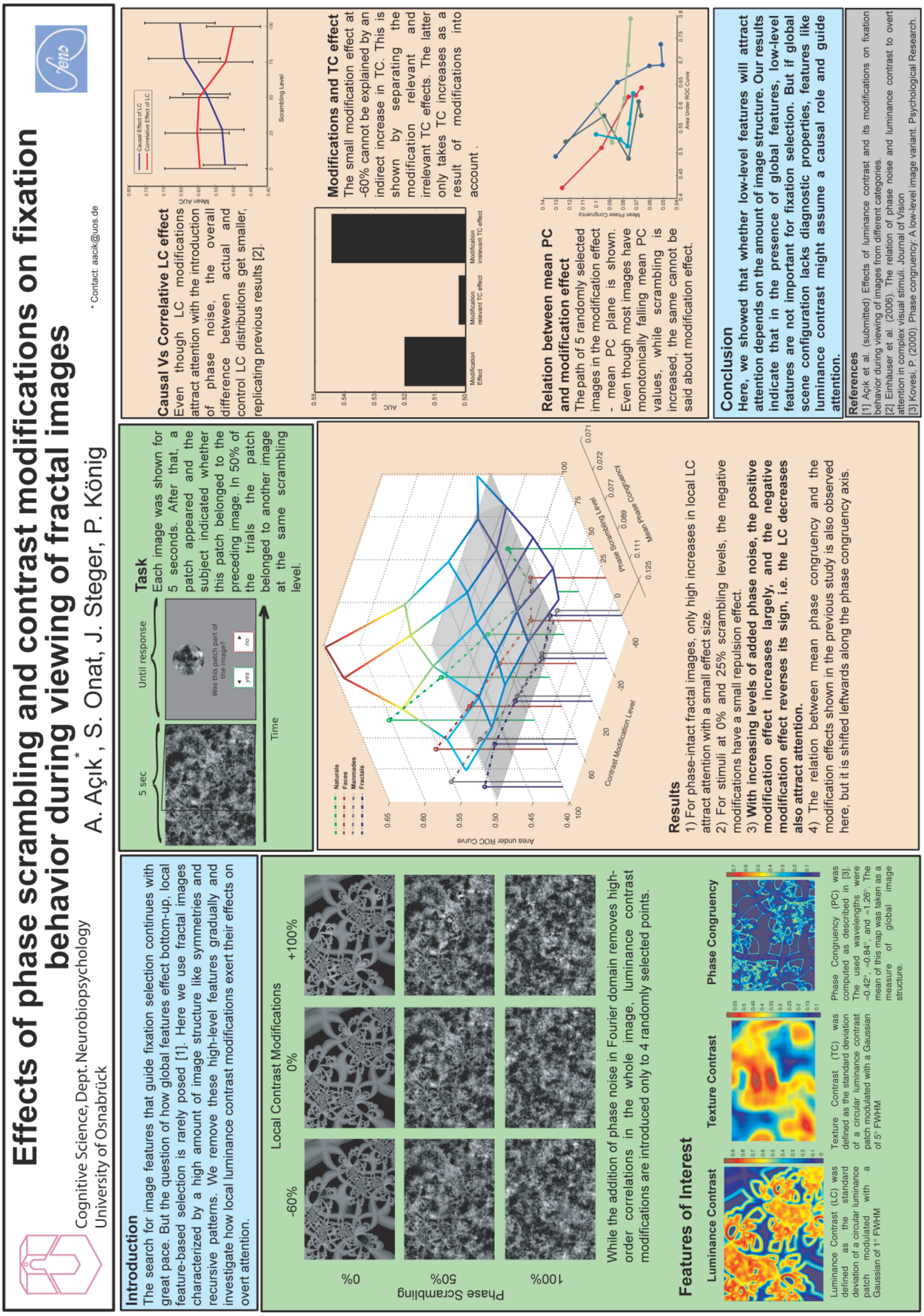
Published articles that are contained in this thesis:

1) Açık, A., Bartel, A., and König, P. (2014). Real and implied motion at the center of gaze. *Journal of Vision*, *14*(1):2, 1-19.
   doi:10.1167/14.1.2
   http://www.journalofvision.org/content/14/1/2

2) Açık, A., Onat, S., Schumann, F., Einhäuser, W., and König, P. (2009). Effects of luminance contrast and its modifications on fixation behavior during free viewing of images from different categories. *Vision Research*, *49*, 1541–1553.
   doi:10.1016/j.visres.2009.03.011
   http://www.sciencedirect.com/science/article/pii/S0042698909000947

3) Açık, A., Sarwary A., Schultze-Kraft, R., Onat, S., and König, P. (2010) Developmental changes in natural viewing behavior: bottom-up and top-down differences between children, young adults and older adults. *Frontiers in Psychology*, *1*:207.
   doi:10.3389/fpsyg.2010.00207
   http://journal.frontiersin.org/article/10.3389/fpsyg.2010.00207/full

4) Onat, S., Açık, A., Schumann, F., and König, P. (2014). The contributions of image content and behavioral relevancy to overt attention. *PLoS ONE*, *9*(4): e93254.
   doi:10.1371/journal.pone.0093254
   http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0093254



Açık, A., Steger, J., and König, P. (2008). Effects of phase scrambling and contrast modifications on fixation behavior during viewing of fractal images *FENS Abstr., vol.4, 056.1, 2008.*