Cogn Comput (2011) 3:94–104
DOI 10.1007/s12559-010-9087-7

An Image Statistics–Based Model for Fixation Prediction

Victoria Yanulevskaya · Jan Bernard Marsman · Frans Cornelissen · Jan-Mark Geusebroek

Received: 29 March 2010 / Accepted: 26 November 2010 / Published online: 14 December 2010

Abstract The problem of predicting where people look at, or equivalently salient region detection, has been related to the statistics of several types of low-level image features. Among these features, contrast and edge information seem to have the highest correlation with the fixation locations. The contrast distribution of natural images can be adequately characterized using a two-parameter Weibull distribution. This distribution catches the structure of local contrast and edge frequency in a highly meaningful way. We exploit these observations and investigate whether the parameters of the Weibull distribution constitute a simple model for predicting where people fixate when viewing natural images. Using a set of images with associated eye movements, we assess the joint distribution of the Weibull parameters at fixated and non-fixated regions. Then, we build a simple classifier based on the log-likelihood ratio between these two joint distributions. Our results show that as few as two values per image region are already enough to achieve a performance comparable with the state-of-the-art in bottom-up saliency prediction.

Keywords Natural image statistics · Visual saliency · Weibull distribution

V. Yanulevskaya (✉) · J.-M. Geusebroek
Intelligent Systems Lab Amsterdam, Informatics Institute
University of Amsterdam, Postbus 94323, 1090, GH,
Amsterdam, The Netherlands
e-mail: V.Yanulevskaya@uva.nl

J. B. Marsman · F. Cornelissen
Laboratory for Experimental Ophthalmology,
School of Behavioural and Cognitive Neurosciences,
University Medical Center Groningen, PO Box 30.001,
9700, RB, Groningen, The Netherlands

Introduction

While observing the world around us, we constantly shift our gaze from point to point to visually sample our surrounding. These shifts are not random but are driven by visual stimuli, like simple variations in contrast or colour [1, 19, 26, 30], or the presence of faces [5]. The visual projection of the world on our eye is not random either, but highly organized and structured. The latter is reflected in the spatial statistics of the perceived scene, whose regularities are captured by the statistical laws of natural images [11]. Therefore, one would expect eye-fixations to be closely connected with the laws of natural image statistics. In this work, we study in how far a direct connection can be established between image statistics and locations of eye-fixations.

Low-level visual features are the basis from which many saliency indicators have been derived. Itti et al. [19], followed by others [15, 22, 31], construct a biologically inspired saliency map by considering colour, contrast, and orientation features at various scales. The model combines a total of 42 feature maps into a single saliency map, resulting in the labelling of regions that deviate from the average for these features. Their influential approach has set a standard in saliency prediction. However, it is unclear how much these 42 features contribute to the fixation prediction and whether it is necessary to consider all of them.

Reinagel and Zador [30] take the fixation locations as a starting point for analysis. They consider the difference between the image statistics of fixated and non-fixated image locations. The issue here is how to choose plausible image features from which to derive eye movements. A number of image regularities have been considered, see [1] for an overview. Most researchers [29, 30, 38] confirm

that contrast and edges yield significant difference between their statistics of fixated and non-fixated locations.

In the field of natural image statistics, Geusebroek and Smeulders [14] have shown the two-parameter Weibull distribution to describe the local contrast statistics adequately. They show that both contrast and edge frequency are simultaneously captured by the Weibull distribution, conjecturing that its parameters might be relevant in fixation prediction. Scholte et al. [34] examined to which degree the brain is sensitive to these parameters and found a correlation of 84 and 93%, respectively, between the two Weibull parameters and a simple model of the parvo- and magnocellular system. Given these results, one would expect image contrasts around fixation locations to reflect these Weibull statistics.

The central issue addressed in this paper is the following: *Do the parameters of the Weibull distribution predict locations of eye-fixations?* If so, the Weibull distribution can be used as, or might even be ground for, a simple predictor of fixation locations.

Our approach elaborates on the work of Zhang et al. [41]. They infer bottom-up saliency from the information gain between the local contrast in a given image when compared against the average statistics over a larger image collection, as parameterized by a Generalized Gaussian distribution—a "cousin" of the Weibull family [14]. Our approach aims at learning the parameters of local statistics as parameterized by the Weibull distribution at fixated and non-fixated locations. As such, saliency is expressed by the likelihood of the parameters of the distribution to occur in scenes, the parameters being tuned to the statistics of local scene content. We show that, using as few as two parameters of such a simple Weibull model, we obtain prediction of fixation locations comparable with the state-of-the-art in bottom-up saliency [4].

Methods

We treat eye-fixation prediction as a two-class classification problem. The salient class consists of fovea-sized (1°, which is 30 pixels in our experiments) regions around fixated locations, and the rest of the image is considered as the non-salient class. Our approach is based on the assessment of local image statistics which are learned for salient and non-salient classes. Particularly, we model the distribution of the regional colour gradient magnitude responses with the Weibull distribution as discussed below. The classification decision is based on the log-likelihood ratio with null hypothesis that the Weibull parameters describe the salient region, and alternative hypothesis that the Weibull parameters describe the non-salient region. The proposed method is summarized in Fig. 1.

To determine the non-fixated locations for an image, we follow [1] and randomly select the fixated locations from different images, which are at least 1°, i.e. fovea size, apart from the fixations on the current image. As a result, we have the same number of fixated and non-fixated regions per image. This way of selecting non-fixated locations ensures similar distributions of fixated and non-fixated regions [1].

Feature Extraction

In our approach, we model local colour contrast statistics with the Weibull distribution. After that, we estimate the joint distribution of the Weibull parameters at the fixated and non-fixated regions.

Colour Contrast

Colour contrast of an image is determined by the gradient magnitude, calculated using Gaussian derivative filters,

$$\begin{aligned} G_x(x,y,\sigma) &= \frac{-x}{2\pi\sigma^4}\exp^{\frac{-(x^2+y^2)}{2\sigma^2}}, \\ G_y(x,y,\sigma) &= \frac{-y}{2\pi\sigma^4}\exp^{\frac{-(x^2+y^2)}{2\sigma^2}}. \end{aligned} \tag{1}$$

We follow [13] and convert RGB values to an opponent colour space with decorrelated colour channels,

$$\begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.3 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{pmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \tag{2}$$

The weights in this conversion are optimized for the human visual system [12]. Colour gradient magnitude is obtained by,

$$\|\nabla E(x,y,\sigma)\| = \sqrt{E_{1x}^2 + E_{1y}^2 + E_{2x}^2 + E_{2y}^2 + E_{3x}^2 + E_{3y}^2}. \tag{3}$$

where E_{jx}, E_{jy}, $j = 1,2,3$ are the Gaussian derivative filter responses of the corresponding decorrelated opponent colour channels in the x and y direction, and $\|\nabla E(x,y,\sigma)\|$ is the resulting colour gradient magnitude of an image. Besides estimating the local intensity edges, this operator also emphasizes chromatic contrasts. To estimate the distribution of the colour gradient magnitude, we construct a weighted local histogram of colour gradient magnitude responses within an image region, where weights are determined by a Gaussian windowing function ($\sigma = 1°$, i.e. 30 pixels) located at the centre of the region. Hence, pixels close to the centre location will contribute more to the histogram than pixels further away, effectively localizing measurements at the centre of the image region which for fixated region is the fixation location itself.

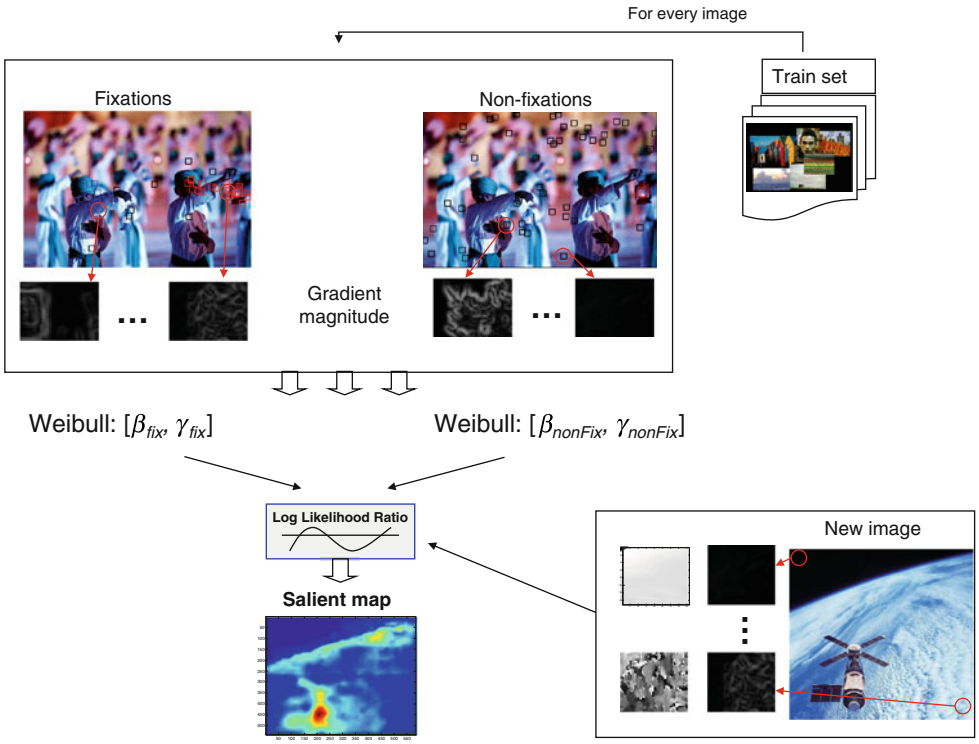

Fig. 1 Experimental setup. We use statistical machine learning techniques to learn to differentiate between fixated and non-fixated locations. *Training phase*: We extract fovea-sized regions around fixated and non-fixated locations from the images which belong to the training set. We represent the local edge distribution as a histogram of colour gradient magnitude responses of the fixated/non-fixated regions. We parameterize these histograms of the colour gradient magnitude responses with the Weibull distribution. The two parameters of the Weibull distribution β and γ are determined by the maximum likelihood estimation. Hence, we obtain two sets of the β and γ parameters: one at fixated patches $[\beta_{fix}, \gamma_{fix}]$ and another at non-fixated patches $[\beta_{nonFix}, \gamma_{nonFix}]$. We estimate class-conditional probability density functions $P(\beta, \gamma|\, fix)$, and $P(\beta, \gamma|\, nonFix)$ and calculate the log-likelihood ratio between these two probability density functions. *Construction of a saliency map for a new image*: we parameterize all fovea-sized regions extracted on a dense regular grid within the image in the same way as above. Then, we classify each region based on the value of the log-likelihood ratio of its Weibull parameters

Scale Selection

For the colour gradient operator from Eq. 3, the parameter σ has to be determined, indicating at which scale edges are detected. Here, we follow [9] and use the minimal reliable scale selection principle. The minimal reliable scale depends on the sensor noise characteristics and the local intensity contrasts. For high contrast, signal to noise ratio will be locally high, so a small scale is sufficient to detect an edge. For low contrast, a large scale is required to distinguish the signal from the noise. Doing so, the method selects the optimal scale for edge detection at each pixel. Specifically, the method of [9] assesses the likelihood that the gradient magnitude of intensity is being caused by noise. The likelihood diminishes when the Gaussian derivative scale σ increases for the gradient operator. The smallest scale at which the gradient magnitude is more likely (significance level $\alpha = 0.05$) to be generated by a true edge rather than sensor noise is considered the minimal reliable scale. We have extended the method to colour gradients introduced previously. We assume noise independence per colour channel, and model the effect of sensor noise on the nonlinear colour gradient response Eq. 3. In our experiments, we assume Gaussian sensor noise with a standard deviation of 5% of the dynamic range of the intensity. Furthermore, we logarithmically sample the scales using the same intervals as in the successful SIFT descriptor [24]. In total, we consider the following 15 scales: 1.519, 1.952, 2.490, 3.160, 4.000, 5.055, 6.380, 8.047, 10.147, 12.790, 16.119, 20.312, 25.595, 32.250, 40.634.

Weibull Statistics

To parameterize the colour gradient magnitude responses, we use the Weibull distribution. The Weibull distribution as a parameterized model provides a good fit to the edge distribution of natural images at a local and global scale [14]. The probability density function (pdf) of the Weibull distribution is given by,

$$P(x) = \frac{\gamma}{\beta}\left(\frac{x}{\beta}\right)^{\gamma-1} \exp^{-\left(\frac{x}{\beta}\right)^{\gamma}}, \quad (4)$$

where $x > 0$ is the value of the gradient magnitude, $\gamma > 0$ is the shape parameter, and $\beta > 0$ is the scale parameter of the distribution. These two parameters catch the structure of the image texture [14]. The scale β represents the width of the distribution and reflects the (local) contrast. The shape γ represents the slope of the distribution and is sensitive to the (local) edge frequency [14]. We determine the Weibull parameters by the maximum likelihood estimation method [20] resulting in the equations

$$\beta - \sqrt[\gamma]{\frac{\sum_{i=1}^{n} x_i^{\gamma}}{n}} = 0, \quad (5)$$

and

$$\begin{aligned} f(\gamma) = n &+ \sum_{i=1}^{n} \ln\left(\frac{x_i^{\gamma} n}{\sum_{i=1}^{n} x_i^{\gamma}}\right) \\ &- n \sum_{i=1}^{n} \frac{x_i^{\gamma}}{\sum_{i=1}^{n} x_i^{\gamma}} \ln\left(\frac{x_i^{\gamma} n}{\sum_{i=1}^{n} x_i^{\gamma}}\right) \\ &= 0. \end{aligned} \quad (6)$$

where n is the size of the observed data. As Eq. 6 is transcendental, we solve it numerically using the standard iterative Newton–Raphson method [2]:

$$\gamma_{k+1} = \gamma_k - \frac{f(\gamma_k)}{f'(\gamma_k)}, \quad (7)$$

where

$$f'(\gamma) = \sum_{i=1}^{n} \Lambda \frac{\sum_{i=1}^{n} x_i^{\gamma}}{x_i^{\gamma}} - n \sum_{i=1}^{n} \Lambda \ln\left(\frac{x_i^{\gamma} n}{\sum_{i=1}^{n} x_i^{\gamma}}\right) - n \sum \Lambda, \quad (8)$$

and

$$\Lambda = \left(\frac{x_i^{\gamma}}{\sum_{i=1}^{n} x_i^{\gamma}}\right)'_{\gamma} = x_i^{\gamma} \frac{\ln x_i \sum_{i=1}^{n} x_i^{\gamma} - \sum_{i=1}^{n} x_i^{\gamma} \ln x_i}{\left(\sum_{i=1}^{n} x_i^{\gamma}\right)^2}. \quad (9)$$

The pseudo code of the Newton–Raphson method is given in Algorithm 1.

The maximum likelihood estimator $\hat{\gamma}$ is the solution of Eq. 6. Consecutively, $\hat{\beta}$ can be calculated from Eq. 5.

In our experiments, we consider the joint distribution of the two Weibull parameters. It allows to combine contrast and edge frequency information together, taking into account the correlation between these two image features. Figure 2 shows the fitted Weibull pdf for a few examples.

Log-Likelihood Ratio–Based Classification

Given the Weibull parameters describing an image region, we want to decide whether or not this region is salient. We base our classifier on the log-likelihood ratio [32] of the probability of the Weibull parameters occurrence on the fixated region with respect to the probability of the Weibull parameters occurrence on the non-fixated region:

$$LLR(\beta, \gamma) = -2 \log \frac{P(\beta, \gamma | fix)}{P(\beta, \gamma | nonFix)}, \quad (10)$$

where $P(\beta, \gamma|fix)$ and $P(\beta, \gamma|nonFix)$ are class-conditional probability density functions of the Weibull parameters β and γ. These probability density functions are estimated using a two-dimensional histogram of the Weibull parameters occurrence on fixated (salient) and non-fixated (non-salient) regions. We estimate $P(\beta, \gamma|fix)$ and $P(\beta, \gamma|nonFix)$ using images from the training data set.

Algorithm 1 Newton–Raphson algorithm for γ estimation

$\gamma = 1$ *initial value*
$\varepsilon = 0.001$ *accuracy of the calculations*
$\gamma_{next} = \gamma - \frac{f(\gamma)}{f'(\gamma)}$
while $|\gamma_{next} - \gamma| > \varepsilon$
    $\gamma = \gamma_{next}$
    $\gamma_{next} = \gamma - \frac{f(\gamma)}{f'(\gamma)}$
return γ_{next}.

Saliency Map Calculation

To calculate the saliency map for a new input image for which we want to predict eye-fixation locations, we first parameterize all fovea-sized local regions extracted on a dense regular grid within the image. Then, we calculate a saliency score of each region according to Eq. 10. The decision whether the region is salient or not is determined by thresholding the likelihood ratio. If $LLR(\beta, \gamma)$ is above the threshold, the region is accepted as being salient, otherwise it is rejected:

$$\left\{ \begin{array}{l} \text{if } LLR(\beta, \gamma) \geq \tau, \text{patch is salient} \\ \text{if } LLR(\beta, \gamma) < \tau, \text{patch is non-salient.} \end{array} \right\} \quad (11)$$

Because regions are overlapping, we average the result over the pixels they cover. In our experiments, we investigate how our method performs across a set of thresholds.

Evaluation

It is important to investigate the peaks of the saliency map as it is expected that new observers will focus attention there. Therefore, to asses the performance of the proposed Weibull method, we follow [21] and report area under the adapted receiver operating characteristics (ROC) curve. The adapted ROC curve depicts the trade-off between hit rate and the percentage of salient area. Particularly, the hit

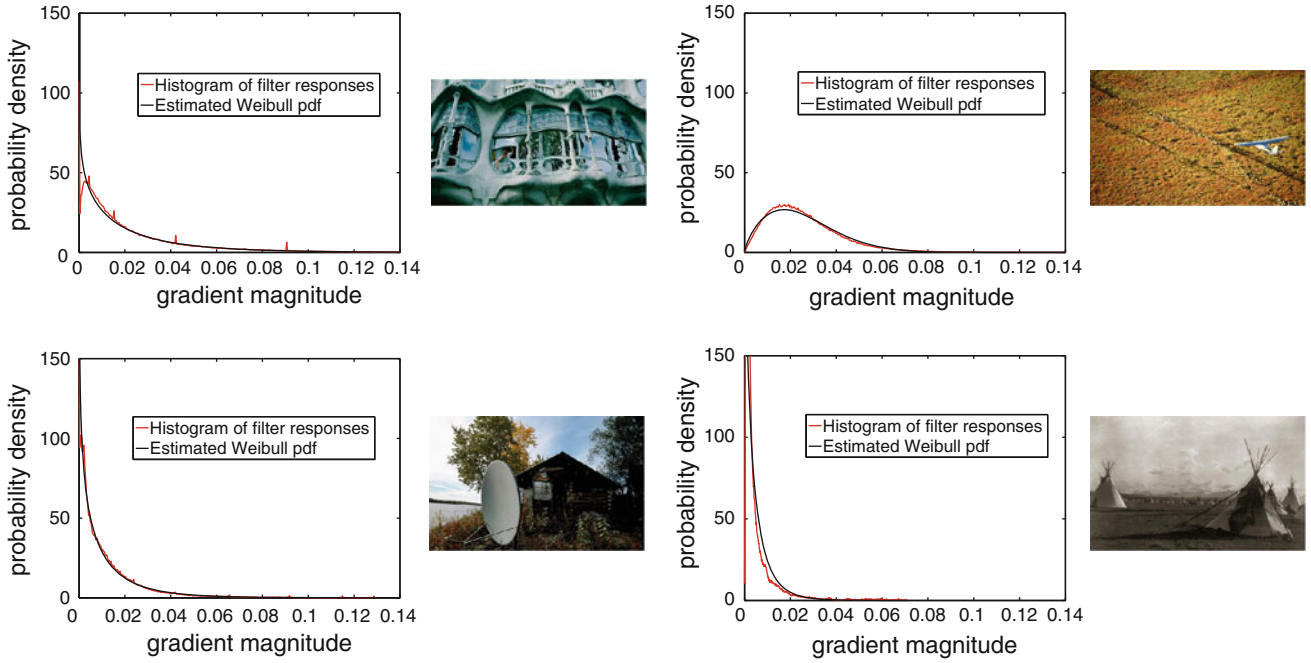

Fig. 2 Examples of the Weibull distribution: the upper row shows images with overall similar contrast variation ($\beta = 4.11$ and 4.12, respectively), but moderate and high edge frequency ($\gamma = 0.95$ and 1.59, respectively). The bottom row shows images with higher and lower contrast ($\beta = 2.01$ and 1.58, respectively), while exhibiting overall moderate edge frequency ($\gamma = 0.701$ and 0.702, respectively). Note that here we consider global image histograms for illustration purposes

rate is the ratio of ground truth fixated locations classified as fixated, and we threshold the saliency map such that a given percentage of the most salient image pixels is predicted as fixated and the rest of the image is predicted as non-fixated. Thus, when the whole image is predicted as fixated, the hit rate reaches its maximum. When we lower the threshold, only peaks of the saliency map are predicted as fixated and the hit rate is changing. The aim of accurate fixation prediction is to achieve a high hit rate with a low percentage of salient area. The adapted ROC curve summarizes the performance of a classifier across all possible percentages of salient area. The area under the adapted ROC curve (AUC) is regarded as an indication of the classification power. For the perfect classifier, the AUC equals to 1, and for the random classifier, the AUC is 0.5.

Experimental Results

To evaluate how well the proposed Weibull method predicts human fixations, we consider two eye-fixation data sets: the standard data set from [4] and an artistic data set recorded by the authors as described in details later. We use human eye-fixations as ground truth data in our experiments. In our experiments, we (1) study the consistency of the eye-fixation pattern of human subjects, (2) prove that our simple method which is based only on two parameters can compete with the state-of-the-art approaches and (3) investigate the generalization of the proposed method on a new data set.

The Eye-Fixation Data Sets

We consider two eye-fixation data sets. The first is the standard eye-fixation data set collected by Bruce and Tsotsos [4]. It contains 120 natural colour images of size 680×510 pixels with eye-fixations collected from 20 subjects. All images depict indoor and outdoor urban scenes. In addition, we recorded eye-tracking data for artistic professional photos from National Geographic wallpapers[1]. Eye-fixations of 17 subjects were collected in a free-viewing setting. All participants were naive to the purpose of the study and had normal or corrected to normal vision. Subjects viewed 49 images of size 800×540 pixels. These were selected from three categories of National Geographic wallpapers: animals, landscapes and people. Typical pictures are shown in Fig. 3. All procedures conformed with National and Institutional Guidelines for Experiments with human subjects and with the Declaration of Helsinki. Eye movements were recorded using an eye tracker (EyeLink II, SR Research Ltd.), sampling pupil position at 1000 Hz. Subjects were seated in a darkened room at 85 cm from a computer monitor and used a chin-rest so that head position was stable. To calibrate eye

[1] http://www.nationalgeographic.com/

Fig. 3 Examples of images from the data sets. The first row illustrates the Bruce&Tsotsos data set; the second row shows images from the National Geographic data set

position and to validate the calibration, subjects made saccades to the 12 fixation spots on the screen, which appeared one by one in random order. During the experiment, images were presented on a 17 inch screen (FlexScan L568) for 5 s. After each stimulus presentation, a fixation spot appeared at a random position of the screen in order to distribute first fixations uniformly over the experiment. These fixations were excluded from the analysis. Fixation locations and durations were calculated online by the eye tracker. The MATLAB psychophysics toolbox was used for stimulus presentation [3]. In addition, the Eyelinktoolbox was utilized to provide communication with the eyetracker [6].

Experiments

In our experiments, we first investigate the variability of eye-fixations across subjects in order to construct a stable ground truth. Then, we evaluate the performance of the proposed Weibull method for each single data set. Finally, we investigate the generalization of the Weibull method by a cross data set analysis. We compare the proposed method with the classical saliency map by Itti et al. [19], and with the state-of-the-art method by Bruce and Tsotsos [4]. Both implementations are unaltered code from the original authors. We assume humans to be an ideal saliency detector. Hence, performance of saliency methods is upper-bounded by the behaviour of an inter-subject model. In this model, the saliency map is generated from the fixations of training subjects, and the result is compared with the same ground truth as used in the cross-validation experiments. To construct inter-subject saliency maps, we convolve fixation locations with a fovea-sized two-dimensional Gaussian kernel ($\sigma = 1°$, i.e. 30 pixels).

Table 1 Comparison of the Weibull model

Method	AUC (SD) for Bruce&Tsotsos data	AUC (SD) for National Geographic data
Itti et al.	0.6951 (0.1048)	0.6649 (0.0893)
Bruce&Tsotsos	0.7636 (0.0831)	0.7115 (0.0784)
Weibull	**0.7639 (0.0866)**	**0.7150 (0.0848)**
Inter-subject	0.8722 (0.0426)	0.7931 (0.0530)

The mean value and standard deviation of the areas under the curve (AUC) for all considered saliency methods

The t-test has shown that AUC for the Weibull model is significantly higher ($P < 10^{-5}$) in compare to other considered models

Inter-Subject Variability

As there is a high inter-subject variability in eye movements, fixation locations are subject dependent. Hence, it is important to consider a sufficient amount of subjects in order to learn consistent patterns in eye-fixation data and to construct a stable ground truth. As we show in the next Sect. 3.4 (Table 1, Fig. 5), subjects have less consistent eye-scanning behaviour on the National Geographic data set in comparison with the Bruce&Tsotsos data set: for the inter-subject model in Table 1, we obtain an AUC of 0.7931 versus 0.8722, respectively. Therefore, we use the National Geographic eye-tracking data set to estimate inter-subject variability. We assess the variation in fixation locations for an increasing number of randomly drawn subjects. Particularly, for each image, we construct a

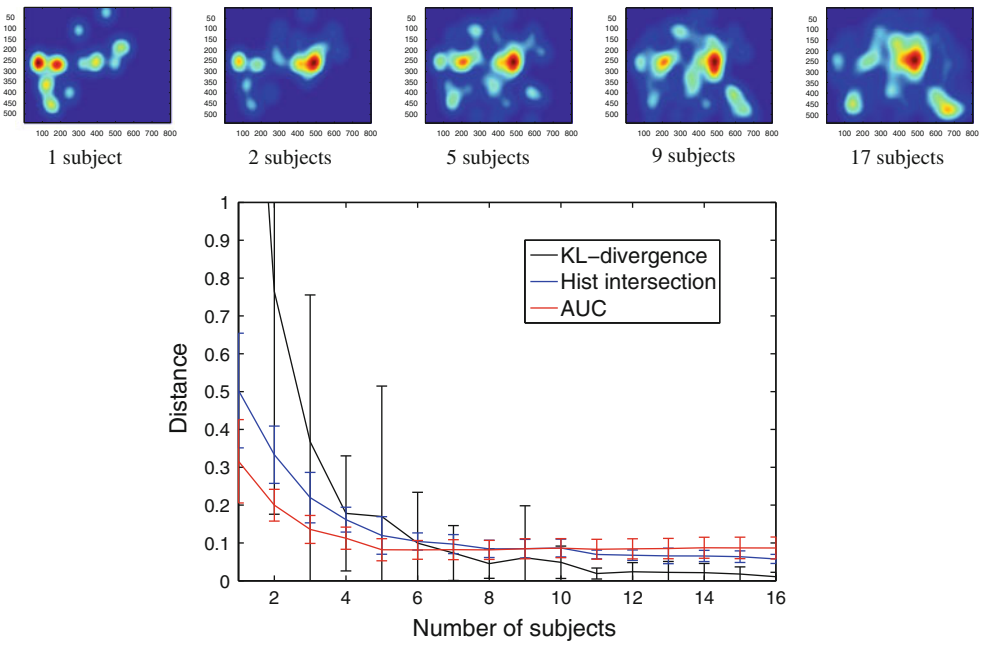

Fig. 4 The top row contains a sequence of fixation maps based on 1, 2, 5, 9 and 17 subjects. At the bottom, the stability of human fixation maps as function of the number of subjects is shown. Results are averaged over 49 images from the National Geographic data set and 30 random draws of subjects from a pool of 17. At nine subjects, the fixation maps stabilize

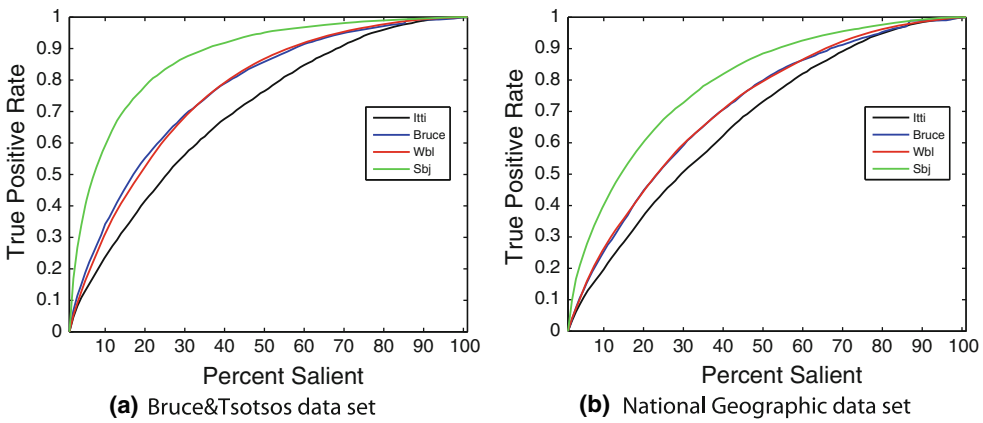

Fig. 5 ROC areas for different saliency models when evaluated on the Bruce&Tsotsos (**a**) and the National Geographic (**b**) data sets separately. The proposed Weibull model performs at the level of state-of-the-art in bottom-up saliency prediction for both data sets

sequence of fixation maps by convolving subject fixations with a fovea-sized two-dimensional Gaussian kernel ($\sigma = 1°$, i.e. 30 pixels). An example of fixation maps of the same image using an increasing number of subjects is given in the top row of Fig. 4. As can be observed from the example, the fixation map becomes stable when based on more subjects. The quantitative evaluation is summarized in the bottom of Fig. 4. The graph depicts three distance measures between fixation maps: symmetrical Kullback-Leibler divergence, (one minus) histogram intersection and (one minus) area under the ROC curve (AUC). Clearly, distance decreases as the number of subjects increases. Results show that for all measures, it is hardly possible to distinguish between fixation maps based on more than nine subjects. Hence, nine subjects are enough to construct a stable ground truth regardless of the type of considered distance measure. Given the similar behaviour of the considered distance measures, we will only report further results for the well-established AUC [4, 38, 41].

Single Data Set Analysis

We start with the evaluation of how well the proposed method predicts human fixations when the Bruce&Tsotsos and the National Geographic data sets are analysed separately. Based on the results of inter-subject variability, as analysed in Sect. 3.3,, we use nine subjects to train the classifier and the remaining subjects to test the results. We consider 5-fold cross-validation and repeat 30 times random drawing of train and test subjects in our experiments. The mean and the standard deviation of the areas under the curves are summarized in Table 1. In more detail, Fig. 5a, b show that the considered saliency methods behave similarly for both the Bruce&Tsotsos and the National

Geographic data sets. The proposed Weibull method performs similar to the state-of-the-art Bruce&Tsotsos algorithm. Both outperform the traditional Itti et al. saliency map. Subject performance is higher for the Bruce&Tsotsos data set than for the National Geographic data set, implying more consistent eye-scanning behaviour for the former data set. One possible explanation is that the artistic images from National Geographic wallpapers have more diverse content in comparison with urban images from the Bruce&Tsotsos data set. Therefore, subjects might use more diverse strategies when viewing images from the National Geographic data set. Hence, the National Geographic data set can be seen as a more difficult data set. Note there is still a large gap between human and algorithm performance, which suggests room for improvement. However, not all fixations can be explained by bottom-up features alone [38].

Cross Data Set Analysis

To investigate the generalization of the proposed method, we perform a cross data set experiment. Here, we train the parameters of our model on one data set and evaluate its performance on the other data set. Here, we again use the National Geographic and the Bruce&Tsotsos data sets. In order to minimize the differences between these data sets caused by the way they were recorded, we do not use the whole data sets in the training phase. Instead, we restrict each data set to only 17 subjects, 49 images and eye-fixations acquired within the first 4-seconds viewing time (minus the first fixation). We start with training the parameters of the Weibull method using the National Geographic data set, while evaluating the performance on the Bruce&Tsotsos test set as used in Sect. 3.4. This ensures comparability with Fig. 5. Next, we switch data sets and train from the Bruce&Tsotsos data set while evaluating the performance on the National Geographic test set. Note that the methods by Itti et al. and Bruce&Tsotsos do not involve a training phase. Therefore, their performance in these settings is the same as for the single data set analysis, see Table 1 and Fig. 5. The results are summarized in Table 2. In more detail, Fig. 6a illustrates that the parameters of the Weibull model learned from the National Geographic data set can be used to predict saliency for the Bruce&Tsotsos data set without any performance loss. Hence, the cross-validation used in Sect. 3.4 does not introduce a positive bias in the results (Table 1, Fig. 5). However, as can be seen in Fig. 6b, when the parameters of the Weibull method are trained on the Bruce&Tsotsos data set, our method does not show the same performance as when the parameters are trained on the National Geographic data set. Again, we attribute this to the higher variation in the content of images from the National Geographic data set. As some feature patterns do

Table 2 Generalization of the Weibull model. The mean value and standard deviation of the areas under the curve (AUC) for single data set analysis versus cross data set analysis

Training data set	AUC (SD) for Bruce&Tsotsos data	AUC (SD) for National Geographic data
Bruce&Tsotsos data	0.7639 (0.0866)	0.6911 (0.0882)
National Geographic data	0.7629 (0.0844)	0.7150 (0.0848)

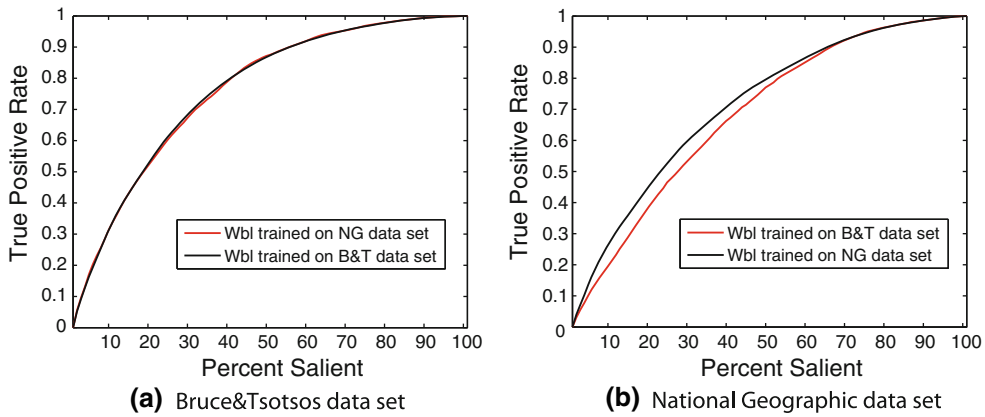

Fig. 6 ROC performance for cross data set analysis. The left graph shows ROC performance of the proposed Weibull method for the Bruce&Tsotsos data set when the proposed method is trained on the National Geographic data set, compared to when it is trained on the Bruce&Tsotsos data set itself. The right graph shows ROC performance for the National Geographic data set when proposed method is trained on the Bruce&Tsotsos data set, compared to when it is trained on the National Geographic data set itself

not occur or occur infrequently in the Bruce&Tsotsos data set, our classifier cannot learn if these are salient or not. Hence, we conclude that our Weibull model has good generalization power when the variation in the training data set is representative for the test data set.

Discussion

In this paper, we explored the link between local image statistics and human fixations by focussing on bottom-up feature-driven saliency. The influence and importance of bottom-up and top-down effects on human attention is an ongoing research question. There are many studies which show that low-level visual stimuli correlate with human fixations much better than expected by chance alone; for a review, see [16]. Moreover, pop-out features like bright spots on a dark uniform background attract attention automatically [40]. In addition to low-level features, human attention depends on high-level information, such as goals, contextual cues, important objects and image interpretation [4, 7, 27, 33, 39]. When eye-fixations are driven by very specific task ("avoid obstacles", "pickup a specific object"), the pure bottom-up saliency fails to predict fixation locations adequately [33]. However, in free-viewing settings or when considering a less specific task ("find interesting objects", "what is important in an image"), low-level features do play a significant role in fixation prediction [8, 35]. Elazary and Itti [8] have shown that interesting objects are collocated with the peaks in their bottom-up saliency map more often than expected by chance alone. Furthermore, objects usually have spatial extension, and low-level features inside the object might still play a role in task-driven saccadic eye movements. Tatler with colleagues [38] proposes the *strategic divergence* framework where people switch strategy over time. They argue that observers start looking at an image with a bottom-up strategy and later switch to more elaborative high-level object-driven strategies, possibly returning to the bottom-up strategy again. To conclude, although bottom-up saliency alone cannot explain fully the richness of mechanisms of human attention, it does play a role in where people look at, and the complete model of attention should incorporate both feature- and task-driven saliency. In this paper, we have explored the link between the location of eye-fixations and natural image statistics modelled by the two parameters Weibull distribution.

Comparison with Previous Works

A number of studies investigate how natural image statistics influence the locations of human fixations [28–30]. Despite the variety of the considered low-level image features, most researchers agree that contrast distribution plays a significant role in guidance of eye movements. Usually, the local contrast is defined as the standard deviation of the image intensities within some small region, divided by the mean intensity within that region, i.e., the local root mean square RMS contrast. However, as the distribution of natural images is non-Gaussian [25, 37], in this paper, we follow [14, 17] and model image contrast with the Weibull distribution. Figure 2 illustrates that the two-parameter Weibull distribution fits the local contrast statistics adequately well. Baddeley and Tatler [1] argue that high-frequency edges turn out to have most impact on fixation prediction, whereas contrast is highly correlated. The next most important feature in their analysis is low-frequency edges. Geusebroek and Smeulders [14] show both contrast and edge frequency to be simultaneously captured by the Weibull distribution. It allows to combine these two image regularities in an elegant way taking into account the strong correlation between them. In our analysis, we investigate a joint distribution of the local contrast and the edge frequency and, thereby, combine low-level image features that are known to be the most powerful in fixation prediction. Moreover, we do not separate high- and low-frequency edges. Instead, we use the minimal reliable scale selection principle [30] as discussed in Sect. 2.1.2 and implicitly consider edges over the available frequency range all together. Inspired by the centre-surround receptive field design of neurons in the retina [18], several successful saliency models are based on comparison of centre-surround regions at each image location [4, 5, 10, 15, 19, 23, 36]. Intuitively, image locations which deviate from their surrounding should be salient. Itti et al. [19] consider visual features salient if they have different brightness, colour or orientation than the surrounding features. Overall, their model combines a total of 42 feature maps into a single saliency map. In contrast, we do not make any assumption about patterns in the spatial structure of feature responses and base our model on comparison of local image statistics with statistics learned from fixation and non-fixation regions. Table 1 and Fig. 5 show that the proposed Weibull method outperforms the method by Itti et al. It might indicate the advantage of direct training of the model parameters from an eye movement data set. Moreover, the higher performance of our method might be due to the explicit modelling of the correlation between image features. Bruce and Tsotsos [4] follow an information-theoretic approach and use *information maximization* sampling to discriminate centre-surround regions. They calculate Shannon's self-information based on the likelihood of the local image content in the centre region given the image content of the surround. Regions with unexpected content in comparison with their surrounding are more informative, and thus salient. As shown in Table 1

and Fig. 5, our model achieves a performance comparable with the elaborate approach by Bruce and Tsotsos, while we use as few as two parameters learned from a set of images with associated eye movements. We have explored the generalization of the proposed method by considering the two eye movements data sets: a standard data set from [4] with urban images, and an artistic photo collection with diverse context from National Geographic wallpapers. Examples of images from both data sets are shown in Fig. 3. Table 2 and Fig. 6 show that training the parameters of our Weibull method on the National Geographic data set and testing it on the Bruce&Tsotsos data gives the same results as both training and testing on the Bruce&Tsotsos data. However, for the National Geographic data set, there is a small drop in performance when the parameters of the Weibull model are trained on the Bruce&Tsotsos data instead of the National Geographic. We attribute this to the higher variation in image content from the National Geographic data. We conclude that the proposed model has good generalization power when the variation in the training data set is sufficiently diverse.

Conclusions

We have presented a Weibull method of saliency prediction based on the local image statistics learned at fixated and non-fixated locations. Our approach combines image contrast and edge frequency as captured by the two parameters of the Weibull distribution into a single statistical model. Using the joint distribution of these parameters and a simple log-likelihood test, we achieve a performance comparable with state-of-the-art bottom-up saliency methods.

Our results show that as few as two values per image region are already enough to indicate saliency, the two values indicating contrast and frequency. Baddeley and Tatler [1] have already shown contrast and frequency to be important for visual saliency. However, the authors highlight that these two features are correlated for natural images and propose high-frequency edges to be the most indicative salient feature. In our Weibull method, we cope with the correlation by considering the joint distribution of the contrast and frequency parameters. Despite its simplicity, our model has good generalization power when its parameters are trained on a diverse data set.

References

1. Baddeley RJ, Tatler BW. High frequency edges (but not contrast) predict where we fixate: a Bayesian system identification analysis. Vis Res. 2006;46(18):2824–33.
2. Bonnans J, Lemaréchal C. Numerical optimization: theoretical and practical aspects. Springer, New York; 2006.
3. Brainard DH. The psychophysics toolbox. Spat Vis. 1997;10(4):433–6.
4. Bruce N, Tsotsos J. Saliency, attention, and visual search: an information theoretic approach. J Vis. 2009;9(3):1–24.
5. Cerf M, Harel J, Einhauser W, Koch C. Predicting human gaze using low-level saliency combined with face detection. Adv Neural Inf Process Syst. 2008;20:241–8.
6. Cornelissen FW, Peters EM, Palmer J. The eyelink toolbox: eye tracking with MATLAB and the psychophysics toolbox. Behav Res Methods Instrum Comput. 2002;34(4):613–7.
7. Einhauser W, Spain M, Perona P. Objects predict fixations better than early saliency. J Vis 2008;8(14):1–26.
8. Elazary L, Itti L. Interesting objects are visually salient. J Vis 2008;8(3).
9. Elder James H, Zucker Steven W. Local scale control for edge detection and Blur estimation. IEEE Trans Pattern Anal Mach Intell. 1998;20(7):699–716.
10. Gao D, Mahadevan V, Vasconcelos N. On the plausibility of the discriminant center-surround hypothesis for visual saliency. J Vis 2008;8(7):1–18.
11. Geisler WS. Visual perception and the statistical properties of natural scenes. Ann Rev Psychol. 2008;59:167–92.
12. Geusebroek JM, van den Boomgaard R, Smeulders AWM, Dev A. Color and scale: the spatial structure of color images. Eur Conf Comput Vis. 2000;1:331–41.
13. Geusebroek JM, van den Boomgaard R, Smeulders AWM, Geerts H. Color invariance. IEEE Trans Pattern Anal Mach Intell. 2001;23(12):1338–50.
14. Geusebroek JM, Smeulders AWM. A six-stimulus theory for stochastic texture. Int J Comput Vis. 2005;62(1):7–16.
15. Harel J, Koch C, Perona P. Graph-based visual saliency. Adv Neural Inf Process Syst. 2007;19:545–52.
16. Henderson JM. Human gaze control during real-world scene perception. Trends Cogn Sci. 2003;7(11):498–504.
17. Huang J, Mumford D. Statistics of natural images and models. IEEE Conf Comput Vis Pattern Recognit 1999;7(6).
18. Hubel DH, Wensveen J, Wick B. Eye, brain, and vision. Scientific American Library, New York; 1988.
19. Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell. 1998;20(11):1254–9.
20. Johnson NL, Kotz S, Balakrishnan N. Continuous univariate distributions. Vol. 1. Wiley, New York; 1995.
21. Judd T, Ehinger K, Durand F, Torralba A. Learning to Predict Where Humans Look. Int Conf Comput Vis. 2009.
22. Kadir T, Brady M. Saliency, scale and image description. Int J Comput Vis. 2001;45(2):83–105.
23. Kienzle W, Franz MO, Scholkopf B, Wichmann FA. Center-surround patterns emerge as optimal predictors for human saccade targets. J Vis. 2009;9(5):1–15.
24. Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis. 2004;60(2):91–110.
25. Mallat SG. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell. 1989;11(7):674–93.
26. Mante V, Frazor RA, Bonin V, Geisler WS, Carandini M. Independence of luminance and contrast in natural scenes and in the early visual system. Nat Neurosci. 2005;8(12):1690–7.

27. Navalpakkam V, Itti L. Search goal tunes visual features optimally. Neuron 2007;53(4):605–17.
28. Parkhurst D, Law K, Niebur E. Modeling the role of salience in the allocation of overt visual attention. Vis Res. 2002;42(1): 107–23.
29. Rajashekar U, van der Linde I, Bovik AC, Cormack LK. Foveated analysis of image features at fixations. Vis Res. 2007; 47(25):3160–72.
30. Reinagel P, Zador A. Natural scene statistics at the centre of gaze. Netw Comput Neural Syst. 1999;10(4):341–50.
31. Renninger LW, Coughlan J, Verghese P, Malik J. An information maximization model of eye movements. Adv Neural Inf Process Syst. 2005;17:1121–8.
32. Ross SM. Introduction to probability and statistics for engineers and scientists. Elsevier, Amsterdam; 2009.
33. Rothkopf CA, Ballard DH, Hayhoe MM. Task and context determine where you look. J Vis. 2007;7(14):1–20.
34. Scholte HS, Ghebreab S, Waldorp L, Smeulders AWM, Lamme VAF. Brain responses strongly correlate with Weibull image statistics when processing natural images. J Vis. 2009;9(4):1–15.
35. Schumann F, Einhauser W, Vockeroth J, Bartl K, Schneider E, Konig P. Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. J Vis. 2008;8(14).
36. Seo HJ, Milanfar P. Nonparametric bottom-up saliency detection by self-resemblance. In: IEEE conference on computer vision and pattern recognition, 1st international workshop on visual scene understanding; 2009.
37. Simoncelli EP, Olshausen BA. Natural image statistics and neural representation. Ann Rev Neurosci. 2001;24(1):1193–216.
38. Tatler BW, Baddeley RJ, Gilchrist ID. Visual correlates of fixation selection: effects of scale and time. Vis Res. 2005;45(5): 643–59.
39. Torralba A, Oliva A, Castelhano MS, Henderson JM. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. Psychol Rev. 2006;113(4):766–86.
40. Treisman AM, Gelade G. A feature-integration theory of attention. Cogn Psychol. 1980;12(1):97–136.
41. Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW. SUN: a Bayesian framework for saliency using natural statistics. J Vis. 2008;8(7):1–20.