Available online at www.sciencedirect.com
SCIENCE DIRECT®
Vision Research xxx (2006) xxx–xxx
Vision Research
www.elsevier.com/locate/visres

High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis

Roland J. Baddeley [a,*], Benjamin W. Tatler [b]

[a] Department of Experimental Psychology, University of Bristol 8, Woodland Road, Bristol BS8 1TN, UK
[b] Department of Psychology, University of Dundee, Dundee DD1 4HN, UK

Received 19 August 2005; received in revised form 13 February 2006

Abstract

A Bayesian system identification technique was used to determine which image characteristics predict where people fixate when viewing natural images. More specifically an estimate was derived for the mapping between image characteristics at a given location and the probability that this location was fixated. Using a large database of eye fixations to natural images, we determined the most probable (a posteriori) model of this mapping. From a set of candidate feature maps consisting of edge, contrast and luminance maps (at two different spatial scales), fixation probability was dominated by high spatial frequency edge information. The best model applied compressive non-linearity to the high frequency edge detecting filters (approximately a square root). Both low spatial frequency edges and contrast had weaker, but inhibitory, effects. The contributions of the other maps were so small as to be behaviourally irrelevant. This Bayesian method identifies not only the relevant weighting of the different maps, but how this weighting varies as a function of distance from the point of fixation. It was found that rather than centre surround inhibition, the weightings simply averaged over an area of about 2 degrees.

Keywords: Salience; Eye movements; Bayesian inference; System identification; Reverse correlation

1. Introduction

Humans have a large over representation of the centre of their visual field (the fovea) relative to the periphery. Therefore, if we want to see the fine details of an object, we point our eyes at it. For many tasks such as reading, or skilled visuo-motor tasks such as threading a needle, directing our eyes accurately at a particular target is not only desirable but is necessary. As a result, if we are to understand visual processing, we must characterise not only how we process individual fixations (the most studied aspect of vision), but also why we choose to fixate the locations we do.

For some time it has been clear that the task performed by a subject is important in determining fixation behaviour (Buswell, 1935; Land & Hayhoe, 2001; Yarbus, 1967). We do not simply reflexively fixate particular regions of our visual world. For example the pattern of fixations when a person has to ascertain the weather is very different from that made when they attempt to infer the thoughts of people depicted in a scene (Nelson, Cottrell, Movellan, & Sereno, 2004).

In contrast, much recent work has emphasised the role of low-level visual features in choosing fixation locations. Here, it is proposed that a map is constructed that represents the "salience" at given locations in visual space (Itti & Koch, 2000; Kadir & Brady, 2001; Parkhurst & Niebur, 2003; Renninger, Coughlan, & Vergheese, 2005), with the map being based on the low-level visual characteristics of a scene. The important factors in determining the low-level "salience" of various proposed features have been explored computationally by two related approaches.

* Corresponding author.
E-mail address: roland.baddeley@bristol.ac.uk (R.J. Baddeley).

0042-6989/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2006.02.024

The first approach explores a given proposal for a salience map by combining visual characteristics that are known to be extracted by early cortical areas. For instance, Itti and Koch (2000) constructed a salience map by first extracting representations of colour, contrast, and orientation. After appropriate normalisation and contrast enhancement, these representations were then combined to make an overall salience map. Essentially this implements the sensible strategy of labelling any parts of the scene that are different from the average as salient. It can then be shown that regions that are more "salient" are more often fixated by observers. Unfortunately, it is unclear which of the many architectural assumptions are important in generating predictions. For example, the model assumes that both contrast and edges are important, yet this may not be the case.

A more systematic approach has been adopted by a number of researchers who have looked to see if there are any differences between visual characteristics at locations that were fixated by observers and those of locations which were not (Parkhurst, Law, & Niebur, 2002; Reinagel & Zador, 1999; Tatler, Baddeley, & Gilchrist, 2005). The basic picture that emerges (Tatler et al., 2005) is: (1) contrast, luminance, orientation energy, and chromaticity all differ between fixated locations and non-fixated locations; (2) these differences are larger for orientation and contrast than chromaticity and luminance, and larger for higher frequencies than lower; (3) however, while these differences are hugely statistically significant (because of large numbers of measurements), the magnitudes of these effects are modest at best, with a large overlap between the statistics at fixated and non-fixated locations. Lastly, although the consistency in where people look changes over time, (indicating that observers' strategies may change over time), the difference in the image statistics at fixated and non-fixated regions do not change, (indicating that the low-level representation of salience does not appear to change over time).

This pattern of results is, unfortunately, rather difficult to interpret. The main problem arises because a given feature will tend to be correlated across neighbouring locations, similarly different features will be correlated within a given location (for example, in the case of luminance, see Baddeley, 1997, or edges, see Elder & Goldberg, 2002). For a particular location, different features will also tend to be correlated (e.g., edges will be associated with high contrast). In short, for natural images, almost every regularity is strongly associated with every other regularity. This means that if we find that the contrast at fixated regions is greater than at non-fixated regions, it could be that this contrast is contributing to salience, or it could be that what is really driving the system is some other regularity, such as edges, to which contrast just happens to be correlated.

A second related problem is that a plausible model of salience requires more than the simple specification of which visual regularities contribute to it; it must also specify *how* they contribute. For instance in Itti and Koch (2000), the model specifies how each regularity is spatially integrated; how feature contrast is calculated within each feature map; how the outputs of 42 different feature maps are combined into a single salience map; and how the location of maximum salience is calculated. Unfortunately, by necessity many of these architectural assumptions are somewhat arbitrary. We cannot identify the architecture of the model simply by calculating the difference between the statistics of fixated and non-fixated locations.

In this paper, we present the results of applying a statistical system identification technique using a regularised generalised linear model (McCullagh & Nelder, 1989). This method has the possibility of capturing relatively sophisticated mappings between image characteristics and "salience," and can deal with the problem of correlations in the image features. Here by "salience" we simply mean the conditional probability of fixation given the image characteristics and leave discussion of what this implies to the discussion. By recording eye movements when viewing natural images, and operationalising salience as the conditional probability of fixation given the image statistics at that location, it is possible to estimate a model of the mapping between images and fixation probability. The nature of the best model will tell us about what and how low-level features contribute to fixation behaviour.

Unfortunately, in order to be able to have a model flexible enough to capture both the contribution of multiple regularities (luminance, contrast, edges, etc.), at different spatial scales, and also capture the spatial integration operating in these maps, a potentially very large number of parameters needs to be identified. Even if we only look at three regularities, each at two spatial scales, and parameterise each using a 12×12 spatial integration window, this requires a model with 864 parameters. Though our data sets are very large, standard maximum likelihood system identification techniques fail to give a good characterisation of the system (the models over fits the data, fitting both the signal and noise, and hence fails to generalise to new data). We therefore used a Bayesian regularisation technique employing a family of priors (the so called bridge priors, Fu, 1998), that have recently been employed in state of the art statistical pattern recognition problems (Frank & Friedman, 1993; Fu, 1998; Hastie, Tibshirani, & Friedman, 2001; Ng, 2004). We use this flexible family of priors since they can deal with both "distributed" mappings, where many features make small contributions to salience, and also "sparse" mappings, where the salience is dominated by a few highly important features. Previous methods used to constrain high dimensional mappings (including singular value decomposition, principal component analysis, and Fourier-based techniques) implicitly bias the identified model to "distributed" difficult to understand solutions even if the reality is simple. By implicitly estimating the sparsity from the data, we avoid the problem of inappropriate biases towards distributed or sparse mappings: if the evidence is that the mapping is sparse, a sparse constraint or prior is used, whereas if the mapping is distributed, a distributed (Gaussian) prior is used.

For simplicity, we investigated three regularities that represent a subset of those calculated in V1. More specifically, we calculated maps to extract luminance, contrast and edges, each at two spatial scales. Though potentially an important feature, colour was not modelled because previous work indicated that it was not a strong contributor to fixation selection (Tatler et al., 2005), and because there are technical difficulties associated with the representation of colour (the appropriate colour space and monitor calibration issues).

With this data set of images, together with the associated eye movements, this system identification technique allows us to identify (1) the relative contributions of the different maps; and (2) the spatial weighting function that is used in integrating these feature maps. We found that using these six potential feature maps, the mapping is dominated by high frequency edges. This is a solution that could not have been found using previous techniques without vastly larger data sets. Previously claimed correlates of fixation selection (such as contrast) are shown simply to be artefacts of their correlation with edges. We argue that contrast does not contribute to "salience" and relate this claim to the previous literature.

2. Methods

2.1. Images and data collection

Fourteen observers viewed 48 images at a distance of 60 cm. The images subtended approximately $30° \times 22°$ of the observer's field of view and had a maximum luminance of 22 cd/m^2. The observers were instructed to perform a memory task, with questions asked at the end, and the viewing time was sampled from a uniform random distribution from 1 to 10 s. Details of the images and memory task, together with technicalities of the eye movement recording can be found in Tatler et al. (2005). This procedure was used to collect 8843 fixations.

2.2. Visual features at fixation

This paper uses Bayesian generalised linear model regression techniques to investigate the relationship between the image statistics and fixation. As the input to this regression we need parameterisations of the image that make various image characteristics explicit. We chose to use three different regularities (edge content, contrast, and brightness), and each of these regularities was calculated at two different spatial scales. Details of the image processing used in extracting these input maps are given in Tatler et al. (2005), and are only briefly covered here. For the "brightness" map we convolved each image with a Gaussian (with a standard deviation of 2.7 cycles per degree for the high frequency filter and 0.675 cpd for the low frequency filter). The mean value was then subtracted, all values squared, and then standardised by dividing by their standard deviation for a given image. For the contrast, images were transformed by convolving with a difference of Gaussian filter (with standard deviations of 2.7 and 0.675 cycles per degree for the centre Gaussians), and again subtracting, squaring, and standardising the results as before. Lastly, for edges, the images were convolved with four (odd-phase) Gabors, with envelope standard deviations of 2.7 and 0.675 cpd for the two spatial frequencies modelled, orientated at 0, 45, 90, and 135 degrees. Again the output was squared, the maximum output of the four different orientations calculated, and the maps again standardised by subtracting the mean, and dividing by the standard deviation.

After processing the 48 images, we extracted the image features in a 3 by 3 degree patch centred at fixation for all the 8843 fixations. The values of the image characteristics in this patch were then parameterised by a 12×12 pixels matrix of values by down sampling the original patch (hence each parameterised pixel corresponds to 4 image pixels). We also collected image patches from the same 8843 locations, but from different images (corresponding to locations not actually selected for fixation by the observers). This matching of the spatial sampling distribution for selecting non-fixated image statistics is important because it removes artefacts that arise from spatially non-uniform sampling of scenes by the eye. For a detailed explanation of these issues see Tatler et al. (2005).

2.3. Regression

This paper sets out to get an explicit mapping between image properties and the probability that a location was fixated e.g., $P(f_i = 1|I_i)$, where $f_i = 1$ if the ith location was fixated and 0 if it was not. I_i is a vector of parameters representing the characteristics at the fixated or non-fixated location (see below). Since a location can either be fixated or not and we are attempting to estimate this probability, it is inappropriate to use a least squares criteria (which implies a normal distribution), and instead we use a generalised linear model with a logistic link function (logistic regression), implying a Bernoulli distributed response variable. Hence:

$$P(f_i = 1|I_i) = 1/(1 + \exp(-S_i)),$$

where $S_i = \sum_j w_j \cdot I_i^j + c$, and $\mathbf{w}$ is a vector of weights that map between the image characteristics and can be thought of in terms of a receptive field, c is a constant, and the sum is over all j components of the image parameterisation. A good introduction to the assumptions behind this approach from a Bayesian perspective can be found in Bishop (1996).

The most common way to find the optimal $\mathbf{w}$ is by maximising the log likelihood. This can work when; (1) the number of dimensions in I and w is small; (2) the amount of data is very large; and (3) the variables in I are relatively uncorrelated. Unfortunately the first and third criteria are violated in our data, as the dimensionality is very high (up to 865 dimensions), and all aspects of natural images tend to be highly correlated. Though the data set is large, the first two problems mean that although the log likelihood is high when trained on a given data set, this is uninformative since the models found fail to generalise to new data sets (see section on cross validation). This problem of over fitting can be greatly reduced by the appropriate use of priors. Rather than maximising the log likelihood, we in addition minimise some function of the parameters w (in Bayesian terms, we place a prior on the weights). Specifically, we maximise the function:

$$L(w|I, \lambda, \beta) = L(w|I) - \frac{1}{\lambda}|w|^\beta,$$

where $L(w|I)$ is the log likelihood of the data given the parameterised image, $|w|^\beta$ is a function that encourages "simple" parameters, λ measures the relative importance of this simplicity term compared to the data term, and $1 \leqslant \beta \leqslant 2$ (the bridge parameter (Fu, 1998)), specifies the relative importance of large parameters as opposed to small. Two special cases of β are of interest. If $\beta = 2$, the prior minimises the length of the regression vector w. If the input is corrupted by independent identically distributed noise, minimising this term will minimise the noise. This form of regression is known as ridge regression in the statistics literature, or weight decay in neural networks. It is closely related to principal components, and Fourier-based methods that have previously been applied to the identification of neuronal responses (Sen, Theunissen, & Doupe, 2001). It corresponds to a Gaussian prior and, importantly, it favours distributed global solutions with a large number of small parameters, as opposed to sparse solutions with a small number of large weights. If many maps make small but significant contributions to the salience, then values of $\beta \approx 2$ should perform best. If instead $\beta = 1$, the so called lasso solution, the sparse local mappings are favoured. This corresponds to a double sided exponential (or Laplace) prior. If the mapping between image characteristics and salience is sparse with only a few image features making large con-

tributions to salience, $\beta \approx 1$ will perform best. Details of ridge, lasso, and bridge priors together with practical advice on fitting such models can be found in (Hastie et al., 2001).

We now have a (set of) models, and a family of evaluation functions but we still need to find an optimal model consisting of $P+1$ weight parameters, (where P is the number of input variables and can be as large as 865 different inputs including a bias term), a sparsity parameter λ, and β, the relative contribution of the data and simplicity terms. This was achieved by initially optimising the model with $\beta = 0$, using $10 \times P$ iteration of scaled conjugate gradient descent; β was then increased by a small percentage and the model re-optimised using P iterations of conjugate gradient descent, 100 iterations of quasi Newton, followed by another $1.5 \times P$ iterations of scaled conjugate gradient descent. This process of increasing β and reoptimising the parameters using the previous values as a starting point was repeated until cross validation performance was found to increase. This results, for a given input parameterisation and λ, in a set of models, each associated with different values of β. Whilst this optimisation technique is more than required for models where $\lambda = 2$ (and standard optimisation methods work well), when $\lambda \to 1$, the fact that the differentials are not well behaved means that the rather computationally intensive search method was required.

To compare different models, 5-fold cross validation of the log likelihood was used. The model is trained on 4/5ths of the data, and then the log likelihood for the excluded 1/5 was calculated (note we only compute the log likelihood, and do not include the prior in the model evaluation). This was repeated five times, each time excluding a different 1/5th of the data. This results in five unbiased estimates of the log likelihood. Model comparison was carried out by performing a matched sample t test on the log likelihoods. Model comparison using chi squared approximations to the log-likelihood ratios gave similar results. All results reported here are highly significant ($P < 10^{-4}$).

All model comparisons are reported in terms of the average log-likelihood ratio between either two hypotheses, or a given hypothesis and chance. We report the average (cross-validated) log-likelihood ratio (in bits) rather than the total ratio as the former measures the magnitude of the difference, whilst the later is more appropriate for assessing the statistical significance. This figure is zero if two models make equivalent predictions, and gives the average ratio of the likelihood for a given point. Note all logarithms were to the base two.

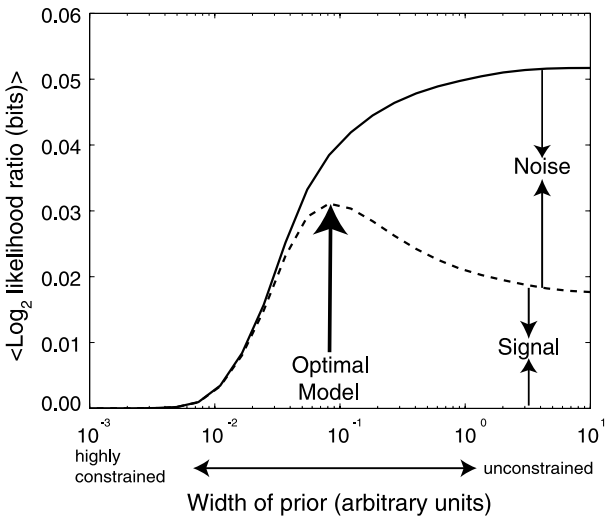

Fig. 1. This figure shows the need for priors when performing high dimensional system identification with realistically sized data sets. Shown is the average log-likelihood ratio for a model both for the data that was used to train it (solid line), and for new test data (dashed line) generated by the same subjects at a later time. This is plotted as a function of the width of prior imposed on the system. With a narrow and constraining prior, the model has no flexibility and performance on both training and test is effectively at chance for both training and test data. With effectively unconstrained models, given the number of parameters, the model is able to not only model the signal (the component shared by training and test data), but also the vagaries of the noise. This means that though the model when evaluated on the training data appears to be good, this performance is artificial and fails to generalise to any new data. In between the over and under-constrained cases is one where the model has sufficient flexibility to fit the data, but not enough to fit the noise. This compromise will constitute a much better characterisation of the system than the unconstrained maximum likelihood case, and the use of such priors or constraints has proved vital in much of statistical pattern recognition. All subsequently reported models are optimised for both the strength and sparsity (lambda) of the prior.

3. Results

A constraint on the parameters is necessary to find a good characterisation of the mapping. This was done by fitting the generalised linear model to estimate the probability of fixation given all of the image feature maps as the input. An initial crude search of different sparsity values was performed comparing test performance when $\beta = \{1, 1.5 \text{ or } 2\}$. Using this limited set of sparsities, performance was optimal when $\beta = 1.5$ (when only this limited range was tested), and all results reported later are for $\beta = 1.5$ unless otherwise stated (see later). In terms of a prior, this is far sparser than a Gaussian.

The model performance was then evaluated for a range of constraints from one extreme where models have no effective constraints ($\lambda = 0$ or the width of the prior is infinite), to the other where the only effective free variable is the bias ($\lambda \to \infty$ or the prior is infinitely narrow). Fig. 1 shows the average log likelihood of both training and test data as a function of the degree of constraint with all maps as input. As can been seen, the likelihood of the training set data steadily increases as the level of constraint decreases. In contrast, although it increases the flexibility of the model, decreasing λ initially improves the cross validation performance. However, beyond a certain level of flexibility, performance decreases as the increased flexibility of the model allows it to fit the noise and therefore, tells us little about the underlying system. For the rest of this paper, we report models where both the strength of the prior (favouring simple over complex models) and its nature (favouring sparse or distributed solutions) are optimised for the ability to predict unseen (cross-validated) data.

Fig. 2A shows the performance of the individual maps in isolation at predicting the probability of fixation, and reveals a pattern familiar from previous studies: high frequency edges and contrast on their own are the most discriminatory. Note that these simple results extend previous work in that the spatial weighting function has also been estimated (see later). This does not however appear to have a significant effect on the pattern of relative contribution.

While the previous calculation takes into account the within map correlations, by only analysing each map individually, we ignore the potentially large effects of between map correlations (e.g., an edge will usually involve lumi-

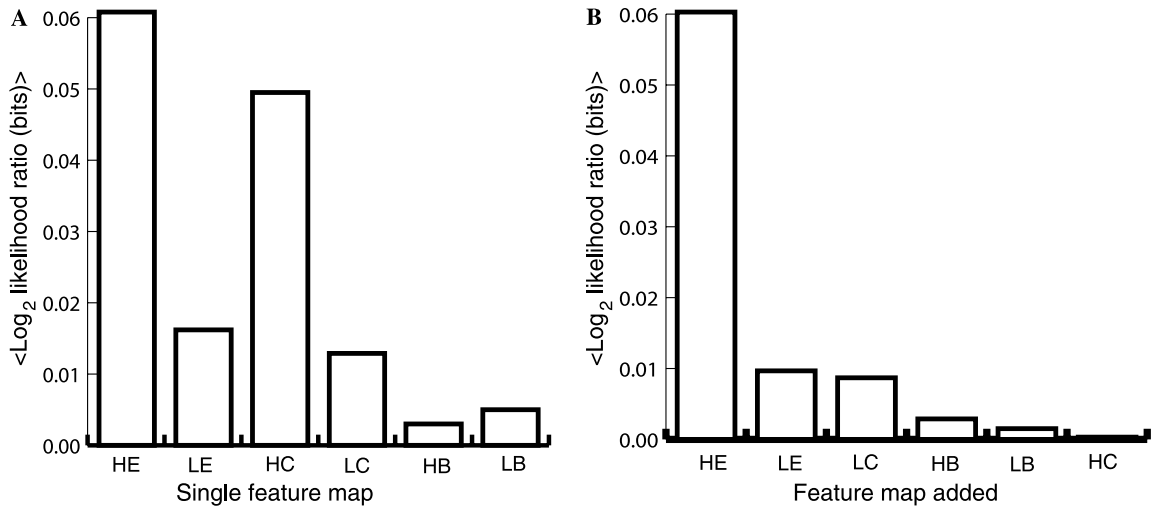

Fig. 2. The information contributed by each feature map individually (A) and in combination with other maps (B). (A) Model performance after the spatial weighting for each map is optimised individually and shows a result familiar from previous studies: when treated individually, high spatial frequency edges and contrast contribute strongly, and brightness at both low and high spatial frequencies is relatively unimportant. The performance is described in terms of the average log-likelihood ratio to the null model ($P = 50\%$ independent of input). In contrast, (B) the additional information contributed by each regularity when added in a stepwise manner (e.g., the best single regularity is added first, followed by the next). In this case the mapping is dominated by high frequency edges with high spatial frequency contrast contributing effectively nothing.

nance contrast). To take these into account, we initially used a stepwise procedure. First, the map that provided the most information was calculated (this is high frequency edges). Then the map that, when combined with the first, provided the most additional information was found (with the maps reoptimised). This procedure was repeated successively adding each map until all six were added: all maps added statistically significant (but small) amounts of information, with the exception of high frequency contrast.

This stepwise procedure, that takes into account between and within map correlations, results in a very different picture from that found when the maps were analysed individually. Although high frequency contrast is the second most informative feature on its own, it is statistically irrelevant when combined with other features (high frequency edges in particular). The simplest explanation of this pattern of results is that only (high frequency) edges are predictive of fixation location. However, contrast is able to act as a poor edge detector in the absence of an explicit representation of edges. This interpretation is supported by the pattern of log likelihoods: if we take the high frequency edge map (average log-likelihood ratio of 0.0603), adding contrast has a minimal effect (mean log-likelihood ratio of 0.0604). Similarly, high frequency contrast alone acts as a poor edge detector (mean log-likelihood ratio of 0.049), but adding the better edge detecting ability of the high frequency edge map results in a large improvement in performance (increasing from 0.049 to 0.0604). In short, the information present in the high frequency contrast map is a subset of the information present in the equivalent edge map. This emphasises the critical role of taking into account the between map correlations: if we simply consider the maps on their own, then we could have (and have previously) claimed, an important role for contrast. In fact this is simply due to the fact that whenever we have a luminance defined edge, by necessity, we have contrast.

While the contribution of high frequency contrast was not significant, the contributions of the other maps were. One possible reason for this is that the mapping between images and fixation probability is dominated by edges, but we have failed to capture some important non-linearity in the edge system. The contribution from these other maps may simply be the model trying to compensate for this poorly fitted non-linearity (an effect sometimes known as leakage). While it is not possible to explore all possible non-linearities, one family seems worth exploring. The particular edge detection system we used squares the outputs of the Gabor filters [an approximation of the local energy, Morrone and Burr (1988)]. Although this is a sensible guess, it is far from clear from the literature whether or not this is the best choice. We therefore explored applying various power functions to the output of the filters.

As can be seen from Fig. 3A, for this problem the squared non-linearity is far from optimal (this would correspond to a power of one). Since the best results are obtained using an exponent of 0.25 operating on the squared output, the best guess at the non-linearity is that it is compressive (approximately square root). The effect of getting the non-linearity correct is highly significant: the improvement added by optimising the non-linearity on the high frequency edge detector, (0.06 bits e.g., from a log-linear ratio of 0.06 bits for the edge only model to 0.118 for the non-linear model), is larger than adding all the other maps, (improvement from 0.06 to 0.0822 bits). Exploring the use of this non-linearity on the other maps showed no significant improvement.

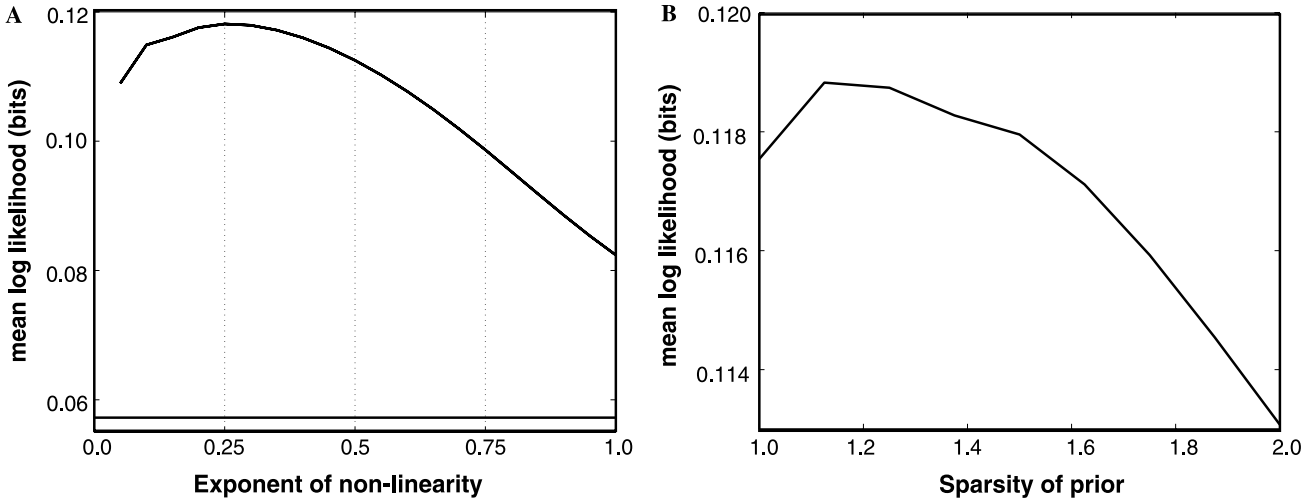

Fig. 3. Optimising aspects of the model. (A) The average log-likelihood ratio as a function of the exponent of a power law non-linearity applied to the output of the edge filter map. This shows that the mapping between visual characteristics and fixation probability is much better captured if the edge filters are subjected to a compressive non-linearity. Since the filters originally had their outputs squared, the exponent of 0.25 of the best fitting model corresponds to an approximately square root non-linearity. (B) The effects of different types of priors. The extreme right of the graph corresponds to an beta exponent of two (a Gaussian prior). This very popular prior encourages distributed solutions and also approximates most methods previously used in reverse correlation. As can be seen, this solution is substantially improved upon by a sparser prior with a value of lambda of 1.2 being near optimal. Such a prior encourages a few large weightings as opposed to many small ones, and this seems a better characterisation of the mapping than the more distributed solutions proposed previously.

The results shown so far have shown which features contribute to increasing fixation probability. However, the method can also be used to derive the spatial integration characteristics of feature selection around the centre of fixation. We therefore optimised the mapping between all the maps taking the above findings into consideration. First, we optimised the sparsity more accurately. A highly sparse value of $\lambda = 1.1$ was found to be best (Fig. 3B). Second, we used the optimised non-linearity for the high frequency edge map.

Fig. 4 shows the weighting functions for the six maps of the optimal model, and four points are of note: (1) the mapping is (unsurprisingly given Fig. 2B) dominated by an excitatory mapping to high frequency edges. (2) This weighting function is purely excitatory and does not show an inhibitory surround or "contrast enhancement," a feature of a number of models of low-level salience (for instance Parkhurst et al., 2002). (3) The area of integration is about 1.5–2 degrees; approximately the size of the fovea. (4) The second most informative map (low frequency edges) is inhibitory which is also true of low frequency luminance. This inhibition is difficult to see since, although it is the second largest contributor to salience, its absolute magnitude is far smaller than that of high frequency edges. Therefore Fig. 5 shows just this low frequency map (Fig. 5B) together with the map that would have resulted if the correlations with other maps were ignored (Fig. 5A). The inhibitory nature of low frequency edges (low frequency luminance was also found to be inhibitory), again emphasises the importance of using a multivariate approach. If we had analysed each map on its own (as we and others have done previously: Parkhurst et al., 2002; Tatler et al., 2005) we would have concluded that low frequency edges indicated positive salience whereas in fact, the opposite is true. Therefore our results, and those of other groups' are an artefact of not taking into account the contribution of high frequency edges.

4. Discussion

Previous explorations of visual features at fixation suffer from the limitation that they do not appropriately account for correlations within and between the features that they evaluate. Using a generalised linear model with an empirically optimised prior on the regression weights, we are able to take these correlations into account and have shown that: (1) the mapping between image statistics and the probability of fixating a location for the memory task investigated is dominated by high frequency edges. (2) The edge map outputs are averaged over an approximately two degree area, (3) there seems little evidence for within map contrast enhancement, but (4) strong evidence for a compressive (square root) non-linearity, and (5) between map inhibition (from low frequency edges in particular).

The strongest result found here is that, to first approximation, the only image characteristic that is a reliable predictor of where we are likely to look is the presence of high frequency edges. In particular, high frequency contrast does not discriminate fixated and non-fixated locations. This is at variance with a number of models (such as that of Itti & Koch, 2000) where contrast is one of the major input features to the final salience map. It at first seems incompatible with the various studies that have shown that contrast is significantly higher at fixated locations than

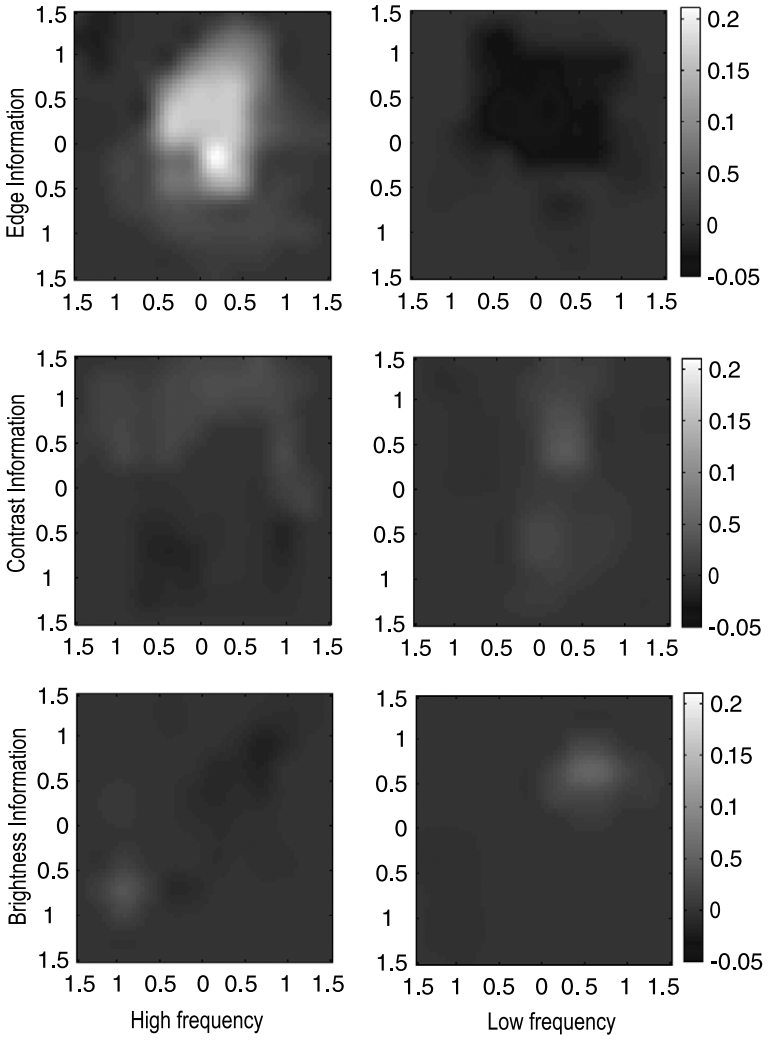

Fig. 4. The spatial weighting functions of the best fitting generalised linear model. As can be seen, this mapping is dominated by high frequency edge information, and this information is averaged over about 1.5–2 degrees. The weightings have been median filtered to remove salt and pepper noise and to ease interpretation. X and Y axes indicate degrees from point of fixation.

non-fixated locations. It also seems counterintuitive: surely extreme contrast is usually informative? A recent experiment (Ludwig, Gilchrist, McSorley, & Baddeley, 2005) has confirmed that it is not that we do not typically fixate high contrast targets, rather that we cannot, even if it is required for the task.

In Ludwig et al.'s (2005) study, participants were presented two Gaussian blobs that varied randomly in their luminance and hence also in their contrast with the background. In this condition, subjects could quickly and easily fixate the brighter (higher contrast) blob. If instead the two blobs had the same average luminance for the first 100 ms, and only differed afterwards, participants' performance dropped to chance. From this data it is clear that it is only the luminance or contrast onset that subjects have available. Since our experiments deal with the viewing of static scenes, and the first fixation after the images are displayed is thrown away, there is no onset information. We are currently exploring eye movement behaviour to moving images, and in this case we believe the rapid (less than 100ms) contrast and luminance onsets will be highly predictive of fixation locations. Again Einhäuser and König (2003) found that changing the contrast in natural images did not affect the pattern of eye movements.

4.1. What the identified system tells us about "salience"

Having the "salience" system invariant to high frequency contrast also makes ecological sense. Given the massive effect of variable illumination, simple contrast differences will not be informative about surface properties. Whilst a sharp edge is unlikely to be due to chance illumination variations, the extremes of contrast in images are often due to either light sources or specular highlights, and hence are not usually informative locations to fixate.

Although high frequency edges are the largest contributor, two other maps make significant, but inhibitory, contributions to predicting the probability of fixation, and illumination may again provide insight into this. The inhibitory contribution from low frequency edges is only

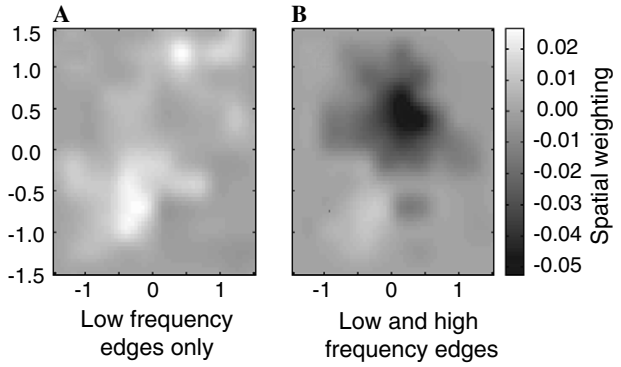

Fig. 5. The importance of doing multivariate statistics: (A) the spatial weighting function of for a low frequency edge map when this is used as the only source of information about the fixation probability. From this result, it appears that low frequency edge information is facilitatory, and this is the conclusion arrived at by all previous studies, since they tested mapped each feature map individually. (B) This is false. When the mapping is optimised with both low and high frequency edge information (the best model using only two feature maps), low frequency edges *inhibit* fixations. Thus conclusions opposite to the true situation have been arrived at, simply because inappropriate single map measures have been used. Axes are again in degrees from point of fixation.

observed when optimised in conjunction with high frequency edges, and image locations that show high power at both low and high frequency will tend to be blurred edges. At least in overcast or hazy conditions, edges due to shadows tend to be more blurred than those due to surface reflectance changes, and therefore signal that the edge is probably less informative about surface properties. The low frequency luminance channel is inhibitory, both when optimised in conjunction with the other maps, and when optimised on its own. Extremes of low frequency luminance tend to either be due to light sources (the sun, or specular highlights), or regions where shadow means no visual information can be extracted. Again, for a system interested in surface properties rather than illumination-based features, a strategy of avoiding extremes make sense.

The system identification not only estimates the relative importance of the different image features, but also how they are spatially integrated. All spatial weighting functions appear to be noisy local averages over an area of about two degrees, essentially the size of the fovea. Eye movements are made to place the object of interest somewhere in the fovea, but not always exactly in the centre. The fact that we found no evidence for centre surround inhibition argues against, but does not rule out surround inhibition as a mechanism. There are a number of reasons why this could be so: first, our results are averaged over saccades of a wide range of magnitudes (up to 20 degrees). If the surround inhibition was both weak, and varied as a function of the eccentricity of a fixation (Tatler, Baddeley, & Vincent, 2006), then by averaging over fixations of all distances, the inhibition could have been averaged out. Second, the eye movement system is inaccurate, with accuracy proportional to the size of saccade (Van Opstal & van Gisbergen, 1989). Again, this could lead to an averaging out of any (small) inhibitory surround. Lastly, the strongest regularity (high frequency edges), and hence the regularity within which we would most expect to find any evidence for centre surround inhibition, is formed by combining edge detectors of four different orientations. Conceivably there could be centre surround contrast enhancement operating with each orientation that is concealed by taking the maximum response of the four orientations. Nonetheless, if within feature map centre surround contrast enhancement is a feature of ''salience'' maps, it appears not to be a strong or robust one.

4.2. The system identification method

We have used participants' responses to natural images to characterise the salience system, and the method has strong similarities to the highly used reverse correlation method (de Boer, 1967; de Boer & de Jongh, 1978; Lee & Schetzen, 1965). The main differences are: (1) we do not estimate using linear regression but instead use logistic regression, and (2) we place a flexible prior on the regression parameters rather than using either no constraint, or using a constraint on the parameters that strongly biases the identified system to global solutions.

Let us consider why we estimate using logistic regression rather than linear regression. The linear regression approach used in reverse correlation is a sensible approach to understanding the relationship between the world (e.g., natural stimuli) and participants' discrete responses in two different situations. First, if the mapping that is estimated is that between response and stimulus (the reason for reverse correlation's name), the input statistics can be designed to have Gaussian statistics (white noise for instance), and the assumptions of linear regression are not violated (as they would if for instance natural images were used). Though this is *not* estimating a receptive field, this mapping between response and stimulus can be very useful for many purposes. It can estimate the amount of information transmitted by a representation since information is symmetric (Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997). This is also very useful for comparison to Bayesian ideal observers.

A second situation in which reverse correlation can be used is if not only the linear component is estimated, but enough higher order moments for the (Taylor) series to start to converge. The mapping cannot be linear, but the non-linearity may be approximated by a Taylor series expansion. Unfortunately, if the dimensionality of the input has more than two or three dimensions, this requires impractical amounts of data. For many problems, the Taylor series may also not converge. Note that only in the second situation can the results of reverse correlation be thought of as a (linear) approximation to the receptive field.

A further limitation of reverse correlation for characterising salience at fixation arises because we do not have a good characterisation of the input, in this case natural

images. In contrast to white noise analysis where we know the statistical distribution of the signal, in our case we have a poor characterisation of the signal (the distribution of filtered natural images). The (binary) responses are very likely to be Bernoulli distributed and, for regression, it is the distribution of the dependent variable which is important. Performing logistic regression also has the advantage that it is an estimate of the receptive field of the system, and not simply something that is related to it only if a large number of assumptions about the system are true. The downsides of using a generalised linear model are large. Linear regression has a closed form solution, and efficient online versions exist that require essentially no memory. In contrast, for logistic regression, only iterative algorithms exist, and convergence can be very slow. Despite this, if one is interested in mappings between poorly characterised inputs (e.g., natural images), and participants' responses (well characterised Bernoulli responses), one wants a characterisation that can be thought of as a receptive field, and one has a reasonably fast computer, then a generalised linear model is the way forward.

The second way that our system analysis differs from previous approaches is that we estimate the sparsity of the mapping. This is important for two reasons: first, previous methods of constraining the mapping (say regressing against a Fourier representation, or Principal components) can greatly distort the estimated mapping unless unrealistically sized data sets are collected, and second, as long as the mapping is sparse (and most biological mappings will be), placing a sparse constraint means that the number of training examples only grows as a logarithmic function of the number of irrelevant features, rather then a linear function for unconstrained mappings (or any other rotationally invariant constraint (Ng, 2004)). This second point is by far the most important. Perhaps the most important constraint on the application of system identification techniques to understanding biological processes is the amount of data required to make an effective characterisation. What Ng (2004) showed is that as long as the mapping is based only on a small (but unknown) subset of the inputs then, by placing a non-rotationally invariant prior on the mapping, the amount of data required to identify this mapping can be orders of magnitude smaller than for either an unconstrained or rotationally invariant constraint. In fact, for a given data set it may be that a non-sparse mapping is present, and in this case our approach offers no real benefit. However, for most biological systems the mapping will consist of a few important inputs out of a potentially very large number. In this case the number of problems that can be approached is greatly increased by the more realistic data requirements of sparse priors.

4.3. Does this argue for a salience model?

We have shown that statistical features of the input can predict whether given locations are fixated or not. Such demonstrations have often been argued to provide evidence for a low-level feature map account, but correlation in no way implies causation. It could simply be that eye movements are directed to the boundaries of objects, and edges tend to occur there. Our work, and work of this type, cannot distinguish between the direct causation of image statistics implied by a salience approach, and the indirect role in a high-level vision approach. That is, the preferential fixation of high frequency edges might arise because the system is driven by the presence of this feature, or because the system is driven by a high-level constraint that happens to result in fixating locations of higher than average edge content (e.g., objects). Nonetheless, we have shown that: (1) it is possible to estimate very high dimensional mapping between inputs and participants' responses (in this case eye movements); (2) carrying out the mapping we investigated, using a method that takes correlations into account produces very different answers from methods of analysis that do not; and (3) the results obtained provides a useful first approximation to what is special about fixated locations: they have high frequency edges.

Acknowledgment

This research was partly funded by the EPSRC sponsored REVERB (reverse engineering the vertebrate brain project).

References

de Boer, E. (1967). Correlation studies applied to the frequency resolution of the cochlea. *Journal of Auditory Research, 7*, 209–217.

de Boer, E., & de Jongh, H. (1978). On cochlear encoding: potentialities and limitations of the reverse-correlation technique. *Journal of the Acoustical Society of America, 63*, 115–135.

Baddeley, R. J. (1997). The correlational structure of natural images and the calibration of spatial representations. *Cognitive Science, 21*(3), 351–372.

Bishop, C. M. (1996). *Neural networks for pattern recognition*. New York: Oxford University Press.

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

Elder, J. H., & Goldberg, R. M. (2002). Ecological statistics of Gestalt laws for the perceptual organization of contours. *Journal of Vision, 2*(4(5)), 324–353.

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

Frank, I. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools—response. *Technometrics, 35*(2), 143–148.

Fu, W. J. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics, 7*(3), 397–416.

Hastie, T., Tibshirani, R. J., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer-Verlag.

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

Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision, 45*(2), 83–105.

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

Lee, Y., & Schetzen, M. (1965). Measurement of the Wiener kernels of a non-linear system by crosscorrelation. *International Journal of Control, 2*, 237–254.

Ludwig, C. J. H., Gilchrist, I. D., McSorley, E., & Baddeley, R. J. (2005). The temporal impulse response underlying saccadic decisions. *The Journal of Neuroscience, 25*(43), 9907–9912.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman and Hall.

Morrone, M. C., & Burr, D. C. (1988). Feature detection in human vision: a phase-dependent energy model. *Proc. Roy. Soc. B, 235*, 221–245.

Nelson, J. D., Cottrell, G. W., Movellan, J. R., & Sereno, M. I. (2004). Yarbus lives: a foveated exploration of saccadic eye movement. *Journal of Vision, 4*(8), 741a.

Ng, A. Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. In: Proceeding of the 21st international conference on machine learning, Banff, Canada.

Van Opstal, A. J., & van Gisbergen, J. A. M. (1989). Scatter in the metrics of saccades and properties of the collicular motor map. *Vision Research, 29*(9), 1183–1989.

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision, 16*(2), 125–154.

Parkhurst, D. J., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*(1), 107–123.

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

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge, Mass: MIT Press.

Renninger, L. W., Coughlan, J., & Vergheese, P. (2005). An information maximization model of eye movements. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.). *Advances in neural information processing systems* (Vol. 17). Cambridge, MA: MIT Press.

Sen, K., Theunissen, F. E., & Doupe, A. J. (2001). Feature analysis of natural sounds in the songbird auditory forebrain. *Journal of Neurophysiology, 86*(3), 1445–1458.

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*, 1857–1862.

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