ROTATIONAL INVARIANCE AT FIXATION POINTS
Experiments using Human Gaze Data

Johannes Steffen, Christian Hentschel, Afra'a Ahmad Alyosef,
Klaus Toennies and Andreas Nuernberger
Otto-von-Guericke University Magdeburg, Magdeburg, Germany

Keywords: Eye-Tracking, Early Vision.

Abstract: An important aspect in machine vision concerns the extraction of meaningful patterns at salient image regions. Invariance w.r.t. affine transformations has usually been claimed to be a crucial attribute of these regions. While continuing research on the human visual cortex has suggested the correctness of these assumptions at least in later stages of vision, only lately the availability of accurate and cheap eye tracking devices has offered the possibility to provide empirical evidence to these claims. We present an experimental setting that is qualified to analyse various assumptions on human gaze target properties. The proposed setting aims at reducing high-level influence on the fixation process as much as possible. As a proof of concept we present results for the assumption human fixation targeting is rotational invariant. Even though high-level aspects could not be completely suppressed, we were able to detect and analyse this relation in the gaze data. It was found that there is a significant correlation between fixated regions within stimuli over different orientations.

1 INTRODUCTION

By analysing the visual field at different resolutions and sampling points the human visual system (HVS) is able to rapidly perceive and understand a complex visual stimulus. This ability is yet unmatched by any technical approach to this problem. Being able to understand the methodologies the HVS employs should help to improve existing approaches to machine driven image analysis and understanding. While it is still not fully known, which information is processed at which level of resolution various studies have given evidence that certain tasks such as reading and visual search require a higher resolution than others such as simple scene classification. The machine vision community has mainly focused on the extraction of local features at salient regions in images. Algorithms that aim at the detection of these regions usually strive to maximize invariance (e.g. w.r.t. affine transformation) in order to provide detectors robust to varying image acquisition conditions. Whether or not invariance is equally important in human gaze targeting is the focus of this paper.

We conceived of an experimental setting that is qualified to analyse the impact of invariance at human fixation point selection by recording gaze data for stimuli carefully selected to reduce high-level influence on the recognition process as much as possible. While this setting is intended for later use in more complex scenarios we show its general validity by analysing whether the specific example of *rotational* invariance is a coherent property of early human gaze targeting.

In the following sections we will first give an overview of the related literature in this field and provide a motivation for our own work. We then present the experimental setting and the stimuli data used during our tests. We will further present a questioning scheme we developed to reduce high level influences during the experiment. Finally, we provide the obtained results and give some evidence on the extendibility of the presented experimental approach to broader and more complex scenarios.

2 RELATED WORK

The selection of fixation target by the Human Visual System (HVS) has been subject of research for years and two major factors for eye movements have been identified (Henderson, 2003). Bottom-up factors are stimulus intrinsic attractors for fixation targets mean-

Steffen J., Hentschel C., Ahmad Alyosef A., Toennies K. and Nuernberger A. (2012).
ROTATIONAL INVARIANCE AT FIXATION POINTS - Experiments using Human Gaze Data.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 451-456
DOI: 10.5220/0003782104510456

ing that the selection of these targets is solely driven by neuronal analysis of the visual features within the stimulus. On the other hand, top down factors introduced by the observer's "internal state" and imposed by a search task or prior knowledge about the stimulus likewise affect the selection of fixated regions (Hopfinger et al., 2000; Corbetta et al., 2000). While the latter factors are more challenging to model, models that predict visual saliency in compliance with human bottom-up factors have been subject of research as they are easier understood.

Early psychophysical experiments with human observers (Treisman and Gelade, 1980; Bergen and Julesz, 1983) have proven that a limited set of bottom-up features (among these color, orientation of line segments, certain shape parameters such as curvature) are detected at an early stage of human vision. Later in the visual process, by combination of these simple features, more complex neurons respond to higher-level features such as corners and junctions (Hubel and Wiesel, 1965; Pasupathy and Connor, 1999), edges (Marr and Hildreth, 1980), curved segments (Dobbins et al., 1989) and key points (Heitger et al., 1992; Rodrigues and du Buf, 2006). Being biologically plausible, the empirical support for the importance of these features in human vision was given in (Biederman, 1987) where it is shown that human recognition performance is decreasing largely when the corners of an object are removed. In (Koch and Ullman, 1985), the first explicit model that fuses the response of several early visual features (intensity, color, orientation and temporal change) into a single saliency map was described. A computational implementation of this model was derived later (Niebur and Koch, ; Itti et al., 1998). While not necessarily primarily with the aim to model properties of early human vision, the computer vision community likewise has developed a vast corpus of local feature detection algorithms, designed to detect edges, blobs, key points etc. as basic elements for machine vision. An extensive overview can be found in (Mikolajczyk et al., 2005) and (Tuytelaars and Mikolajczyk, 2007).

The availability of inexpensive and rather accurate eye trackers that record the direction of gaze of a human observer while regarding a visual stimulus (typically an image on a computer screen) recently led to approaches that analyse the neighborhood of fixation targets and try to model saliency as targets of overt attention. The results of the conducted experiments give support to the biologically inspired models of saliency. Edge density was reported to be higher at fixation points (Mannan et al., 1996) and in (Krieger et al., 2000) two-dimensional image features like curved lines and edges, occlusions, isolated spots, etc. have been identified as important fixation candidates. Later studies of image statistics at human gaze data (Parkhurst and Niebur, 2003; Rajashekar et al., 2007) obtain similar results. Finally (Parkhurst et al., 2002) describes a significant correlation between computed visual saliency (by advancing from the aforementioned biologically plausible models) and human eye movement data.

In a previous work (Alyosef, 2011), we analysed the consistency of five of these computational models for local feature selection with the fixation targets of human observers solving a retrieval task. The intention was to investigate to what extent these detectors are qualified to predict bottom-up as well as top-down gaze targets. While our results (the key point localization step of the SIFT algorithm performed best) correspond to the findings in (Harding and Robertson, 2009) and (Rajashekar et al., 2007) we identified a strong interplay of top-down cues (i.e. the retrieval task was too simple w.r.t. the presented stimuli).

We therefore decided to restrict ourselves to analyse solely bottom-up properties of human gaze targeting. Invariance w.r.t. varying image acquisition conditions has been considered an important property of most of the aforementioned algorithms in machine vision. Similarly, transformation invariances have been considered to be important by neurally motivated models of vision (e.g. see (Wallis et al., 1993; Deco and Rolls, 2004)) . Whether or not invariance is important in human gaze targeting is the focus of this paper. We conceived of an experimental setting that is qualified to analyse the impact of invariance at human fixation points. While this setting is intended for later use in more complex scenarios we prove its general validity by analysing the human gaze data on stimuli that are qualified to prove or dismiss the theory that selection of fixation targets of the HVS is *rotational* invariant.

3 GAZE DATA ACQUISITION

3.1 Experimental Setup

For obtaining our sampling data we used the T60 eye tracker from Tobii Technology[1]. The T60 is a table-mounted video-based eye tracker that uses an infrared camera system to record gaze data. The cameras are built into a 17" TFT screen with a native resolution of $1280x1024$ pixels that is used to present the stimuli data. Being non-intrusive, the tracker offers freedom of head movements (within an eye tracking

[1]Tobii Technology: http://www.tobii.com/

box of W:44cm x H:22cm x D:30cm at 70cm from the eye tracker) which allows the participants to behave as naturally as in front of any other computer screen. No additional chin or forehead rest was used. The tracker collects raw eye movement data points every 16.6 ms (i.e. the sampling data rate is 60Hz) and provides a tracking accuracy of 0.5°. Drift effects (caused e.g. by varying pupil size due to varying screen illumination levels) are reduced to $< 0.3^\circ$.

We set up the T60 on a blank desk and removed all possibly irritating objects behind it. We then mounted a fixed chair that we adjusted for each participant assuring an eye-screen-distance of about 70 cm. To avoid different illumination on the scene and assuring consistent ambient illumination we closed all curtains and used artificial light sources leading to low light conditions instead.

Stimuli presentation, tracker calibration and gaze data acquisition was done using the freely available OGAMA (OpenGazeAndMouseAnalyzer[2]) software.

3.2 Stimuli Description

We rendered 10 black polygons on white background with 3 to 8 edges. To make fixation points at polygons' corners clearly distinguishable polygons were selected to exhibit visually separable vertices (w.r.t. distance) and a minimal and maximal opening angle between two edges. Polygons were scaled to fit most of the trackers' screen, yielding to an average visual angle of approx. 18°. The polygons are depicted in Fig. 2. Each polygon was rotated around the center of the screen. Rotation angles of multiples of 60° were used. Thus, we obtained 6 different projections of the same polygon – the original and 5 rotated versions.

An additional set of 20 polygons was rendered in the same way to be used as settling data and was presented to the participants before the actual experiment started. In order not to bias the experiment, this was not communicated to the participants. The last set of 16 randomly rendered polygons was used to distract the participant from the fact that they are observing 10 unique polygons that are just rotated differently and were presented in random order in the experiment data after the settling set. That leads to 96 images.

The motivation for this step was to avoid that a participant could make up any assumption on what to look for within a presented polygon and would therefore be biased while examining succeeding images. This set was selected as to convince the participants of the random structure of the presented stimuli and to avoid that a participant would look for any regularities or any repetitive character of the presented images throughout the rest of the experiment. Any such assumption of a participant would represent a top-down bias for the process of gaze targeting. The participants were unaware of this step and were left under the impression of examining a total of 96 images.

[2]OGAMA: http://www.ogama.net/

3.3 Task Description

As reducing the impact of top-down factors on the participant's fixated regions was crucial throughout the design of our experimental setting we conceived of a generic viewing task that was intended to help achieving this goal. To do so we created a set of 30 questions and the participants were asked to answer one of them *after* each presented stimulus.

The questions were designed to be very simple assuring a) that the cognitive load of the participants stays at a constant level and that they are not too distracted understanding the question and b) that the questions can be answered in a quick manner to reduce the experiment duration. In addition to that we shuffled the questions randomly to make them as unpredictable as possible and thus eliminated bias that could have raised by knowing which parts of the stimulus is or could be important for answering the consecutive question on it. The correct response to these questions was neither given to the participants nor was it of any importance for the outcome of the experiment. As mentioned before the overall aim of the questions was rather to ensure a level of overt attention spatially equally spread over the entire stimulus. Here are some examples of the mentioned questions:

- "Have there been more than 4 edges?"
- "Have you seen circular edges?"
- "Have you seen exactly one object?"

3.4 Participants and Process

Fixations have been recorded for 15 participants who examined each of the $20 + 76$ polygons described in 3.2. After the calibration process, the participants were instructed to view each presented image carefully in order to be able to answer the subsequently posed questions. All 96 images were presented in a slide-show-like character to each of the participants.

Each image was presented for a duration of 5 seconds. After each presentation a random question was selected from the catalog (see Section 3.3). The participant was asked to answer the question. While the correct answer did not matter for the experiment itself, a new image was not shown before the subject gave any answer at all. To avoid any distraction of the participants by asking orally we decided to present

the question as a part of the slide show directly after a stimulus. Thus, the participant had to answer the question by remembering the presented stimulus without seeing it, which otherwise would have meant a high-level top-down influence in the fixation selection process (e.g. a task specific search on the image). Additionally, presenting a neutral gray image before each new stimulus avoided that different stimuli images interfered during the recording process. Once the question was answered orally the next image was presented.

3.5 Gaze Data Analysis

All gaze movements of a participant were recorded using OGAMA. As a post processing step participants' fixations were filtered containing information about the fixation duration, start time, end time, and position for each trial within the experiment. Parameters for calculating a fixation out of all gaze point were set fix for all participants and every trial.

Because eye gaze fixation are not steady we considered a gaze point as a part of a real fixation if the maximum distance between its coordinates and the coordinates of the average fixation point did not exceed 20 pixels in total. Furthermore, the minimal number of gaze point samples to be considered as a fixation was set to 5. Finally, all calculated fixations that are within the maximum distance of any neighboring fixation point were merged to be just one fixation point. To avoid having overlapping fixations when switching to a consecutive trial (e.g. when we switched from the neutral grey background image for questioning to the next trial) the first fixation point was discarded if it was at the same place as the last fixation point of the foregoing trial with a duration less than 200 ms.

For analysing the gaze data, saliency maps were generated using an aggregated Gaussian distribution (Vosskuehler, 2009):

$$f(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (1)$$

where $x,y \in [-s,s]$ and $\sigma = s/5$. All multiplied kernels per trial are then added to a new image having the same size as the given stimulus (1280x1024 pixels) and are finally normalized. We set the Gaussian kernel size to $\sigma = 40 pixels$.

In the next step the average saliency map for each of the 60 stimuli over all 15 participants was generated. For better comparability the saliency maps were rotated back to the original orientation of 0°. To give an impression on the resulting data, figure (2) shows the stimuli polygons as well as their corresponding 6 saliency maps colored using a rainbow gradient.

To measure the similarity between two saliency maps A and B the two-dimensional correlation coefficient p is calculated using the Pearson's linear correlation as described in (Engelke et al., 2010):

$$p = \frac{\sum_m \sum_n (A_{mn} - \overline{A_{mn}}) - (B_{mn} - \overline{B_{mn}})}{\sqrt{\sum_m \sum_n (A_{mn} - \overline{A_{mn}})^2} - \sqrt{\sum_m \sum_n (B_{mn} - \overline{B_{mn}})^2}} \quad (2)$$

where $m \in [1,M]$, $n \in [1,N]$ are the pixel coordinates, and $\overline{A_{mn}}$, $\overline{B_{mn}}$ denoting the mean pixel value. The larger the value of the correlation coefficient p gets, the higher is the correlation between two saliency maps, with $p \in [-1,1]$.

4 EXPERIMENTAL RESULTS

We computed the correlation p between the saliency maps of any two rotated projections of each of the 10 polygons $A_x, B_x, ..., J_x$ (where $x \in [0°, 60°, 120°, ..., 300°]$ denotes the rotation angle). Figure 1 shows the aggregated correlation coefficient for each polygon over all orientations.

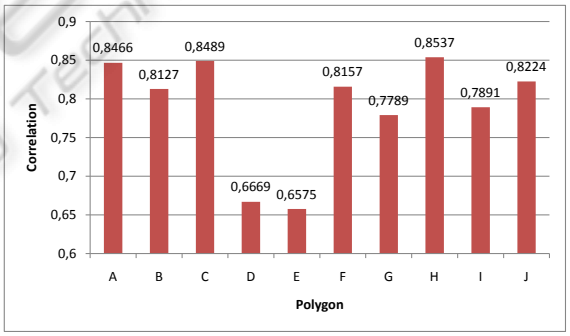

Figure 1: Mean correlation coefficient for all polygons over all orientations and participants

Assuming the participants treat the rotated projections of each polygon as unseen stimuli, the correlation has to be considered as rather high. The smallest correlation $p = 0.5132$ was found comparing the saliency maps of polygon E for the rotation angles $x = 180°$ and $x = 300°$ whereas the highest value $p = 0.9211$ was obtained comparing the rotated projections of polygon H with $x = 240°$ and $x = 300°$. Furthermore, considering the immediate "neighbor" of each projection (i.e. $x \pm 60°$) the correlation was found highest for 83% of all neighboring projections. Considering the mean correlation coefficients (see Fig. 1) the smallest correlation can be

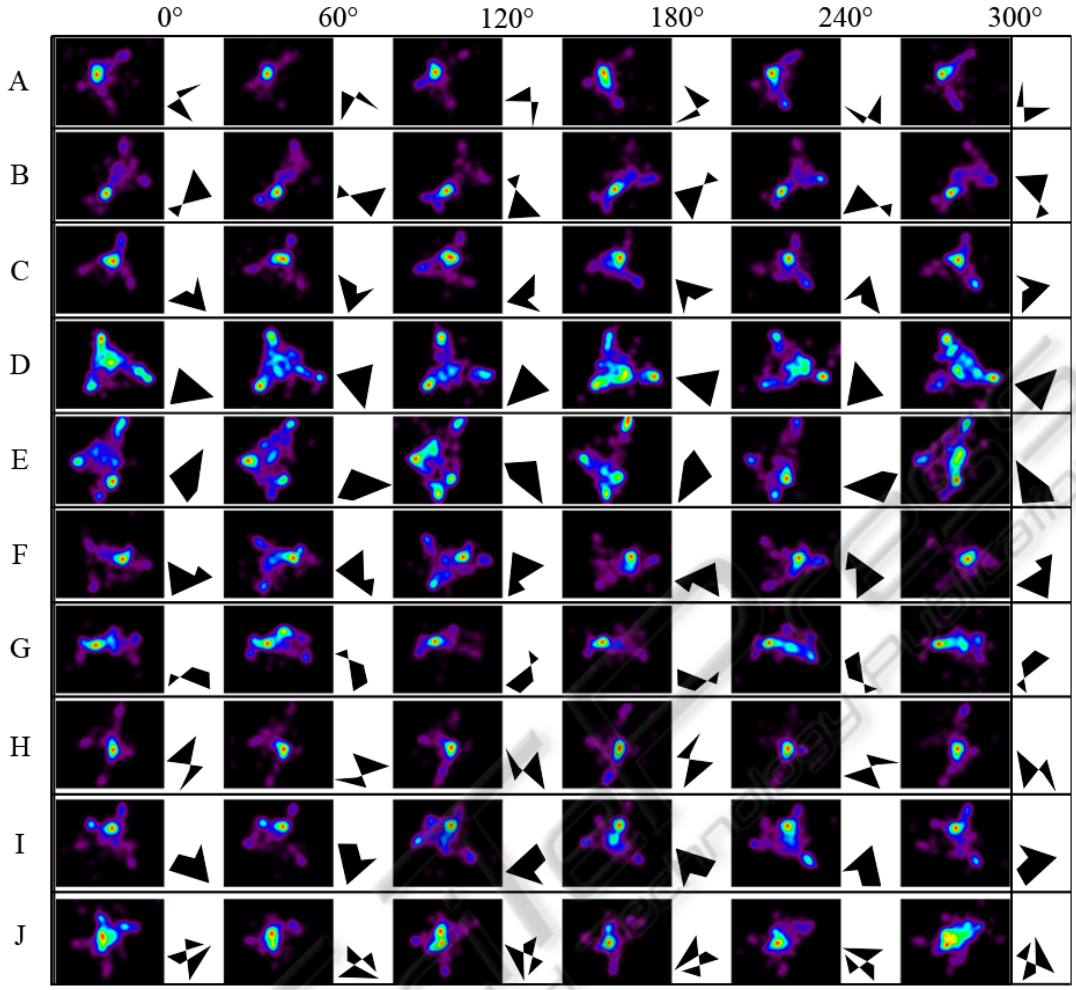

Figure 2: Overview of all 10 polygons and their corresponding colored saliency maps (rotated back to their initial position of 0° degree for better visually comparison)

found for polygon D and E. As these were the simplest polygons, we assume that most of their properties could be perceived without overt attention and details were not fixated directly. When sorting the correlation coefficients p with increasing degree of rotation angle the absolute values of p were found to be increasing too. We assume this is due to the fact that there is at least one factor influencing the fixation selection process of the participants, which is yet unrepresented in the experimental setting. Considering that the segregation or dichotomy of the ventral "what" and dorsal "where" stream (two-stream hypothesis) is highly in doubt (Farivar, 2009), we can not assure that our results only represent outcomes of the low-level "where" stream. Due to the heavily interconnection of both streams, we have to act on the assumption that fixation positions are always a result of both processing streams.

Another outcome worth being mentioned is that the highest density of fixation points over all participants was found at regions with high local complexity. Thus, interesting characteristics like local convexity, line-crossings, and junctions trigger the attention of every observer resulting in very high peaks at the corresponding saliency maps. As can be seen in Fig. 2 whenever there is a concave region or a region showing a high local complexity participants tend to fixate this region more often than outer corners and edges (e.g. polygon D compared with H).

4.1 Conclusions

As a summary, evidence for rotation invariance at fixation points could be given on a certain level. Assuming that the presented design avoids most of the high level influences built on experience held before or gained throughout the experimental process, the analysed data indicates that rotational invariance exists

considering the given simple stimuli. The strong correlation of neighboring orientations raise the assumption that it was not possible to suppress all top-down influences as initially intended. Moreover, given the fact that there is a significant difference comparing fixation data of simple convex polygons with other more complex concave versions further investigation on convexity and concavity should be performed. More experiments with different polygon prototypes, which should be further reduced regarding their complexity (i.e. the number of vertices and nodes) should be carried out. Another important fact, which is currently not represented by the proposed experimental setting is that humans usually do not perceive objects and scenes in discriminative steps as we simulated.

REFERENCES

Alyosef, A. A. (2011). Comparison of interest points of computer vision detectors with human fixation data. Master's thesis, University of Magdeburg, Germany.

Bergen, J. R. and Julesz, B. (1983). Parallel versus serial processing in rapid pattern discrimination. *Nature*, 303:696–698.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147.

Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., and Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature neuroscience*, 3(3):292–7.

Deco, G. and Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision research*, 44(6):621–42.

Dobbins, A., Zucker, S. W., and Cynader, M. S. (1989). Endstopping and curvature. *Vision Research*, 29(10):1371–1387.

Engelke, U., Liu, H., Zepernick, H.-J., Heynderickx, I., and Maeder, A. (2010). Comparing two eye-tracking databases: The effect of experimental setup and image presentation time on the creation of saliency maps. *International Picture Coding Symposium*.

Farivar, R. (2009). Dorsalventral integration in object recognition. *Brain Research Reviews*, 61(2):144 – 153.

Harding, P. and Robertson, N. (2009). A comparison of feature detectors with passive and task-based visual saliency. *LNCS*, 5575:716–725.

Heitger, F., Rosenthaler, L., von der Heydt, R., Peterhans, E., and Kübler, O. (1992). Simulation of neural contour mechanisms: from simple to end-stopped cells. *Vision Research*, 32(5):963–981.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Neuroscience*, 7(11):498–504.

Hopfinger, J. B., Buonocore, M. H., and Mangun, G. R. (2000). The neural mechanisms of top-down attentional control. *Nature neuroscience*, 3(3):284–91.

Hubel, D. and Wiesel, T. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, 28.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259.

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

Krieger, G., Rentschler, I., Hauske, G., Schill, K., and Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial vision*, 13(2-3):201–14.

Mannan, S., Ruddock, K., and Wooding, D. (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.

Marr, D. and Hildreth, E. (1980). Theory of Edge Detection. *Proceedings of the Royal Society B: Biological Sciences*, 207(1167):187–217.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., and Gool, L. V. (2005). A comparison of affine region detectors. *International Journal of Computer Vision*, 65:43–72.

Niebur, E. and Koch, C. Control of selective visual attention: modeling the" where" pathway. *Advances in neural information processing systems*, pages 802–808.

Parkhurst, D., Law, K., and Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–23.

Parkhurst, D. J. and Niebur, E. (2003). Scene content selected by active vision. *Spatial vision*, 16(2):125–54.

Pasupathy, a. and Connor, C. E. (1999). Responses to contour features in macaque area V4. *Journal of neurophysiology*, 82(5):2490–502.

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

Rodrigues, J. and du Buf, J. (2006). Multi-scale keypoints in v1 and beyond: object segregation, scale selection, saliency maps and face detection. *BioSystems*, 86.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136.

Tuytelaars, T. and Mikolajczyk, K. (2007). Local Invariant Feature Detectors: A Survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280.

Vosskuehler, A. (2009). Ogama description (version 2.5).

Wallis, G., Rolls, E., and Foldiak, P. (1993). Learning invariant responses to the natural transformations of objects. *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*, 2:1087–1090.