Journal Of Eye Movement Research
2(2):2, 1-11

Semantic Override of Low-level Features in Image Viewing – Both Initially and Overall

Marcus Nyström
Lund University

Kenneth Holmqvist
Lund University

Guidance of eye-movements in image viewing is believed to be controlled by stimulus driven factors as well as viewer dependent higher level factors such as task and memory. It is currently debated what proportions these factors contribute to gaze guidance, and also how they vary over time after image onset. Overall, the unanimity regarding these issues is surprisingly low and there are results supporting both types of factors as being dominant in eye-movement control under certain conditions. We investigate how low, and high level factors influence eye guidance by manipulating contrast statistics on images from three different semantic categories and measure how this affects fixation selection. Our results show that the degree to which contrast manipulations affect fixation selection heavily depends on an image's semantic content, and how this content is distributed over the image. Over the three image categories, we found no systematic differences between contrast and edge density at fixated location compared to control locations, neither during the initial fixation nor over the whole time course of viewing. These results suggest that cognitive factors easily can override low-level factors in fixation selection, even when the viewing task is neutral.

Keywords: Image viewing, contrast manipulation, semantic information dispersion, bottom-up, top-down

Introduction

The human visual system (HVS) is equipped with a high resolution fovea where detailed information about the visual environment is acquired, and a less sensitive periphery which samples the visual input very sparsely. In order to fully comprehend the visual world, thus, we move the eyes to provide the fovea with detailed information. Typically, the eyes move about three to four times per second by employing fast ballistic movements called saccades. In between the saccades, the eye is virtually stable in what is referred to as a fixation.

The guidance of eye-movements is generally attributed to bottom-up and top-down processing. Bottom-up processing implies that gaze guidance is controlled by low-level primitives such as contrast, luminance, and edge density (Treisman & Gelade, 1980). It is generally described as a fast and involuntary process. Top-down processing is somewhat more difficult to define precisely, but can be thought of as a semantic interpretation of the scene reflecting the interplay between higher cognitive factors such as a viewer's task, goals and familiarity with similar types of scenes (Sarter, Givens, & Bruno, 2001).

Gaze behavior and guidance of eye-movements in image viewing have been studied as early as in the 1930s by Buswell (1935), and later by Yarbus (1967). In particular Yarbus' work is well cited in the literature. Two of his main observations were that the task at hand heavily influences where people look and that 'informative' regions are looked at more than other regions. More recently, there have been a series of studies building on these pioneering works seeking to gain a deeper understanding of the mechanisms behind eye-guidance through eye-tracking experiments. For example, it has been investigated how different tasks affect eye-movements, and what makes a region informative.

To address eye-movement guidance from a bottom-up perspective, statistical differences in image content around visually attended regions and control regions have been compared. For example, Reinagel and Zador (1999); Parkhurst and Niebur (2003) report of higher luminance contrast around gaze positions than control regions, and Baddely and Tatler (2006) conclude that high-frequency edges are good predictors of fixated locations. The influence of bottom-up features on eye-movements has also been studied through computational frameworks defining the *salience* at different image locations through combinations of a number of low-level primitives (Itti, Koch, & Niebur, 1998). Salience has been shown to correlate with gaze positions better than at random (Parkhurst, Law, & Niebur, 2002), and has recently been reported to coincide with image regions deemed as important by human viewers (Elazary & Itti, 2008). Furthermore, it has been analyzed how low-level primitives relate to fixated image regions over the time course of viewing. Specifi-

DOI 10.16910/jemr.2.2.2

ISSN 1995-8692

cally, Parkhurst et al. (2002) and Itti (2006) report that bottom-up processing is more influential early after stimulus onset. However, these findings are not supported by Tatler, Baddeley, and Gilchrist (2005), who argue that bottom-up features are equally influential over time, whereas top-down influences increase as a function of viewing time. It is further known that bottom-up control of eye-movements is less influential as the saccadic amplitude increases (Tatler, Baddeley, & Vincent, 2006); the landing positions of long saccades are hard to predict given the feature content available to a viewer when the saccade is initialized (Rajashekar, Linde, Bovik, & Cormack, 2007).

Complementary to the empirical evidence supporting that gaze guidance is controlled by the physical properties of a stimulus, there are several examples of cognitive factors known to influence where people look, of which some are listed by Henderson and Ferreira (2004): Short-, and long-term episodic scene knowledge, scene schema knowledge, and task knowledge. Some of these factors have shown to override prediction based on saliency; in line with Yarbus' work, certain task instructions have been shown to critically influence where people look, reducing saliency map predictions to a chance level (Underwood, Foulsham, Loon, Humphreys, & Bloyce, 2006; Rothkopf, Ballard, & Hayhoe, 2007). Henderson, Brockmole, Castelhano, and Mack (2007) showed that image patches extracted around fixated locations not only contained lower intensity as well as higher contrast and edge density than control locations, but these image patches were also deemed more semantically important. This raises the question whether semantic importance, instead of saliency, dominantly influences fixation selection. Interestingly, Einhäuser and König (2003), found no significant change in where people looked as an effect of moderate manipulations in local image contrast at a number of random image locations.

Besides that image properties and cognitive factors influence eye-movements, it is known that eye-movement parameters such as fixation durations, saccade lengths and saccade directions also depend on previous and future eye-movements. For example, Tatler and Vincent (2008) argued that knowledge about such systematic tendencies in eye-movement behavior could, together with bottom-up and top-down processing, be an important factor toward a more coherent theory about eye-movement guidance.

Clearly, there is a large body of research on gaze behavior and fixation selection in images, of which some are listed above. The efforts to pursue these issues have grown rapidly over the last years partly due to cheaper and more accessible eye-tracking technology. Most certainly, computational models of visual attention have helped in boosting this interest. Despite recent research efforts, and even though the mechanisms driving gaze guidance slowly are starting to unravel, the unanimity in results is surprisingly low. One reason for this may be that earlier studies dominantly use real-world photographs as stimuli, in which semantically interesting regions coincide with low-level features in a manner that is not under any experimental control. As a consequence, it is hard to conclude whether the reported high feature densities at fixation are *causal* or *correlative*. A causal effect would imply that fixation locations are chosen as a direct consequence of the signal strength of one or a set of combined low-level primitives. A correlative effect, on the other hand, would mean that fixations land on regions that happen to contain high feature densities, but are in fact guided to these regions by other, higher level mechanisms. For example, objects may be fixated since they contribute to the semantic representation of the scene, and not because they happen to contain e.g., high contrast.

In this paper, we will investigate how contrast and edge density contribute to fixation selection, and how this effect varies over time. Unlike the majority of previous studies, test images are contrast manipulated prior to display. Meanwhile, we aim to keep their semantic content intact. We believe that by decoupling objects (or regions) from their low-level signal strength, an analysis is more likely to elicit causal relationships between where subjects fixate and the reason why they choose to look there. Besides manipulating the image statistics, three image categories are used: Images naturally embedding faces, images with man-made objects, and images depicting scenes with neutral semantics (trees, leaves, etc.). Each class is chosen to represent images with different *semantic information dispersion (SID)*. We define this concept as follows:

Definition 1. *Semantic information dispersion (SID) measures how spread out the information is that subjectively best conveys the information of the whole image.*

For example, a face generally contributes more to the core meaning of an image then does a leaf on a tree. Consequently, an image has a low SID if a small part (such as a face) of the image is judged to contain the majority of conveyed information. The rationale for using different image categories is to introduce a varying top-down influence without using an explicit task. For example, the task *look at regions with uniform texture* would yield a low correlation between e.g., edge density and fixated image content, but would hardly reveal much about the mechanisms behind gaze guidance. To verify that the images chosen for the experiment indeed represent different levels of SID, an experiment is performed where subjects are asked to identify a fixed size region that best conveys the information of the whole image. The average overlap between the regions chosen by the subjects is then used to estimate the SID.

DOI 10.16910/jemr.2.2.2

ISSN 1995-8692

Figure 1. Test images comprise three semantic categories: Face images (top two rows), images with neutral semantics (row three and four), and images containing man-made objects (bottom two rows). Two contrast manipulated versions of each image are used in the experiments.

(a) (b) (c)

Figure 2. Contrast manipulation for face images. (a) shows the original image. In (b), the contrast is decreased away from the marker in (a), positioned over the woman's face. The figure in (c) illustrates the case where contrast instead is reduced toward the face area by inverting the contrast manipulation function in (b).

(a) (b) (c)

Figure 3. Contrast manipulation for images not containing faces. Figure (a) shows the original image with four candidate markers. One of these markers is chosen at random, and (b) illustrates the case when contrast is reduced away from this marker (in upper left corner). In Figure (c), the marker diagonally toward the randomly picked one is instead used as the point from were contrast is reduced.

Methods

Test images

Three semantic image categories are used. In the first category, we use images containing faces; it is known that faces are very semantically important image regions and therefore frequent fixation targets (e.g. Yarbus, 1967). The second category comprises images with neutral scene semantics and depicts scenes with motives from nature such as trees and bushes (from Einhäuser & König, 2003), grass, and a picture of a brick wall. The last category falls between the first two categories and contains man-made objects embedded in natural environments. Six images from each category are used. Images were converted to eight bit gray scale and resized to dimension 1024×768 through the Matlab functions `rgb2gray` and `imresize` (bilinear), respectively. The test images are shown in Figure 1. As can been seen, each image comes in two versions where contrast has been modified differently.

Face images are modified to form two subcategories. In the first subcategory faces were retained in high contrast, whereas other regions were gracefully reduced in contrast away from the facial region. In the second subcategory, these contrast modifications were inverted; only the facial regions were reduced in contrast. Figure 2 exemplifies this. For the other two categories, each image was transformed into two different versions as follows: Four candidate positions, same for all images, were available as shown in Figure 3. One of these positions was selected at random, and the first version was generated by reducing the contrast smoothly away from this position. The other version was generated in a similar manner, but now with the contrast being reduced away from the point diagonally opposite to the randomly selected position.

Image manipulation

Contrast manipulation was implemented by means of variable resolution image processing using Gaussian pyramids. A five level pyramid was created by iterative lowpass filtering and downsampling of the original image, followed by upsampling and (bilinear) interpolation back to the original image resolution (1024×768). Lowpass filtering was implemented by an ideal filter with a cutoff frequency adjusted to avoid aliasing given a subsampling factor of 2 pixels. These operations resulted in a collection of images where the original image comprised the bottom layer and higher layers were copies of the original image with increasingly lower contrasts. To create images with variable contrast, high resolution regions were selected from the bottom layer of the pyramid, whereas low resolution regions originated from the higher layers in the pyramid. Regions from different levels were then synthesized through a Gaussian shaped blending function. Let $I_\ell(m,n)$ denote an image at level ℓ in the lowpass pyramid. m and n span the image dimensions and $\ell = \{1,2,3,4,5\}$, where $\ell = 1$ denote the bottom layer comprising the original image. Then the implementation can be described by Algorithm 1. $I(m,n)$ is the output image, and $G(m,n)$ denotes a Gaussian func-

DOI 10.16910/jemr.2.2.2

ISSN 1995-8692

Algorithm 1 Implementing a variable contrast

1: $I(m,n) = I_1(m,n)$ {Initialize}
2: **for** $\ell = 2$ to 5 **do**
3: $\quad I(m,n) \leftarrow I(m,n) \cdot G(m,n) + I_\ell(m,n) \cdot (1 - G(m,n))$
4: **end for**

tion

$$G(m,n) = e^{-\left(\left(\frac{m-m_i}{\sigma}\right)^2 + \left(\frac{n-n_i}{\sigma}\right)^2\right)} \quad (1)$$

where (m_i, n_i) represents the point where the Gaussian function is centered, i.e., the point from where the image is increasingly reduced in contrast. To introduce a noticeable amount of blur, σ was set to $\sigma = 50$ pixels. σ was chosen simply by pilot testing where contrast reduction was deemed as significant without changing the semantics of the image. It has been pointed out in an earlier study (Parkhurst & Niebur, 2004), that when using contrast manipulations to study fixation selection, it is important to implement smooth contrast degradations to avoid undesired variation in higher order image statistics, which could explain possible changes in fixation behavior. To account for this observation, we implement very smooth, although noticeable contrast reductions.

Contrast manipulation for face images was implemented with the above parameters when contrast was reduced away from the face. However, in the opposite case, when contrast was reduced toward the face region (the face was blurred), then the blending function was modified to

$$G(m,n) = 1 - e^{-\left(\left(\frac{m-m_i}{\sigma}\right)^{1.5} + \left(\frac{n-n_i}{\sigma}\right)^{1.5}\right)} \quad (2)$$

in order to better limit the contrast reduction effect to the facial region.

Subjects

13 naive test subjects (25.7±4.9 years old, one female) were recruited to participate in the experiment. Their visions were normal or corrected to normal. Compensation was given in the form a lottery ticket and subjects consented to use of their data by signing a form.

Experiment I: Viewing contrast manipulated images

Contrast manipulated images from all three categories were shown one at the time in full screen. Before the presentation of an image, a central dynamic fixation marker in the form of solid black circle was shown on a mid-gray screen. The diameter of the circle was decreasing as a function of time. After one second, the circle disappeared and an image was displayed in full screen during a time randomly drawn from the interval $t = [3,4,5,6]$ seconds. This procedure was repeated for all images, which were shown in random order. Varying display time was used to prevent subjects from adopting top-down strategies such as systematic scanning of the images. Prior to each image was displayed, subjects were asked to look at the fixation marker.

The instruction given to the subjects was to *please study the images carefully*. Supposedly, being a fairly general instruction, it prevents subjects to adopt individual viewing strategies trying to guess the purpose of the tests. For example, we saw in an earlier study (Nyström & Holmqvist, 2007), where subjects were given the more neutral instruction solely to *watch the images*, that subjects adopted a top-down strategy avoiding to look at the blurred regions a bit into the presentation. We believe that the task instruction used in this paper will alleviate this undesirable adaption.

Experiment II: Image semantics evaluation

In a second experiment, that followed right after the first, subjects were shown the 18 unprocessed (no contrast manipulation) images (in eight bit gray scale of dimension 1024×768), one by one in full screen. They were not informed about this evaluation until after the first experiment was completed. Superimposed on each image was a quadratic box that could be controlled by the mouse cursor. Subjects were asked to position the box over the area that best capture the core meaning of the image. A mouse click continued this procedure for next the image. The exact instruction was given in writing as: 'Position the box over a region that best conveys the information of the whole image'. The size of the box was chosen large enough to encapsulate whole objects or parts of objects, so that the meaning of the box content would be clear without access to the whole image. We used a box size that spanned four degrees (128×128 pixels).

Eye-tracking

Eye-tracking was preformed monocularly during both experiments with an SMI iView X Hi-Speed 1250 Hz system. Subjects were seated 0.67 m away from a 19 Inch Samsung GH19PS screen with the resolution and update rate set to 1024×768 pixels and 60 Hz. The physical dimension of the screen was 380×300 mm, spanning 32×25 degrees of visual angle. Each recording started with a 13-point calibration. Stimuli presentation, communication with the eye-tracker, and data analysis were performed with Matlab and the Psychophysics Toolbox Version 3 (PTB-3) (Brainard, 1997). A saccade based detection scheme developed by SMI (IDFconvert.exe) was used to filter out event based measures such as fixations and saccades. Gaze positions were classified as saccades if the eye velocity was $\geq 75°/s$ and if the saccade duration lasted ≥ 10 ms. If these assumptions were violated, and the eye was stable for $\geq$50 ms, a fixation was detected.

DOI 10.16910/jemr.2.2.2

ISSN 1995-8692

Analysis and results

In this section data is visualized and analyzed. The analysis addresses the following questions: 1) Are contrast and edge density different at fixated regions compared to control regions for contrast manipulated images? 2) Do contrast manipulations change where people look? 3) Is the magnitude of change related to image semantics, and in terms of semantic image dispersion (SID)?

What do we look at? - Feature analysis

It is known from several previous studies that certain low-level features are elevated at fixated positions. For example, fixated locations tend to have higher contrast and edge density than non-fixated, control regions. We begin our analysis by testing whether these observations still hold using contrast manipulated images. Contrast at the image location (m,n) is defined as the standard deviation within a 3×3 neighborhood centered at (m,n). Edge density is extracted by convolving the image separately with horizontal and vertical Sobel operators, and then computing the average of these filtered outputs.

In the analysis, an approximately 1 degree (32×32 pixel) region is extracted from the feature maps around each fixation location. For comparison, equal sized regions are also extracted from control locations, and the difference between fixated and control feature contents is analyzed. Instead of using uniform sampling over the image area to simulate a random viewer, we use control fixations collected from other images used in the experiment. This way, a simulated 'random' fixation pattern coincides with the distribution of fixations, which is known to be non-uniform with a bias to the center of the display. It has been argued that the central bias may give rise to artificially high features values at fixation (e.g., Tatler et al., 2005), and should therefore be carefully accounted for in the analysis. More insights about the central bias effect are given by Tatler (2007).

An increasingly popular method to estimate the degree to which fixated and control feature content can be differentiated from each other is the receiver operating characteristics (ROC) analysis (e.g., Hanley & McNeil, 1982). A ROC curve plots the fraction of *true positives* (TP) against the fraction of *false positives* (FP). In our case, TP consist of fixated feature content, whereas FP comprise feature content at control locations. The area under the ROC curve varies between zero and one, and is a robust measure of how well image features can be discriminated between fixated and control locations; if the ROC area is significantly larger than 0.5, a tested feature is said to discriminate fixated locations from control locations. A ROC area that equals 1 is said to give perfect classification.

Figure 4 plots the average ROC areas for contrast and edge density. Black bars represent results considering the first fixation (from all subjects in all images) only, whereas the white bars represent a similar analysis over all fixations. By the *first fixation*, we mean the fixation following the initial saccade after image onset and not the first registered fixation is the data file, which is constrained to the center of the screen by a fixation marker. As reported by several previous studies, feature densities at fixated locations are significantly higher (ROC area > 0.5) than feature densities at control locations ($p < 0.01$, t-test, for both contrast and edge density). Apparently, this is also true for contrast manipulated images. Moreover, there is a tendency, although non-significant, that initial fixations discriminate contrast and edge density better than fixations do over the whole time course of viewing.

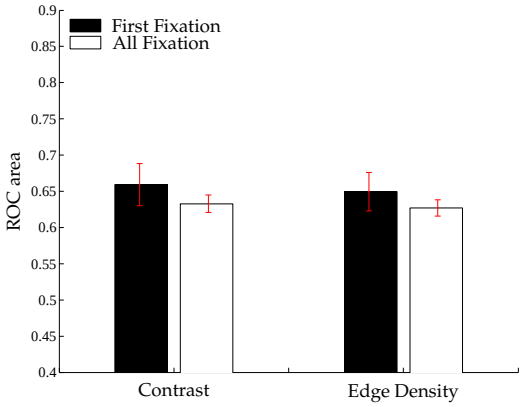

Figure 4. ROC areas for discrimination between image features at fixated and control locations. Black bars show ROC areas for the first fixation whereas all fixations are included in the white bars. Error bars span standard errors of the mean. A ROC area larger than 0.5 indicates a difference.

Do image semantics and feature manipulations influence where we look?

To this point, our empirical findings are in line with previous results emphasizing bottom-up control over fixation selection. The findings show, *on average*, that contrast and edge density are higher at fixated positions than at other, control positions. In this section, it is investigated whether these general tendencies are consistent when analyzing images with regard to their semantic information dispersion (SID) as well as their direction of contrast reduction. What happens with peoples' allocation of fixations, for example, if a region deemed as semantically important is reduced in low-level signal strength? Obviously, a saliency based framework would predict an obligatory shift in fixation density away from this region.

Using data collected from the second experiment, we found the SID for each image, calculated as the average overlap between box locations within an image. Thus, if $B_{i,j}$ denotes a box in the image i positioned by subject

DOI 10.16910/jemr.2.2.2

ISSN 1995-8692

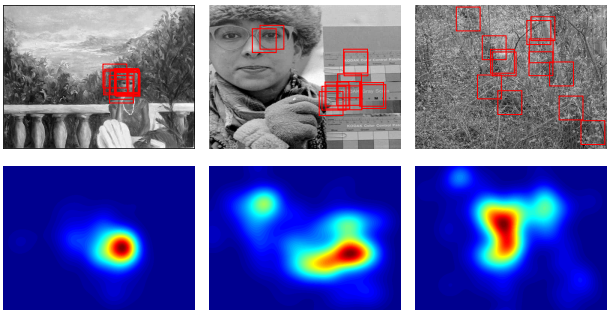

Figure 5. Images in order of increasing semantic information dispersion (SID). The top row shows where subjects have positioned a box that 'best conveys the information of the whole image'. The bottom row illustrates the fixation density of the same subjects while performing this task. As can be seen, the inter-subject agreement between fixation density and the regions judged to best convey the information of the whole image is large.

j, the SID for image number i is defined as

$$\mathrm{SID}_i = \left[\frac{2}{N(N+1) - 2N} \sum_{\substack{j=1,\ldots,N-1 \\ k=j+1,\ldots,N-1}} B_{i,j} \cap B_{i,k} \right]^{-1} \quad (3)$$

where $\cap$ denotes the intersection between the boxes in pixels, and N is the number of viewers. The inverse is computed such that a large SID value represents a spread out semantic information and vice versa. The top row in Figure 5 shows three of the unprocessed test images and the boxes as positioned by the test subjects. Out of the 18 unprocessed images used in the experiment, images with the lowest, midmost, and highest SID are shown in the figure. Unsurprisingly, the image with the lowest SID contains a face, and the image with the highest SID contains rather neutral semantics. For the sake of comparison, the fixation density of the same subjects while performing the SID detection task is given in the second row in the figure. For these images, the overlap between where subjects fixated and where they positioned the box is quite large. As expected, the image categories were tightly couple with SID; five of the six images containing faces were among the images with the lowest SID (boxes were dominantly positioned over the face), and all the six images from the 'neutral' category had the highest SIDs. Consequently, five images from the 'man-made object' class were located in the mid-SID section along with one face image.

Figure 6 illustrates how the fixation density changes as a result of contrast manipulations for images with low, medium, and high SID. The fixation densities are visualized as heat maps, where Gaussian functions have been centered at each fixation location and then superimposed. The variance of each Gaussian function has been set such that the width at half its maximum height approximates the size of the foveal span of a viewer in the current experimental setup. In addition, the height of each Gaussian function has been scaled in proportion to the fixation duration. As a consequence the fixation densities not only reflect where people have fixated, but also their level of cognitive processing during each fixation, hence providing more sensitive and detailed information. Henceforth, we refer to the heat maps as fixation density functions (FDFs), in order to better capture what the heat maps represent. The second column in Figure 6 depicts FDFs for all subjects during the first fixation, and the third column illustrates corresponding fixation densities collapsed over all fixations. This can be compared with the two rightmost column, where contrast and edge density are visualized. A visual inspection of the plots indicates that contrast and edge manipulations clearly influence where subjects look. However, the magnitude of change seems to differ depending on the image type; the images containing faces undergo relatively small changes in fixation placement due to contrast manipulation whereas fixations in the images that contain more neutral semantics seem to be more influenced by the manipulations.

To quantify how fixation locations change as a function of contrast manipulation and SID, the two-dimensional correlation coefficient between FDFs belonging to the two contrast manipulated versions of each image is computed. This metric has been used in other works for the same purpose (Rajashekar, Linde, Bovik, & Cormack, 2008). Although it is not clear how accurately the 2-D correlation coefficient, or any other metric for that matter, captures the difference between people's fixation locations, it gives an estimate that helps us to interpret the magnitude of change. For a reference of other metrics used to estimate the similarity between fixations, see for example S. Mannan, Ruddock, and Wooding (1995); Privitera and Stark (2000); Tatler et al. (2005). Since images' SID-values almost perfectly matched the initial division of images into three semantic categories, the analysis is preformed with respect to the image categories, which henceforth are referred to as 'Face', 'Man-made', and 'Neutral'. Figure 7(a) depicts the average 2-D correlation between FDFs generated from the initial fixation (black bars) and all fixation (white bars) within each category. It can be seen that the image category influences the degree to which contrast manipulations trigger shifts in fixation densities; images containing regions of high semantic importance, such as faces, are less sensitive to the manipulations than other images and in particular those from the 'Neutral' category. This tendency is present for both the initial fixation and for fixations over the time course of viewing.

Another way to represent how fixation locations are affected by contrast manipulations and semantics, shown in Figure 7(b), is to plot the shift in fixation density (2-D correlation coefficient between FDFs) against images' SID. Circles and triangles represent how the

DOI 10.16910/jemr.2.2.2

ISSN 1995-8692

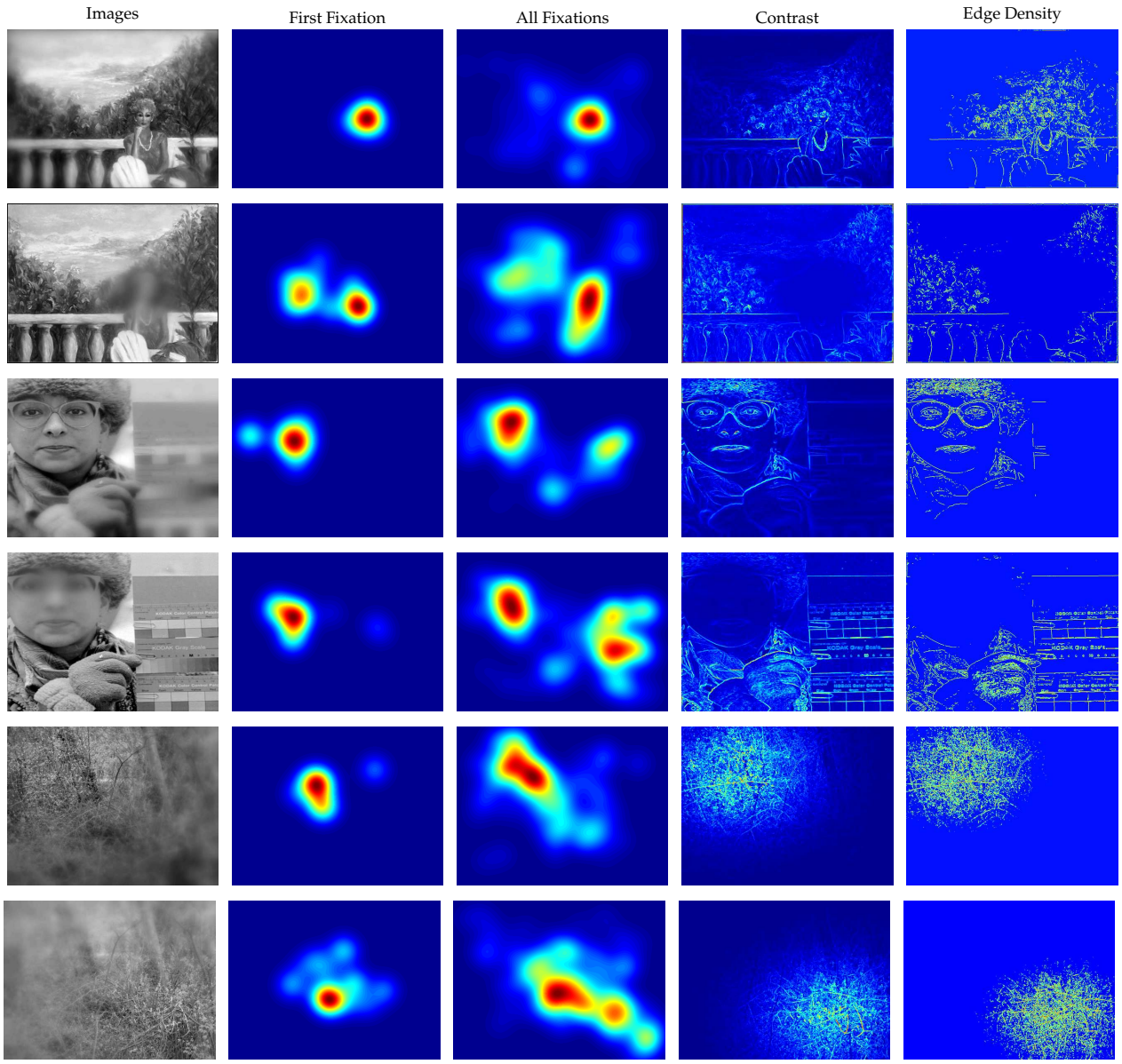

Figure 6. Effect of contrast manipulation on fixation behavior.

initial fixation and all fixations, respectively, are shifted in location as a function of SID. The lines are least square fits to the data points. Considering all fixations, it can been seen that SID clearly influences the magnitude of shift in fixation density, having a correlation of $\rho = -0.62$. This tendency is weak, or hardly present at all, considering the first fixation only. It may be the case since fewer fixations are used to generate the first fixation FDFs, giving individual fixations more weight. Consequently, a fixation that is not aligned with other fixations have a large impact on the shape of an FDF, and therefore also the value of the 2-D correlation coefficient between two FDFs.

In summary, the results from Figure 7 clearly illustrate that the degree to which fixation locations are influenced by contrast manipulations depends on SID and image category.

Since contrast manipulations change where people look with different magnitudes depending on images' SID, one would expect this to be reflected in fixated image content across the image categories. For example, in the category that was least influenced by the image manipulations, we would expect a lower discrimination for contrast and edge density between fixated and control locations than for the other two categories. Figure 8(a) plots average ROC areas for contrast and edge density over the three image categories. Results for both the first fixation and all fixations are given for each feature and category. As expected, the discrimination of features between fixated and control

DOI 10.16910/jemr.2.2.2

ISSN 1995-8692

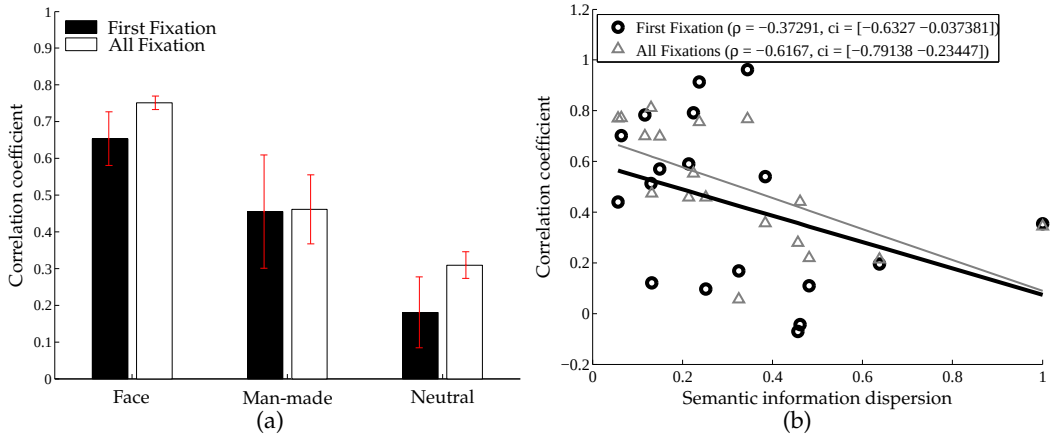

Figure 7. Influence of fixation selection on image category, SID, and contrast manipulations. (a) Bars represent the average shift in fixation density due to contrast manipulations within each image category. (b) The images are presented in order of increasing semantic information dispersion (SID). The solid lines are least square fits to the data points. Error bars span one standard error around the mean. 95% confidence interval of the correlation coefficient ρ are generated using bootstrapping with 1000 resampled sets (Matlab's `bootstrp` function).

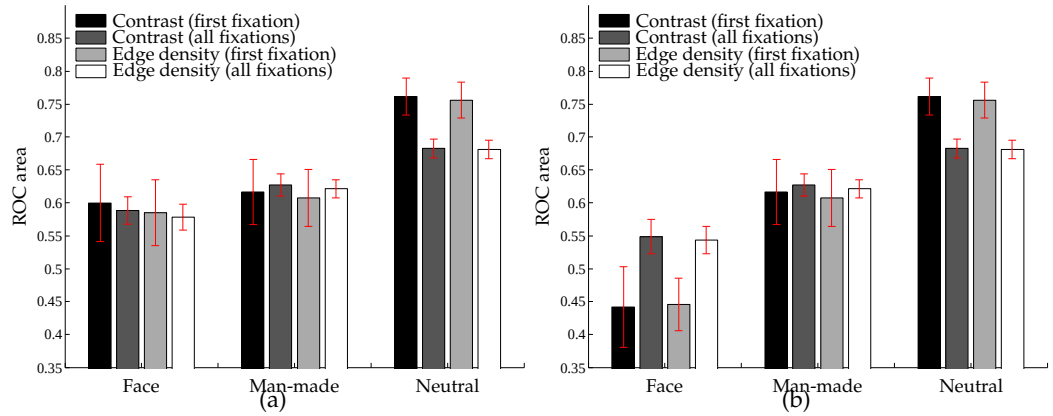

Figure 8. ROC analysis of contrast and edge density over different image categories. (a) All images from each category are included. (b) From the 'Face' category, only images with blurred faces are included.

locations were the lowest in the 'Face' category and increasingly higher for the 'Man-made' and 'Neutral' categories. However, it was significantly ($p < 0.05$, *t*-test) better than chance (ROC area > 0.5) in all cases. Also notice how ROC scores in the 'Neutral' category are significantly ($p < 0.05$) higher for first fixation than all fixations, whereas this tendency was not significant in the other two categories.

Figure 8(b) differs from Figure 8(a) in the way that only those images from the 'Face' category where contrast was reduced toward the face, i.e., where the faces were blurred, were included in the analysis. Since people still looked at the face regions after being reduced in contrast, the discrimination was reduced to a chance level, both considering the first and all fixations. Interestingly, discrimination was worse for feature content fixated at the initial fixation, contrary to the finding by Parkhurst et al. (2002).

Discussion

It has previously been reported that image features such as contrast and edge density are higher around fixated locations then control locations, and thus are likely to contribute in guiding eye-movements. We found this effect to be highly dependent on images' semantic content and how this content was distributed over the image. The effect was replicated in images with neutral semantics, but was not present in images where semantically important regions where reduced in low-level signal strength. These results suggest that semantic interpretations of an image easily can override bottom-up control of eye-movements.

There is a continuously increasing debate about what controls where we look in real-world photographs. In particular, it is debated how bottom-up and top-down factors interact to guide eye-movements toward regions of particular interest or relevance. On

DOI 10.16910/jemr.2.2.2

ISSN 1995-8692

the one hand, the physical properties of an image have shown to correlate well against locations fixated by human viewers, motivating computational approaches to fixation prediction by combining a set of image features with known high correlation (e.g., Itti et al., 1998; Itti & Koch, 2000). On the other hand it is known from, e.g., Buswell (1935) that higher level factors such as a viewer's task heavily influences where people look. For example, several recent studies have shown that some tasks easily overrides a prediction based on a saliency map (see Underwood et al., 2006; Rothkopf et al., 2007; Einhäuser, Rutishauser, & Koch, 2008). Currently, the nature of bottom-up and top-down control of eye-guidance is largely an open problem.

We investigated the contribution contrast and edge density as well as image semantics to the selection of fixations in images. Unlike the majority of earlier studies, the statistics of the tested images were experimentally manipulated, and the degree of change in fixation locations caused by the manipulations was measured. The degree of change was quantified over three image categories: Images containing faces, images with man-made objects, and images depicting scenes with rather neutral semantics. The semantic information dispersion (SID) was calculated for each image by letting subjects position boxes over the one region in each image that best conveyed the information of the whole image. The average spatial box-overlap across subjects defined the SID. Face images had the lowest SID followed by man-made objects and the neutral category. In the images containing faces, contrast was either reduced away from, or toward the face area. In the other two image categories, contrast was reduced away from one of four candidate points.

ROC analysis was used assess whether contrast and edge density around fixated image patches could be discriminated against the same features at control locations. In agreement to several previous studies (e.g., Reinagel & Zador, 1999; Parkhurst & Niebur, 2003), and despite the contrast modifications introduced in the images, both contrast and edge density were found to be elevated at fixated positions. Moreover, another previously reported observation (Parkhurst et al., 2002), that the correlation between features and fixation locations are highest early after image onset, was also found in our analysis, although this tendency was weak. Interestingly, these overall results showed vary across the image categories.

To quantify how contrast manipulations affect where people look, each image was presented in two versions to the subjects; each version had its contrast reduced at different image regions. The 2-D correlation coefficient between the fixation densities of people looking at the two versions was computed to measure the magnitude of change in fixation locations. Fixation locations in images from the 'face' category, with the lowest SID, were least affected by contrast manipulations, whereas fixation locations in images from the 'man-made' and 'neutral' categories where increasingly affected as a function of increasing SID. To investigate whether these observations across image category were reflected in terms of fixated image content, further ROC analyses were performed on contrast and edge density for each image category separately. ROC analyses revealed that the ability for contrast and edge density to discriminate fixated locations from control locations varies heavily on image category (and thus SID). Images from the 'neutral' category has the best discrimination with up to 70%, whereas the other two categories had lower discriminabilities, but still above change performance. The poorest discrimination was found when analyzing only those images where faces were reduced in contrast; in this case, contrast and edge density were not different between fixated and control locations. In summary, we found to systematic discrimination in contrast and edge density between fixated and control locations, and the degree to which these features influenced fixation selection showed to vary across image category.

Recently, it has been debated whether low-level factors more heavily influence where people fixate early after image onset compared to later in viewing. For example, Parkhurst et al. (2002), suggested that saliency (see Itti et al., 1998) contributes more to fixation selection during the first fixation and thereafter slowly decreases over successive fixations, however still above chance level. To address this issue, we compared the influence of contrast and edge density on the initial fixation location. Again, the results varied heavily over the image categories. The effect reported by Parkhurst et al. (2002) was found in the low-SID, 'neutral' category. However, it was not consistent over the image categories. In fact, the opposite effect was found when regions rated as semantically important, such as faces, were reduced in contrast; subjects' initial fixations instead landed on regions with low contrast and edge density. Overall, this argues against a causal link between low-level features and fixation locations, at least for images with low SID. For high-SID images with neutral semantic content it may be argued that, since objects with high-level semantic interest are largely absent, bottom-up features to a larger extent influence where people look. However, even though the correlation between bottom-up features and fixated locations is higher in this case, it cannot be ruled out that other high-level mechanisms still control fixation selection.

According to a bottom-up, saliency framework, reducing the low-level signal strength would yield an obligatory shift in fixation density toward regions with a retained, high signal strength. Even though this effect was present overall in our data, it was not stable over different image categories and different types of contrast manipulations. Rather, the results found in this paper suggest that regions with high semantic importance attract fixateions regardless of their saliency. However, even though our results do not support a

DOI 10.16910/jemr.2.2.2

ISSN 1995-8692

causal link between saliency and fixation locations, it has been shown that what is judged as interesting coincides with regions of high saliency in typical real-world scenes (Henderson et al., 2007; Elazary & Itti, 2008). This may in part explain previous results emphasizing bottom-up control of eye-movements.

An issue related to this study concerns how low and high spatial frequencies are related to fixated image content. We have in this paper analyzed fixated content at rather high spatial frequencies. For example, the filters we used were of size 3×3 pixels and operated on images of size 1024×768 pixels. Consequently, only image variations with high detail were extracted, whereas coarser variations were not captured by these filters. S. Mannan et al. (1995); S. K. Mannan, Ruddock, and Wooding (1996) investigated how low pass filtering of an image affects where people look. They found that during the first 1.5 seconds of viewing, people fixate the same locations in the original image as in the low pass filtered version of this image. Since only the low frequency content is shared between these versions, this suggests that a representation based on low spatial frequencies could be responsible to guide fixations. In this sense, a saliency map operating on lower spatial frequencies could account for the results found in this paper. This line of argumentation has some support considering images from the 'face' category only; contrast manipulations dominantly attenuating higher frequencies have little influence on where people look, and faces are looked at regardless of their contrast levels. However, it seems more plausible that face regions are looked at because of their known semantic importance than because of some low-level account. Moreover, images from the 'neutral' category directly overthrow this assumption since fixation locations showed to be directly affected by contrast manipulations in this case.

Acknowledgments

Benjamin Tatler and the two anonymous reviewers are gratefully acknowledged for their comments on an earlier version of this manuscript. So are the personnel at Lund Humanities Lab.

References

Baddely, 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.

Buswell, G. T. (1935). *How people look at pictures*. University of Chicago Press, Chicago.

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

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), 1-19.

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

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

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 movements: A window on mind and brain* (p. 537-562). Oxford: Elsevier.

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (p. 1-58). New York: Psychology Press.

Itti, L. (2006). Quantitative modeling of perceptual salience at human eye position. *Visual Cognition*, *14*(4-8), 959-984.

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

Itti, L., Koch, K., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *20*(11), 1254-1259.

Mannan, S., Ruddock, K., & Wooding, D. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-d images. *Spatial Vision*, *9*(3), 363-386.

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.

Nyström, M., & Holmqvist, K. (2007). Variable resolution images and their effects on eye-movements. In B. Rogowitz, T. Pappas, & S. Daly (Eds.), *Human vision and electronic imaging.* San Jose, CA: SPIE.

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. (2003). Scene content selected by active vision. *Spatial Vision*, *16*(2), 125-154.

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

Privitera, C. M., & Stark, L. W. (2000). Algorithms for defining visual regions of interest: Comparison with eye fixations. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *22*(9), 970-982.

Rajashekar, U., Linde, I. van der, Bovik, A. C., & Cormack, L. K. (2007). Foveated analysis of image features at fixations. *Vision Research*, *47*(25), 3160-3172.

Rajashekar, U., Linde, I. van der, Bovik, A. C., & Cormack, L. K. (2008). Gaffe: A gaze-attentive fixation finding engine. *IEEE Trans. Image Processing*, *17*(4), 564-573.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Computation in Neural Systems*, *10*(4), 341-350.

DOI 10.16910/jemr.2.2.2

ISSN 1995-8692

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

Sarter, M., Givens, B., & Bruno, J. P. (2001). The cognitive neuroscience of sustained attention: where top-down meets bottom-up. *Brain Research Reviews*, *35*, 146 - 160.

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), 1-17.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, *45*(5), 643 - 659.

Tatler, B. W., Baddeley, R. J., & Vincent, B. T. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, *46*(12), 1857-1862.

Tatler, B. W., & Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, *2(2):5*, 1-18.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97-136.

Underwood, G., Foulsham, T., Loon, E. van, Humphreys, L., & Bloyce, J. (2006). Eye movements during scene inspection: A test of the saliency map hypothesis. *European J. Cognitive Psychology*, *18*(3), 321-342.

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

DOI 10.16910/jemr.2.2.2

ISSN 1995-8692