Available online at www.sciencedirect.com

SCIENCE DIRECT®

Vision Research 45 (2005) 643–659

Vision Research

www.elsevier.com/locate/visres

Visual correlates of fixation selection: effects of scale and time

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

[a] *Sussex Centre for Neuroscience, School of Life Sciences, University of Sussex, Brighton BN1 9QG, United Kingdom*
[b] *Department of Experimental Psychology, University of Bristol, 8 Woodland Road, Bristol BS8 1TN, United Kingdom*

Received 15 June 2003; received in revised form 9 November 2003

Abstract

What distinguishes the locations that we fixate from those that we do not? To answer this question we recorded eye movements while observers viewed natural scenes, and recorded image characteristics centred at the locations that observers fixated. To investigate potential differences in the visual characteristics of fixated versus non-fixated locations, these images were transformed to make intensity, contrast, colour, and edge content explicit. Signal detection and information theoretic techniques were then used to compare fixated regions to those that were not. The presence of contrast and edge information was more strongly discriminatory than luminance or chromaticity. Fixated locations tended to be more distinctive in the high spatial frequencies. Extremes of low frequency luminance information were avoided. With prolonged viewing, consistency in fixation locations between observers decreased. In contrast to [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] we found no change in the involvement of image features over time. We attribute this difference in our results to a systematic bias in their metric. We propose that saccade target selection involves an unchanging intermediate level representation of the scene but that the high-level interpretation of this representation changes over time.

Keywords: Saccadic selection; Image features; Spatial scale; Time course; Intermediate representation

1. Introduction

The way that our visual system samples world is both temporally and spatially constrained; sampling takes place during periods of fixation that typically occur at a frequency of 3–4 per second and is spatially constrained by sampling limits imposed by the retina. Given these constraints the visual system is unable to sample completely and uniformly the complex visual environment. Indeed, it is clear that during activities of daily life there are large proportions of the visual surroundings that we do not direct our eyes toward (e.g. Ballard et al., 1992; Land & Hayhoe, 2001; Land, Mennie, & Rusted, 1999). When viewing paintings and images, visual complexity is greatly reduced; the scene is constrained to two dimensions and spatially limited to a relatively small proportion of the observer's field of view. However, even under these conditions sampling is not complete or uniform, with some regions of the scenes receiving many more fixations than others (Buswell, 1935).

What are the processes that underlie this non-uniform sampling of the environment? Most researchers would argue that eye movement targeting involves a combination of bottom up and top down guidance factors. Some emphasise bottom up processes: implying that the most important factor in non-uniform sampling is the non-uniform distribution of "salience" in the world (e.g. Braun & Sagi, 1990; Kowler, Anderson, Dosher, & Blaser, 1995; Nakayama & Mackeben, 1989). The activity in low-level feature maps has been

[*] Corresponding author. Address: Department of Psychology, University of Dundee, Dundee DD1 4HN, United Kingdom. Tel.: +44 1382 348260; fax: +44 1382 229993.
E-mail address: b.w.tatler@dundee.ac.uk (B.W. Tatler).

0042-6989/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2004.09.017

proposed to underlie saccade targeting (Itti & Koch, 2000; Itti, Koch, & Niebur, 1998; Niebur & Koch, 1996; Olshausen, Anderson, & Vanessen, 1993; Parkhurst, Law, & Niebur, 2002; Treisman, 1988; Wolfe & Gancarz, 1996). As evidence for the contribution of salience, the differences between the image statistics of fixated and non-fixated locations in scenes are emphasised; for example, Reinagel and Zador (1999) showed that fixated locations have higher contrast than non-fixated locations.

Other researchers emphasise the contribution of top down processes: implying that the non-uniform sampling is due mainly to high-level task demands. Pelz and Canosa (2001) suggested that "look ahead" fixations (checking objects that will be manipulated several seconds in the future) provide strong evidence that at least these types of eye movements are not salience driven, but rather are task dependent and driven by top down control. Shinoda, Hayhoe, and Shrivastava (2001) similarly stressed the importance of top down control, finding that detection of traffic signs in a driving simulator was modulated by visual scene context and task instructions.

While evidence that fixated and non-fixated locations differ in their statistics may be seen initially as evidence for the relative importance of low-level salience, this may not be the case. A predominantly top down selection mechanism may also result in non-random selection of low-level features. Most tasks require fixations on a specific set of objects and these objects tend to be distinguished by differences in luminance, colour, contrast and the occurrence of edges. Under this view, differences in image statistics at fixation could be an artefactual result of people fixating objects, which tend to differ from the background. Therefore, simply looking at the statistics at fixated and non-fixated locations cannot differentiate high- and low-level accounts.

One possible source of evidence is to investigate whether any quantifiable characteristics of eye movements change over viewing time. Both Buswell (1935) and Yarbus (1967) found that over time, the consistency between observers in where they fixated decreased. While this was primarily a qualitative observation, if confirmed quantitatively, it could place constraints on the interaction between top down and bottom up processes. Specifically, in the current study we measure not only the consistency of fixation locations, but also the inferred salience at these locations over time. This allows four possible frameworks to be distinguished. We call these four frameworks (1) *salience divergence*, (2) *salience rank*, (3) *random selection with distance weighting* and (4) *strategic divergence*.

The *salience divergence* model proposes that the balance between top down and bottom up control of saccade target selection changes over time. Specifically, the bottom up component is more influential early in viewing, but becomes less so as viewing progresses; this was suggested by Parkhurst et al. (2002). Such a framework could account for an observed decrease in between-participant consistency over time. In addition to a decrease in consistency, this framework predicts that the difference between saliency at fixated locations and at non-fixated locations will be greatest early in viewing.

A second possibility is that there is no change in either the top down or bottom up components of saccadic targeting over time. In the *salience rank* model, locations in the scene are ranked according to their visual salience and the oculomotor system selects targets sequentially according to this ranking; Itti and Koch's model uses a system for selecting successive targets for attention based upon decreasing salience (Itti & Koch, 2000). In any scene it is likely that there will be few locations of high salience, many of medium salience and even more of low salience, if salience is simply related to the output of filters (Field, 1987). Therefore the *salience rank* model predicts a decrease in consistency between participants, and a decrease in the salience of fixated locations over time.

The *random selection with distance weighting* framework for target selection (Melcher & Kowler, 2001) suggests that targets are selected using a proximity-weighted random walk process. This proposes that fixation locations are essentially random with respect to both bottom up and top down processes. The *random selection with distance weighting* proposal predicts that given a common starting location, the between-observer consistency of saccades will decrease over time, but that there should be no systematic change in the visual saliency at fixation.

A fourth possibility is *strategic divergence*. Here the influence of low-level visual feature salience on saccadic targeting does not change during viewing. Instead, the *strategic divergence* account proposes that the strategies chosen by observers have the same bottom up frame of reference for eye movements, but over time observers use different top down strategies. This could predict an increase in the variability of fixation locations, but no change in the saliency at fixation over time.

As can be seen, the four models predict both an increase in between-observer variability over time and different patterns of change in salience over time. We therefore quantified changes in the between-observer consistency in fixation locations as a function of viewing time. Explicitly, we estimated the probability distribution of fixation locations for individual observers. We then used an information theoretic measure (Kullback–Leiber divergence) to quantify the differences between these probability distributions. This quantity was estimated both as a function of fixation number and viewing time.

In order to quantify any difference in the visual saliency of fixated and non-fixated locations, we extracted image features at fixation and compared these to image features at non-fixated locations. The first regularity that we explored was whether there are simple differences in luminance at fixated locations. It could be that eye movements are attracted to extremes of luminance, or potentially, because the brightest regions are often highlights and the darkest are often uninformative shadows, eye movements may avoid such extremes of luminance. The other three image features investigated were based on a subset of properties represented early on in the visual system. Retinal ganglion cells make explicit both contrast (the output of centre-surround receptive fields) and chromaticity, and both may be relevant in determining saccade target locations. In V1 a much greater range of characteristics are made explicit, but the majority of receptive fields are well characterised by Gabors (making orientated edges explicit). Other features are also made explicit such as stereo, motion, and potentially orientation contrast, but for the purposes of this study we concentrate on a representation of orientated edges.

We defined a signal detection measure for characterising the visual salience of fixated locations for each of our four image features. Essentially, this measure quantifies the visual salience difference in terms of how reliably fixated and non-fixated locations can be discriminated based upon the underlying salience measure.

There already exist in the literature techniques for quantifying both between-observer consistency (Mannan, Ruddock, & Wooding, 1995, 1996, 1997) and saliency at fixation (e.g. Parkhurst et al., 2002; Parkhurst & Niebur, 2003; Reinagel & Zador, 1999). Unfortunately previous methods have a number of limitations. Nearest saccade-based measures of consistency, such as those used by Mannan and colleagues, have problems in that they have to exactly specify a function relating distance and similarity; they are insensitive to differences in the probability distribution; and they confound within- and between-observer variability. These limitations are discussed in more detail in the methods section.

Measures of salience at fixation, such as employed by Parkhurst and colleagues, quantify visual salience using a significance test of the difference between statistics at fixated and non-fixated locations. This method also suffers from a number of limitations. First, significance and effect size are confounded: a behaviourally insignificant effect can be highly statistically significant given enough data. Second, a measure based on parametric statistics makes assumptions about the normality of image statistics; image statistics tend not to be normally distributed (Baddeley, 1996). Lastly, in defining measures of visual salience, often arbitrary decisions need to be made about such things as non-linearities. Parametric tests are highly dependent on such arbitrary decisions. In the methods section we identify a further confound, which arises due to non-spatially uniform distribution of saliency in natural scenes. This is important because it could artefactually indicate a change in saliency over time.

The present study describes two measures without the above confounds. These are used to assess between-observer variability as a function of time and the effect of viewing time on saliency at fixated locations. Our results are used to place constraints on possible models of eye movement control.

2. Methods

2.1. Participants

Fourteen participants took part in this experiment. All had normal or corrected to normal vision and had never previously participated in eye movement experiments.

2.2. Images

Forty-eight images of natural scenes were used in the experiment, covering a variety of indoor and outdoor scenes. Images were recorded using a handheld Fujifilm MX-1500 digital camera and were displayed in 800×600 pixel format with 8-bit representation of red, green and blue (a 24-bit image). The images were displayed on a 17″ SVGA colour monitor with a refresh rate of 74 Hz and a maximum luminance of 55 cd m^{-2}. The experiment was carried out in a darkened room. The monitor was positioned at a viewing distance of 60 cm; consequently, the images presented subtended 30° horizontally and 22° vertically.

2.3. Procedure

Each trial began with a central fixation box [1] on a mid-grey background followed by display of one natural image for a period that varied randomly between 1 and 10 s, after which the display returned to the mid-grey of the initial background. Presentation times were varied to reduce predictability and prevent the employment of unnatural strategies, such as systematically working through the image. Images were blocked into three sets and the order of these image-sets was varied systematically between participants, to minimise any potential order effects on fixation patterns. Given any top down effects, different strategies can result in different viewing patterns. During free viewing, the number of strategies is effectively uncontrolled, with different observers employing different strategies and effectively performing different tasks. In order to minimise this variability, we

[1] A second experiment was conducted to validate our results, in which the position of the fixation box was varied randomly between trials. See results section for details of this experiment.

asked observers to perform a memory task while viewing. Performing such a memory task does not reduce high-level strategic factors, but is likely to promote the employment of broadly similar high-level strategies between participants. Following each image presentation, participants were asked two questions about the image just viewed. Questions covered a range of possible aspects of the scene: what the image depicted, whether items were present or not and a range of details about objects including absolute position, relative position, colour and shape. Objects tested varied in size from 0.04% to 32% of the total screen area. The position of the item tested was also varied between questions and scenes. Though not through design, this distribution was slightly centrally weighted. Responses to these questions were not used in the analyses.

2.4. Eye movement recording

Eye movements were recorded using an EyeLink I eye tracker, which uses infrared pupil tracking to sample eye position data at 250 Hz and compensates for head movement. Eye position data were collected binocularly and analysed for the eye that produced the better spatial accuracy. A 9-point target display was used for calibration of eye position. A second 9-point display was used to validate the calibration and return the mean spatial accuracy of the eye tracker calibration. Further 9-point validations of the calibration were carried out at regular intervals throughout the experiment. If the validation showed that the spatial accuracy of the eye tracker had deteriorated to worse than ±1°, the eye tracker was re-calibrated as described above. In this study, the mean spatial accuracy of the eye tracker calibration was 0.40°, with a standard deviation of 0.10°.

Analysis of the eye movement record was carried out off-line after completion of the experiments. The timings of eye movement and display events were extracted from the raw data record along with the co-ordinates of saccade and fixation start- and end-points. Extraction was carried out using software supplied with the EyeLink I eye tracking system. Saccade detection required a deflection of greater than 0.1°, with a minimum velocity of $35° \mathrm{s}^{-1}$ and a minimum acceleration of $9500° \mathrm{s}^{-2}$, maintained for at least 4 ms. The extracted event data was used in all subsequent MATLAB-based analysis protocols, which were written specifically for these analyses. Trials were discarded if the eye tracker set up resulted in a spatial accuracy poorer than ±1°. As a result all trials for one participant were discarded, but no other trials failed to meet this criterion, leaving 624 usable trials for analysis.

2.5. Feature modelling

In order to assess quantitatively the extent to which image properties are selected by the eye movement system, four candidate image features were chosen: luminance, chromaticity, contrast and edge-content. The models used to construct the feature maps for each of these properties are described below. The smallest filter used for modelling was 10.8 cycles per degree; this was the highest frequency that could be displayed reliably, given that the images were presented at 26 pixels per degree. Feature maps were constructed at thirteen spatial scales for each of the four image characteristics, ranging from 0.42 to 10.8 cycles per degree (for edge information, the spatial scale refers to the peak of the Gabor carrier).

The processes involved in the construction of the salience maps for each of the features are illustrated in Fig. 1. The first step in the construction of feature maps for luminance, contrast and edge-content was to convert the colour bitmap viewed by participants into a greyscale version of the image using the built-in MATLAB conversion.

For all feature maps, it was important to minimise any edge effects at the image boundaries (which would result in false "activation" of the filters when they overlapped the edges of the image) in later stages of image filtering; we found that the best method to achieve this was as follows. At the start and end of each row or column in the (greyscale) image, the intensities of the five pixels closest to the edge were averaged and this mean intensity was streamed out from the end of the column or row. At the corners, pixel intensity was calculated by the nearest neighbouring pixel intensities. [2] Using this streaming technique the images were extended by eight times the standard deviation of the filter used in the subsequent convolutions (see below) in all directions. After convolution, the images were cropped to the original image size, removing the extended margins.

About the best model of receptor non-linearities is the Naka–Rushton equation (Valeton & Vannorren, 1983). While this equation models saturation at high and low light levels, it is essentially well-summarised as a logarithmic relationship over three orders of magnitude; we therefore log-transformed the greyscale (and now extended) image. The extended and transformed images were convolved using filters specific to the feature map under construction. Luminance information was extracted by convolving the images with a Gaussian filter as described in Eq. (1), where x and y specify the co-ordinates of each pixel in the image.

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

[2] We used this streaming technique because we found that the more traditional methods of zero-padding and, to a lesser extent, flipping resulted in serious edge artefacts. This streaming methods greatly reduced these edge effects.

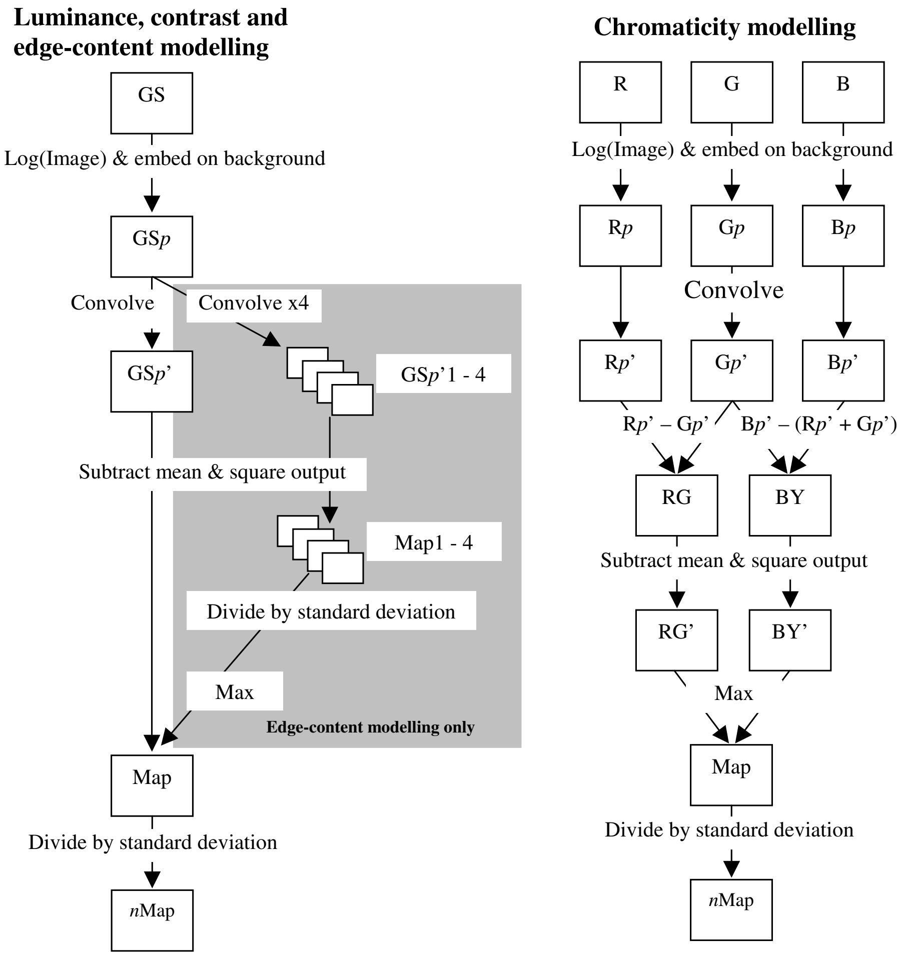
Luminance, contrast and edge-content modelling
GS
Log(Image) & embed on background
GSp
Convolve
Convolve x4
GSp'
GSp'1 - 4
Subtract mean & square output
Map1 - 4
Divide by standard deviation
Max
Edge-content modelling only
Map
Divide by standard deviation
nMap
Chromaticity modelling
R
G
B
Log(Image) & embed on background
Rp
Gp
Bp
Convolve
Rp'
Gp'
Bp'
Rp' – Gp'
Bp' – (Rp' + Gp')
RG
BY
Subtract mean & square output
RG'
BY'
Max
Map
Divide by standard deviation
nMap

Fig. 1. The processes involved in modelling images for each of the four features (full details in main text). For chromaticity modelling the individual red (R), green (G) and blue (B) channels are separated out; for other features the images are converted to greyscale (GS). The first step was to log transform the image (or channel) and embed it on a mean background. Following this the image was convolved with the relevant filter for the feature in question. For edge-content modelling four filters were used to convolve the image and the outputs of these four filters were later combined and normalised. At this point in the chromaticity modelling, the MacLeod–Boynton opponency maps were calculated (RG and BY). Common to all four feature extraction procedures, the next step was to subtract the mean feature value in the models and square the output. This step served to capture unsigned deviation from mean feature salience in the image. In the luminance, contrast and edge-content models this step produced the raw saliency map (Map). However, in the chromaticity modelling procedure the raw map was produced by combining the two opponency maps. The raw feature map was normalised by dividing by its standard deviation, in order to produce the final feature map (*n*Map). The modelling process was repeated for each feature at each of 13 spatial scales.

For contrast information, convolution was carried out using a difference of Gaussian filter, described in Eq. (2). One critical value is the ratio of surround to centre radius. Lee, Kremers, and Yeh (1998) found an average value of 3.28 in the primate retina, but this value was significantly lower than that found by Croner and Kaplan (1994) who found a ratio of 4.8 for M class cells and 6.7 for P cells. As a compromise we used a ratio of 3.88, which incidentally was the average value for this ratio found in the cat retina (Linsenmeier, Frishman, Jakiela, & Enrothcugell, 1982).

$$f(x,y) = \exp\left(-\frac{x^2+y^2}{2\sigma_1^2}\right) - \exp\left(-\frac{x^2+y^2}{2\sigma_2^2}\right) \qquad (2)$$

The image convolution step for the extraction of edge-content information was carried out using each of four oriented Gabor filters, the outputs of which were normalised (by dividing by the standard deviation) and combined (by finding the maximum value in all four convolutions for each pixel in the image) after convolution. The Gabor filters are described in Eq. (3). θ_1 describes the orientation of the Gabor; four values were used: 0, $-\pi/4$, $\pi/2$, and $\pi/4$. The frequency of the carrier is defined by θ_2 and was set at 0.4σ (i.e. 0.4× standard deviation of the Gaussian component). All of the parameters in our Gabor filters were chosen to be within plausible biological ranges (Daugman, 1985).

$$f(x,y) = \sin\left(\frac{1}{\theta_2}x\sin\theta_1 + y\cos\theta_1\right)\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{3}$$

For each of the features we were interested in capturing unsigned difference from average in our feature maps of the images. Thus our luminance feature maps, for example, were designed to capture both "brightness" and "darkness" in the images rather than simply "brightness". In order to achieve this, the mean feature salience value in each image was subtracted from the convolved image and the output squared.

Different images in the experiment were likely to contain very different ranges of luminance, contrast and edge-content information. In order to allow meaningful comparisons of feature maps between images, therefore, the final step in the construction of the feature map was to normalise the model by dividing by the standard deviation of the output.

There exists a large variety of proposed spaces to represent colour. Important issues include not only how the three receptor types are combined in a colour space but also non-linearities, particularly in the yellow–blue system (Wyszecki & Stiles, 2000). Most of these spaces concentrate on exactly matching perceptual differences (for instance the CIE systems—Glassner, 1995) or just-noticeable differences (Vorobyev & Osorio, 1998). Here we are only interested in whether there are gross chromatic differences at saccaded to locations compared to non-saccaded to locations. We therefore employed a crude approximation to the MacLeod–Boynton colour space (MacLeod & Boynton, 1979). In this space, chromaticity is represented by two channels: the difference between the L and M receptors, and the difference between the S and a combined L and M channel. In our study, rather than using cone fundamentals, we used the RGB system provided by our camera and approximate L with the red channel, M with the green and S with the blue channel. While this is only a crude approximation, any large differences in chromaticity in the MacLeod–Boynton space will also be large differences in our space. Both spaces ignore potentially complicating non-linearities (particularly in the blue–yellow system), however this is less of a concern here as our salience metrics are invariant to monotonic transformations of the colour space. As a result, our RGB space is a reasonable way to explore chromaticity based salience. Our chromaticity filters therefore measure difference from average chromaticity in each image, irrespective of the actual colours.

The processes involved in the construction of the chromaticity feature maps were largely similar to those for the other three features. However, rather than initially converting the image to a greyscale version, the image was separated into its individual red, green and blue (RGB) channels. Each channel was then prepared (extended and log-transformed) and convolved independently, using Gaussian filters of the same form as described in Eq. (1) for the convolution process. Following the MacLeod–Boynton colour space, we subtracted the green channel from the red and the sum of the green and red channels from the blue channel (subtraction because the image was log-transformed), thus producing two opponent maps. As for the other models, the next step was to subtract the mean and square the output in order to capture the maximum unsigned difference in the two opponent channels. The final chromaticity feature map was constructed by combining the two opponent convolutions (using maximum values for each pixel) and normalising the output by dividing by its standard deviation.

Examples of the feature maps constructed for each of the four image features at three of the spatial scales are shown in Fig. 2, for a single image used in this study.

2.6. Measuring the difference in image characteristics between fixated and non-fixated locations

Having constructed the saliency maps for each feature and spatial scale, local image statistics at fixation were extracted from these maps. Local statistics were extracted by centring a box with a diameter of 1° around the centre of each fixation made by participants on the original image (extracted from the EyeLink I eye tracker data). Local statistics were defined as the maximum value of the saliency map within this "foveal" patch for the particular feature and spatial scale. Extraction was carried out for all fixations that began after stimulus onset; fixations beginning prior to onset were at the central fixation point that initiated each trial and so were not analysed. After the images had been processed and the saliencies at fixated locations measured, a method was required to characterise how different the image statistics at these fixated locations were from non-fixated locations. Four issues complicate the problem of characterising the relationship between salience and fixated locations.

First, given the large amount of data collected using our protocol (and in similar studies), even very small and behaviourally irrelevant differences in the image characteristics can be highly statistically significant. Therefore, while it is important to check for the significance of any estimated measure, we need a measure of the differences in image characteristics that also captures the interpretable magnitude of the difference. Such a measure should take into account the variability of fixated locations as well as that of non-fixated locations and should also be independent of the number of data points used in its calculation.

A second complicating issue is that the statistics of natural images, and those of saliency maps derived from them, violate two important conditions for the use of parametric statistics. For reasons described in Baddeley

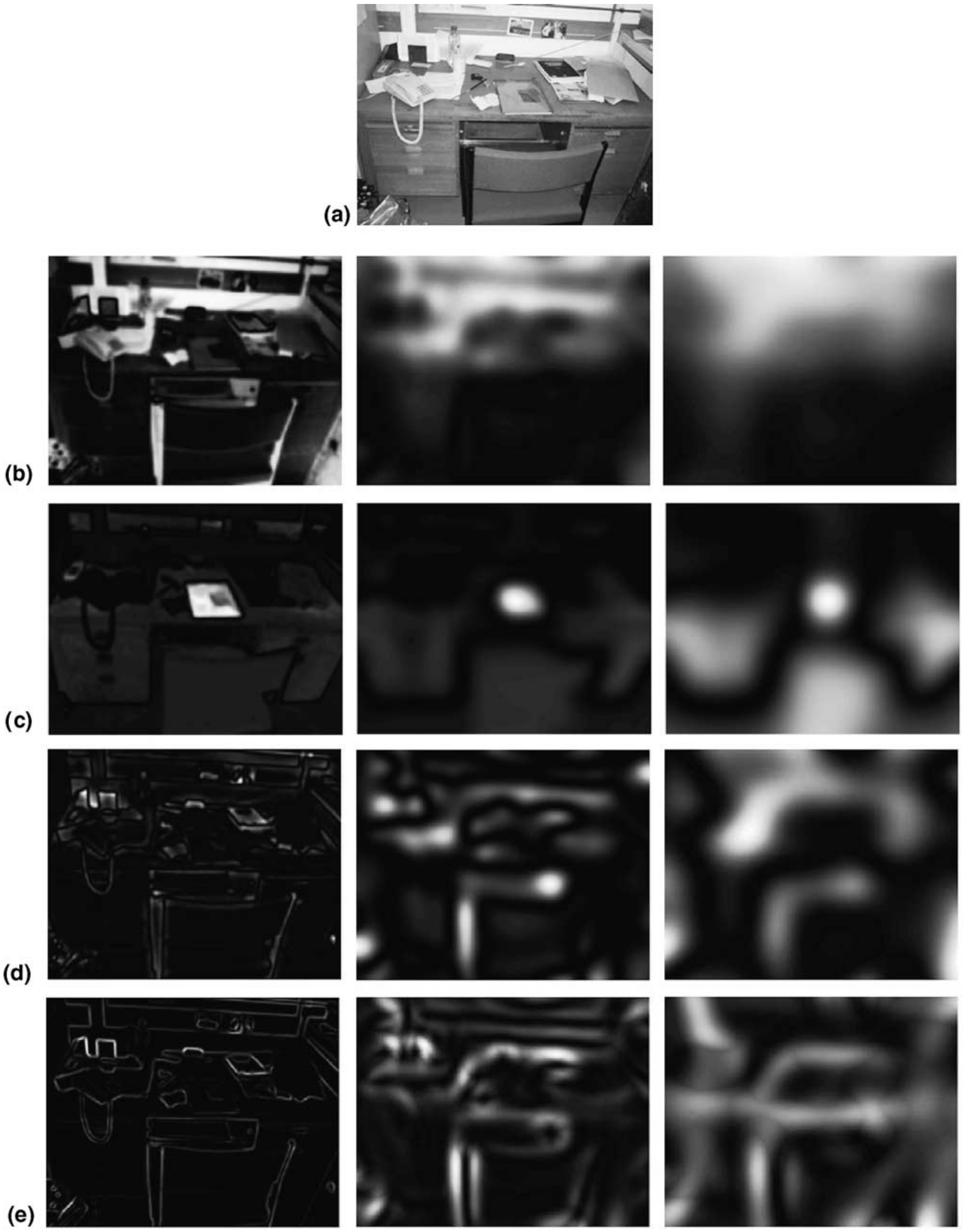

Fig. 2. Examples of feature maps for one of the images viewed by participants. (a) The original image (in greyscale). Feature maps are shown for three of the 13 spatial scales (from fine scale information on the left to coarse scale information on the right) for (b) luminance, (c) chromaticity, (d) contrast and (e) edge-content. Pixel intensity in the feature maps corresponds to the response of the feature filter at that position in the image; hence how much of that feature was present at that location in the image. Feature salience is the squared difference from the mean feature value in the image. For example, intensity in the luminance map specifies how bright or dark each position in the image is with respect to the average luminance in the image; the high spatial frequency filter responds similarly to the dark chair legs and the bright wall.

(1996), many interesting derived characteristics are far from normally distributed, often being Laplace distributed (double sided exponentials). Importantly the variance is also far from spatially homogeneous (the local variance is often approximately Gamma distributed depending on the spatial scale, Baddeley, 1996). Both of these characteristics violate assumptions for ANOVA-based methods in particular, and for parametric methods in general, meaning a non-parametric measure is required.

Third, a related issue is that in generating our salience maps, some arbitrary decisions had to be made. One example is our use of a squared non-linearity. While this is a plausible assumption, equally plausible models based on rectification are possible. If our measure depended critically on the nature of any non-linearities then the result of the analysis would be less reliable. This again argues that the measure must be non-parametric and approximately invariant to monotonic transformation of the salience values.

Lastly, a subtler and therefore more dangerous requirement is that our metric should not be confounded by one particularly subtle bias arising from the interaction between two factors. Most photographs of scenes have a small but reliable bias towards higher salience in the centre than around the edges of the images (Fig. 3a). There are a number of causes of this effect (e.g. sky in the upper visual field, uncluttered ground pane, photographers' tendency to place subjects of "interest" at the centre), but this has been observed even for such simple image features as the power spectra slope (Torralba & Oliva, 2003). On its own this bias would not present a major problem, but there is also a bias toward making early fixations near the centre of an image (Fig. 3b). The central fixation bias may reflect a general tendency for observers to fixate near the centre of scenes, irrespective of salience, or it may be that these two biases are interrelated. However, whether or not these two biases are related, they must be considered and accounted for in any comparison of salience at fixated and non-fixated locations. If centrally biased fixated locations were compared to uniformly sampled non-fixated locations this would result in an artificially high salience (for similar arguments see also Parkhurst & Niebur, 2003; Reinagel & Zador, 1999).

The combination of these two factors has two important effects on measures of salience. First, fixated locations will show higher salience than non-fixated locations, even if salience was irrelevant in selecting these locations. Second, because early fixations show even more of a central bias than later ones, early fixations will have higher salience than later ones, again independent of any real role of salience in saccade target selection. Parkhurst et al. (2002) observed this effect of decreasing salience with fixation number on a scene, but within our data this effect is entirely attributable to the central biases. This problem can be dealt with by correcting any statistical measure for this bias.

Two classes of measures can satisfy the first three of the four constraints discussed above: signal detection

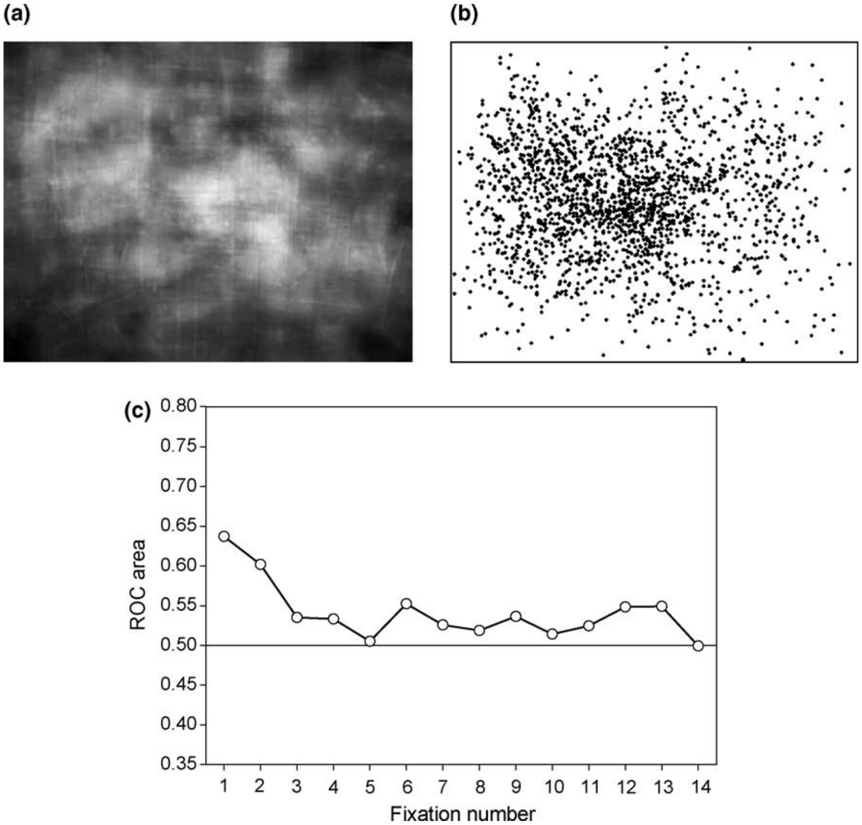

Fig. 3. (a) Contrast feature map averaged across all 48 images and all 13 spatial scales. There was a bias toward contrast salience in the centre of the images. (b) Distribution of fixations that occurred within the first second of viewing, combined across all images and all observers. The distribution was not uniform, but had a central bias. Only fixations that began after stimulus onset were included in this distribution. (c) The artefact produced by selecting non-fixated locations randomly from a uniform distribution. An artificial set of "fixated" patches was generated by taking the fixation locations for a target image but extracting the salience at the corresponding location on an image selected randomly from the remaining image set. Hence the extracted salience did not correspond to anything selected by the eye. This artificial set of statistics was then compared to a set of patches chosen randomly from a uniform distribution on the images, using our signal detection based method (ROC). We can therefore replicate Parkhurst et al.'s (2002) finding of higher salience for early fixations using this artificial dataset and hence show that this effect is an artefact of the uniform sampling of non-fixated locations (cf. our results using non-uniform selection in Fig. 7).

theory or information theory based measures. These two classes essentially differ in the nature of the structure that they would find informative. An information theory based measure of whether salience provides information about whether locations are fixated or not, would be sensitive to any differences between the statistics of fixated and non-fixated locations (for example if they had the same means but different variances). A signal detection measure would only measure those differences that would allow the distributions to be separated by a simple threshold. While both are informative, we propose that a signal detection based measure is a better characterisation of *useable* information and this is the measure we use.

The exact signal detection measure that we chose is the receiver operator curve area (ROC area, Green & Swets, 1966). This metric determines how well fixated and non-fixated locations can be discriminated by their saliencies using a simple threshold. The ROC is a curve that plots the false alarm rate (labelling a non-fixated location as fixated) as a function of the hit rate (labelling fixated locations as fixated). Systematically changing the threshold used to discriminate between fixated and non-fixated locations changes both the hit rate and false alarm rate, and rather than simply choosing an optimal threshold, the ROC area provides a measure summarizing performance across all possible thresholds (the threshold is systematically moved between the minimum and maximum values in the data sets). For two distributions that it is not possible to discriminate, the ROC area will be 0.5. For perfect discrimination, the value will be 1.0, and when the system is predicting worse than chance, the area will be less than 0.5.

This measure is invariant to monotonic transformation of the salience values and takes into account the variability both of saliencies at fixated locations and those at non-fixated locations. Therefore as a measure of strength it has much to commend it. To allow statistical inference to be preformed, we calculate the 99% non-parametric confidence limits of the ROC area by the use of the bootstrap technique (Efron & Tibshirani, 1993). Essentially we created 1000 surrogate data sets of the same size as our original data set. This was done by sampling with replacement from our data, and calculating the ROC areas using these surrogate data sets, and then the distribution of these values was used to calculate the confidence limits.

We are still left with problem of central fixation and salience biases. We approached this problem by not using randomly selected locations to collect the image statistics for non-fixated saliencies, but choosing locations randomly from a distribution of all fixation locations for that observer that occurred at the same time, but on other images (hence corresponding to a location in the current image that was not selected by the observer). This means that both fixated and non-fixated distributions have the same bias and we do not get positive salience differences simply because observers tend to fixate more to the centre of images. This problem we believe generates a number of artefacts in studies such as that of Parkhurst et al. (2002) where they observed an interaction of viewing time and salience. We also observed such an interaction when we failed to correct for central fixation bias (Fig. 3c) but they disappeared when the appropriate correction was used (see results below). In a more recent paper Parkhurst and Niebur (2003) recognised that using appropriately-weighted selection of non-fixated patches produced different results, but they have not applied this technique to an investigation of the time course of selection (such as that conducted in their earlier paper, Parkhurst et al., 2002). Reinagel and Zador (1999) also highlighted the need to bias sampling distributions for selecting non-fixated locations in order to appropriately measure whether salience was selected by the eye. While it is necessary to account for central biases and our approach does this, if the central fixation bias is due to the bias in salience, then our method may underestimate the magnitude of any salience effects.

2.7. Assessing the variability in saccade locations

The second aspect of our data that we wanted to quantify was the consistency between fixation locations: do different observers move their eyes to similar locations, and do eye movements become more variable with time? This specific problem has been investigated previously in a series of papers (Mannan et al., 1995, Mannan, Ruddock, & Wooding, 1996, 1997), but using a measure we believe has a number of limitations. We will now consider briefly these limitations and how they can be overcome by using the measure that we have developed.

The method of Mannan et al. is based upon the sum of squared distances between fixations. Given two collections of eye movements, the difference is calculated by going through every eye movement for observer A, finding the nearest fixation location for observer B, and making a running total of the squared distance to this nearest location. This is repeated for observer B's fixations. After appropriate normalisation, this figure is compared to the value found for "random fixations".

Nearest saccade-based metrics like that used by Mannan et al. and described above are limited by the fact that there is not a natural metric to quantify *how* different two fixation locations are. While two fixations that place the fovea over the same region can reasonably be classed as the same, is a location 20° away twice as different as one 10° away? Should its similarity be scaled according to the cortical magnification factor in the superior colliculus, or is it four times as different, as it would be classified using a sum of squared distance metric? By quantifying difference in terms of the squared distance, the measure is dominated by the extreme fixations. Two essentially identical distributions of fixations, with a single rogue eye movement will be characterised as very different. The most effective solution to this problem is simply to class two fixations that are directed to the same location as the same, and two fixations that are directed to different location as different. Because the fovea subtends 1°–2° we define two locations as the same if they lie within 2° of each other.

The second limitation of nearest saccade-based metrics is illustrated by the simple situation of two observers viewing two objects. If the first observer fixated object A 95% of the time and object B only 5% of the time, while the second observer showed exactly the opposite pattern, it would be reasonable to claim that these two observers displayed different eye movement patterns, and this should be reflected in our measure. However, nearest saccade-based measures would classify these two distributions as identical. A measure based on the probability distribution of locations would not have this problem.

Lastly and most importantly, the nearest saccade-based measure is not simply a measure of similarity because it is confounded by within observer variability. Consider two observers viewing an image; the first observer predominantly fixates the centre of this image whereas the second observer's fixations are distributed randomly throughout the image. A third observer, who predominantly fixates in the top right hand corner, would be classed as more similar to the second observer than the first, even though all three strategies are completely unrelated. Essentially the more evenly distributed the saccades are in one population, the higher the chance that a given saccade will be close to one of them. This effect also leads to the problem that the measure does not scale properly with differing data set sizes. Given an infinite number of fixations for one observer, all locations will have been viewed and a comparison fixation will always be of zero distance to one of them. No amount of normalisation can adequately resolve this problem. For a well-behaved measure, the results should not depend critically on arbitrary decisions or be dominated by outliers. It should be sensitive to differences in distribution as well as simple location, should measure only between observer similarity unconfounded by within observer variability, and its measures should be constant as a function of the amount of data used, simply increasing in accuracy with increasing amounts of data. Nearest saccade-based measures do not satisfy these criteria.

We have developed an information theoretic measure of the difference between fixation distributions that satisfies the above criteria and avoids the limitations of a sum of squared distance measure. Fixation location data were used to estimate their spatial probability distribution, using a binning technique where the bins were chosen to be 2° × 2° squared (a kernel method gave similar results but was slower). As is common with such estimators a prior (corresponding to a Dirichlet prior) was used, implemented by adding a small constant (c in Eq. (4)) to all bins. We used a value of our prior of $c = 10^{-5}$ but given the size of our data set, its value was not critical. The probability of a saccade landing at position x, y was represented as $P(X, Y)$, and the number of saccades occurring in bin X, Y as $F(X, Y)$:

$$P(X,Y) = \frac{(F(X,Y)+c)}{\sum_{X',Y'}(F(X',Y')+c)} \tag{4}$$

Following this an information theoretic measure (the Kullback–Leiber divergence) was used to estimate the difference between two probability distributions. Given two such probability distributions $P_a(X, Y)$ and $P_b(X, Y)$, the Kullback–Leiber divergence is defined as:

$$KL = -P_a(X,Y)\log(P_b(X,Y)) + P_a(X,Y)\log(P_a(X,Y)) \tag{5}$$

The first term on the right hand side is the negative log likelihood of $P_b(X, Y)$ under distribution $P_a(X, Y)$; how probable was the distribution of fixations $P_b(X, Y)$ to be generated under $P_a(X, Y)$. This does measures the difference between the distributions but is confounded by a within observer variability bias. This bias is simply the entropy of $P_a(X, Y)$ (the second right hand side term) and by removing it we make our measure unbiased. The Kullback–Leiber divergence can also be considered to be the number of additional bits of information required to describe distribution $P_a(X, Y)$ given knowledge of $P_b(X, Y)$. Sampling error in this measure is dominated by $P_b(X, Y)$ so rather than compare every fixation distribution to every other distribution, we compared each observer's distribution to the distribution based on all other observers, and averaged over all observers. The pattern of results from this measure is the same as that from comparing every observer to every other observer but the sampling variability is less.

3. Results

3.1. The scale of selection of visual features

We can assess the spatial scale at which the oculomotor system selected image features for fixation by comparing the performance of the 13 different spatial scales of feature modelling (see Section 2). The ROC area statistic reflects the ability to discriminate between fixated and non-fixated regions of the image on the basis of the image feature and spatial scale chosen. By comparing ROC areas for each of the scales, those scales most implicated in selecting fixation locations by the eye can be identified as those for which the ROC area is highest. ROC areas for each of the four image features analysed (luminance, chromaticity, contrast and edge-content) are plotted for each of the 13 spatial scales of salience in Fig. 4. An ROC area of 0.5 indicates that discrimination between fixated and non-fixated regions in the images is at chance. A value greater than 0.5 suggests that the image feature is being selected for fixation. A value below 0.5 suggests that the feature is being avoided.

Generally, ROC areas were higher for the high spatial frequencies than for the lower spatial frequencies. For luminance, ROC areas were below 0.5 for spatial scales coarser than 1.35 cpd. ROC statistics did not fall below 0.5 for any of the other image features.

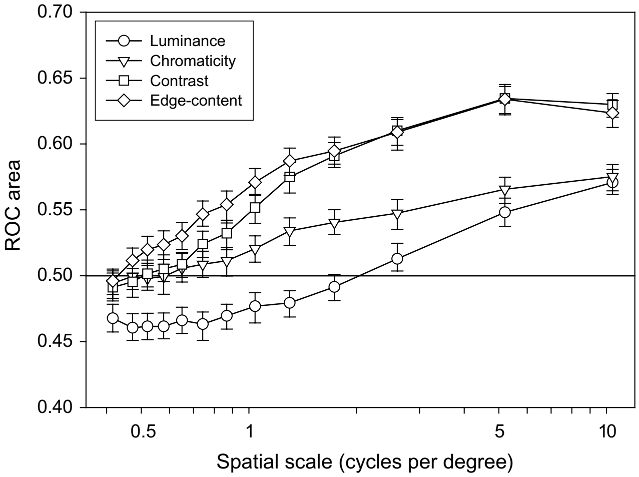

Fig. 4. ROC area values for luminance (○), chromaticity (▽), contrast (□) and edge-content (◇) of fixated locations compared to non-fixated locations as a function of spatial scale. ROC area values measure the difference between the distributions for fixated and non-fixated locations. An ROC area value of 0.5 indicates no difference. The y-scale is much enlarged. Error bars indicate 99% confidence intervals, calculated using a bootstrap technique.

By using ROC rather than ANOVA-based methods we unconfounded the significance and strength of any effect. An extremely strong effect can be non-significant given noisy data and a small effect can be significant given very large datasets (as in the case of our data). In terms of significance, at the highest spatial frequencies discriminability between fixated and non-fixated regions was highly significantly different from chance for all image features ($p < 10^{-9}$). However, while the effect was significant, it was not very strong. For luminance and chromaticity, discrimination of fixated and non-fixated regions was at 57% at its highest (at a spatial scale of 10.8 cpd). For contrast and edge-content, discrimination was at 63% at its highest (at 5.4 cpd). Since ROC captures effect strength, we can use the values in Fig. 4 to effectively rank the spatial scales according to the extent of involvement in target position selection.

3.2. Temporal patterns of fixation target selection

Similar fixation locations were selected by different participants in the first second of viewing (Fig. 5a), but participants selected different fixation targets from each other after several seconds of viewing (Fig. 5b).

3.2.1. Targeting the first few fixations on a scene

Consistency between participants can be assessed using an information theoretic approach. Kullback–Leiber divergence can be used to determine the entropy (hence difference) between probability distributions constructed from the fixation positions of each participant (see methods). The number of bits reflects the degree of difference between the locations targeted by each participant—a higher number of bits indicates a greater difference between participants. Fig. 6 shows the fixation location entropy between participants as a function of fixation number during viewing, for the first 14 fixations after stimulus onset. The actual values of the Kullback–Leiber divergence are not important here because they depend on arbitrary decisions such as the number of

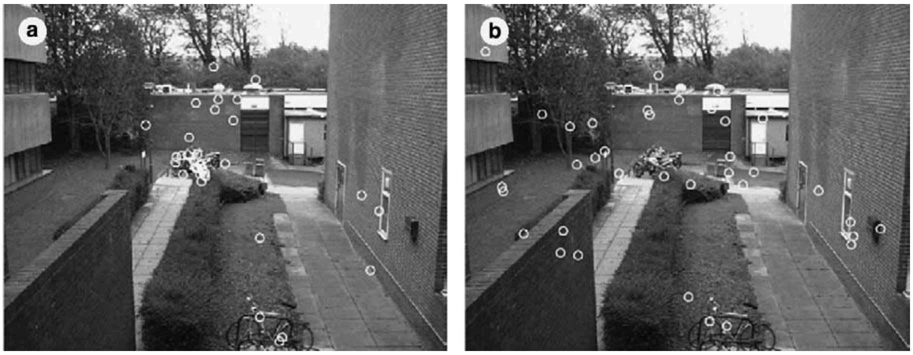

Fig. 5. Fixation locations (indicated by circles) for all observers combined (a) during the first second after stimulus onset, and (b) during the fifth second after stimulus onset, for one of the images viewed. There appears a greater degree of consistency in the locations chosen early in viewing than several seconds later.

participants, the bin size and prior employed in estimating the fixation probability distributions. Rather we are interested in the pattern of change over the 14 fixations.

Difference between participants increased rapidly over the first five fixations on the images but slowed thereafter. One potential problem in our data is that each trial began with a centrally located fixation marker. This common starting point for all participants on each image may in itself account for early central fixation bias (see Fig. 3b) and the greater degree of consistency early in viewing found using our information theoretic measure (Fig. 6a). While studies without a central fixation marker (e.g. Canosa, Pelz, Mennie, & Peak, 2003) still show a central fixation bias, we cannot discount this possible explanation. In order to investigate this issue, we recruited a further four participants to take part in a replication of the main experiment in which the pre-trial fixation marker was randomly positioned before each trial; all other methodological details were identical. The results of this validation experiment are plotted in Fig. 6b. The first fixation shows low consistency between participants, as would be expected given random starting positions, but consistency increases on the second fixation and thereafter follows a very similar pattern of decreasing consistency over several fixations as was found for the experiment in which the pre-trial fixation marker was always central (Fig. 6b cf. Fig. 6a). It would therefore appear that our observed early consistency between participants followed by a decrease in consistency over the next few fixations is not an artefact of the experimental design, but a reflection of the strategies employed by observers when viewing the images.

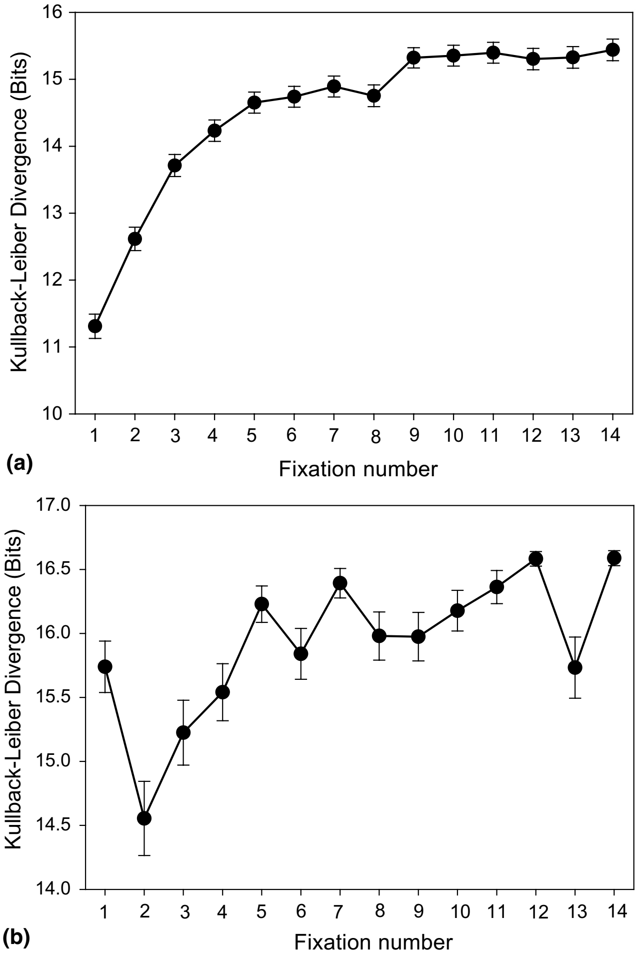

Fig. 6. Mean Kullback–Leiber divergence (±1 SE) in fixation locations between observers as a function of fixation number (a) when the pre-trial fixation marker was always centrally located and therefore all participants shared a common starting point, and (b) when the position of the pre-trial fixation marker was varied randomly. In (a) the difference is least for the first fixation and increases over the following fixations. Fixation location consistency between observers is highest for the first fixation and decreases over the course of several fixations on a scene. In (b) the first fixation shows low consistency (as would be expected given random starting positions, but thereafter the pattern is similar to that shown in (a).

The saliency models can be used to assess whether or not the selection of image features changed over the course of several fixations during viewing. The discriminability between fixated and non-fixated locations is shown in Fig. 7 as a function of fixation number during viewing, for the first 14 fixations by all participants. The magnitude of the ROC value reflects the involvement of the chosen image feature in the selection of the location of each of these fixations. Fig. 7 shows both the data for the main experiment, where the pre-trial fixation marker was always central (Fig. 7a) and for the validation experiment where the pre-trial fixation marker was positioned randomly on each trial (Fig. 7b). As can be seen in Fig. 7, there was no change in the involvement of image features in selecting fixation targets over the course of the first 14 fixations and this was not an artefact of the central pre-trial fixation marker in the main experiment; the curves are essentially flat as a function of fixation number.

3.2.2. *Targeting over the course of prolonged viewing*

Changes in fixation location and selection of image features over the first 14 fixations characterise any changes in targeting by the oculomotor system that occur soon after the onset of viewing. We can extend this investigation to look for any changes over prolonged viewing (up to 9 s). For this analysis we divided the viewing time into nine one-second windows and compared fixation positions using the same information theoretic and signal detection theoretic methods as employed above, but this time over a much extended period of viewing.

Fig. 8 shows the entropy between participants' fixation locations during each of these nine one-second intervals of viewing. The most pronounced increase in entropy (hence decrease in consistency) occurred during the first three seconds of viewing. After this, the differ-

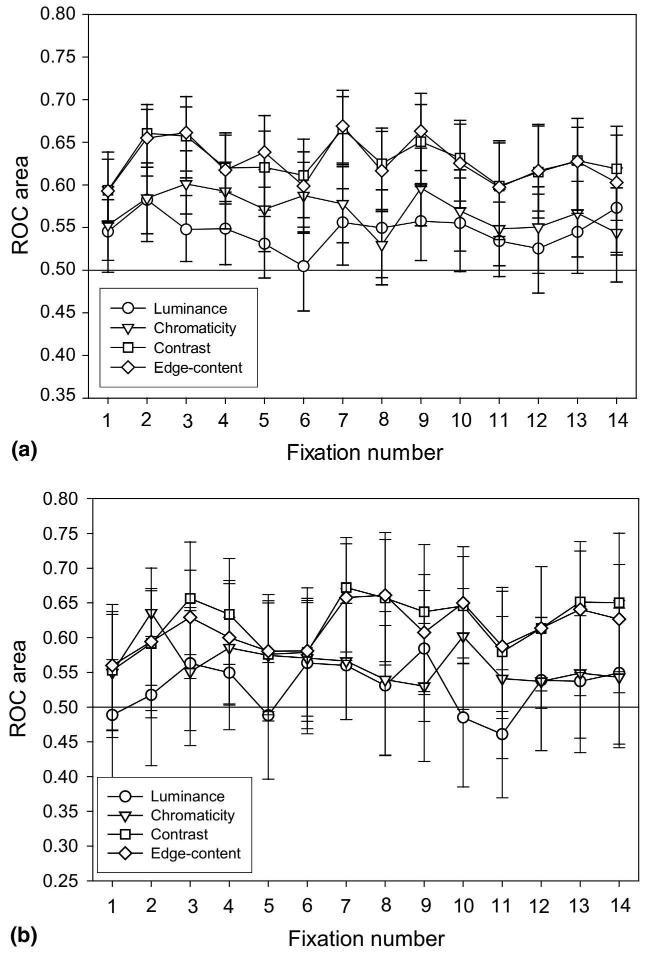

Fig. 7. ROC areas (±99% CI) for each image feature as a function of fixation number (a) when the pre-trial fixation marker was always centrally located and therefore all participants shared a common starting point, and (b) when the position of the pre-trial fixation marker was varied randomly. There is no change in the ROC area with fixation number in either (a) or (b), suggesting that the importance of these image features in selecting fixation targets does not change. ROC areas plotted are for image features extracted at a spatial scale of 5.4 cpd.

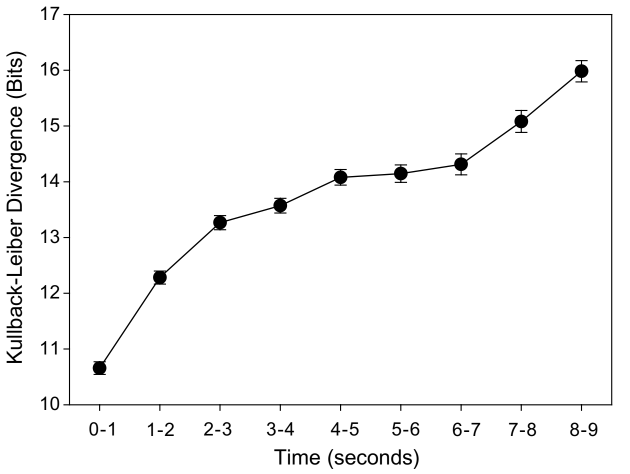

Fig. 8. Mean Kullback–Leiber divergence (±1 SE) in fixation locations between observers as a function of time. Viewing time was divided into nine one-second intervals. The difference between observers is least for the first second and increases over the following seconds. Fixation location consistency between observers is highest for fixations made in the first second of viewing and decreases over the course of several seconds of viewing a scene.

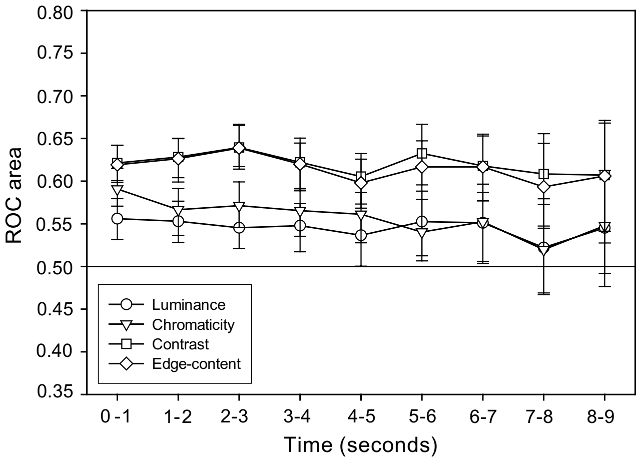

Fig. 9. ROC areas (±99% CI) for each image feature as a function of time. There is no change in the ROC area over time, suggesting that the importance of these image features in selecting fixation targets does not change over the course of several seconds of viewing. ROC areas plotted are for image features extracted at a spatial scale of 5.4 cpd.

ence between participants continued to increase but at a slower rate (although the rate of increase in divergence did appear to increase again toward the end of the viewing period).

The involvement of image features in the selection of fixation target positions over the nine one-second intervals of viewing was characterised by the discriminability between fixated and non-fixated regions extracted from the saliency maps. Fig. 9 shows that there was no change in the selection of image features over the nine second viewing period.

While there was no change in the selection of image features over the course of several seconds of viewing, it is possible that the scale at which selection occurs might vary. One indication of whether or not the scale of features selected by the oculomotor system changes during prolonged viewing is to compare the ROC areas for each of the 13 spatial scales for fixations occurring during the first second of viewing, with ROC areas for fixations occurring during the 9th second of viewing. There was no difference in the pattern of selection of the different spatial scales in the first and ninth seconds (Fig. 10) for any of the four image features. Hence it would appear that the scale of selection of visual features did not change during the time period studied.

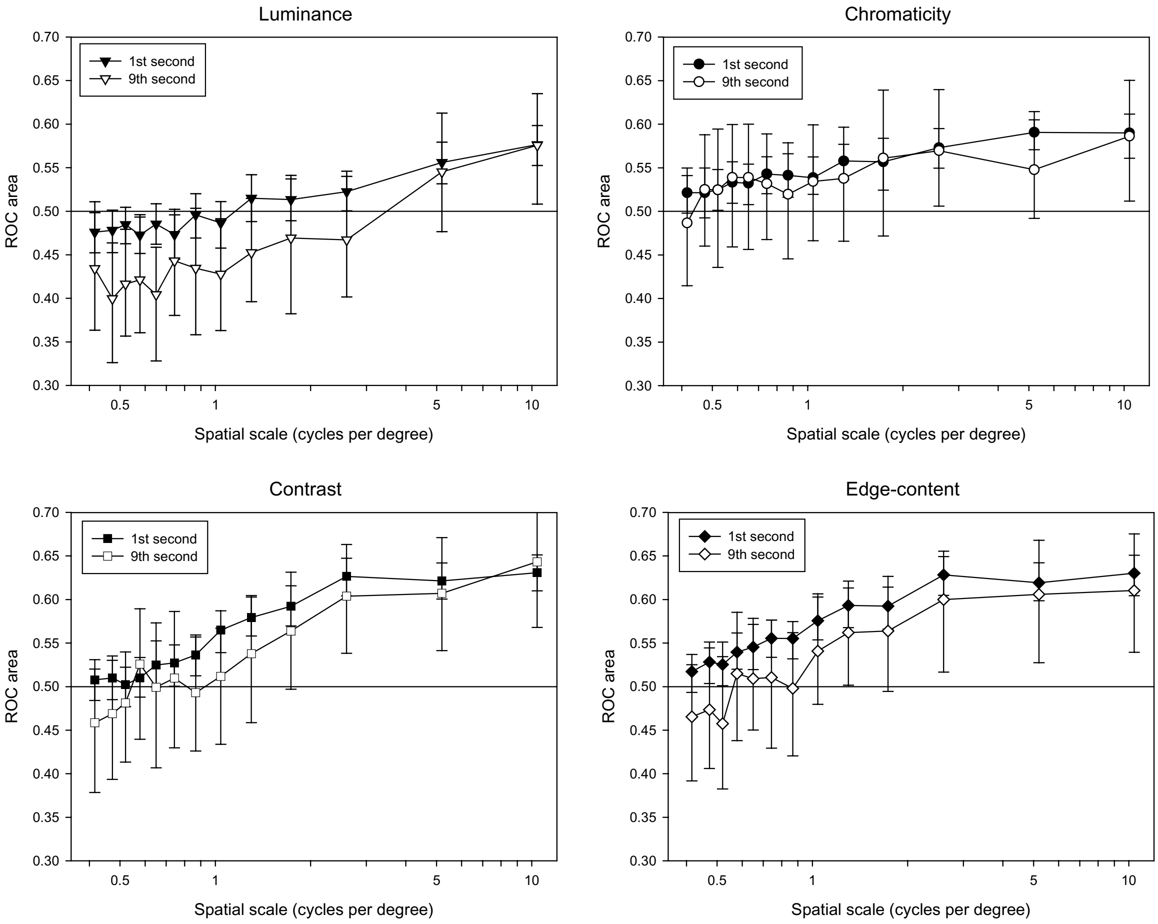

Fig. 10. ROC area (±99% CI) as a function of spatial scale for each feature for fixations occurring during the first (filled symbols) and ninth (open symbols) seconds of viewing. The patterns across spatial scales for each feature are similar for these two time windows and the error bars overlap in almost all cases. We find no evidence for a coarse to fine effect, with the spatial scale profiles being identical at the beginning and end of our viewing period.

4. Discussion

We found that the consistency in fixation locations selected by observers decreased over the course of the first few fixations after stimulus onset; observers were much more consistent in the locations they selected for fixation early in viewing than they were later on. This information theoretic result alone does not reveal the source of the divergence between observers over time. However, in conjunction with the signal detection techniques using the bottom up salience modelling of visual features, we can consider the possible source of change. We can use our findings to evaluate the four models proposed in Section 1.

The *salience divergence* model suggests that the balance between top down and bottom up control of saccade target selection changes over time. Specifically, the bottom up component is more influential early in viewing, but becomes less so as viewing progresses. Such a framework would account for our observed decrease in between-participant consistency over time. In terms of low-level image feature saliency, this framework would predict that the difference between saliency at fixated locations and at non-fixated locations would be greatest soon after viewing began, but decrease thereafter. This pattern of decreasing saliency over a number of fixations on a scene has been reported (Parkhurst et al., 2002). However, the methodological limitations with the technique applied by these authors that we discussed earlier cast doubt on this result. Indeed our image salience measures show that there is no variation in discrimination between saccaded to and non-saccaded to locations over the course of several fixations or even several seconds of viewing. It appears that Parkhurst et al.'s findings are an artefact of their methodology. Because we find no evidence for variation in the discrimination between the salience at fixated and non-fixated locations the current data do not support the *salience divergence* account of saccadic targeting.

In the *salience rank* model locations in the scene are ranked according to the salience of visual features. The oculomotor system then selects targets sequentially according to this ranking (employed by e.g. Itti & Koch, 2000). In any scene it is likely that there will be only a few locations of extremely high salience. These will be selected first during viewing and the limited number of such locations means that there will be a high degree of consistency in the locations selected early in viewing by all observers. Conversely, there are likely to be quite a number of locations with similar, moderately salient characteristics. Hence once the high salience locations have been visited, there exists a much broader range of possible saccade targets. If the oculomotor system selects from among these possible targets at random, then this would give rise to a lower degree of consistency between observers as viewing progresses. The *salience rank* model is therefore able to account for the observed decrease in consistency between participants in where they fixate and does not require any changes to either top down or bottom up selection mechanisms. Once again we can use our signal detection based salience measures to assess this model. Sequential selection of targets based upon visual salience rankings would predict large differences between saliencies at saccaded to locations and those at non-saccaded to locations early in viewing, but smaller differences later on. Our data show that this is not the case; there is no change in the discrimination between the salience at saccaded to and non-saccaded to locations; the current data do not provide support for the *salience rank* model of saccadic targeting.

The *random selection with distance weighting* model of target selection (Melcher & Kowler, 2001) suggests that targets are selected using a proximity-weighted random walk process. Within this model, the selection of locations for fixation is essentially random with respect to both bottom up and top down processes. Such a framework may at first appear to be consistent with our observed data for between-participant consistency and the influence of visual features over time, when considering the results of the main experiment in which all participants commenced their viewing of the images from a common starting point. Constant influence of bottom up features would be expected given random selection with respect to the physical properties of the stimulus. Early consistency between locations fixation by participants, followed by later divergence would also be consistent with a random walk mechanism given common starting points in the image, as in the main experiment reported here. While the data from the main experiment does not allow us to discount the possibility that eye movements on complex natural scenes may be driven by the *random selection with distance weighting* model, the data from the validation experiment can be used in this way. In this validation experiment, starting positions were random and therefore a *random selection with distance weighting* model would not predict inter-participant consistency early in viewing under these conditions. However, we did observe a similar early consistency in this validation experiment and therefore we argue that a *random selection with distance weighting* model cannot fully explain our observations.

The fourth possible framework for saccadic targeting is *strategic divergence*, where the influence of low-level visual feature salience on saccadic targeting does not change during viewing, but the top down strategic component does vary. Within this model, it is this variation that accounts for the observed decrease in consistency between participants in the target locations selected for fixation. This framework is entirely consistent with our findings; fixation location consistency changes between observers over time, but the influence of image features does not. Thus the *strategic divergence* account proposes that the strategies chosen by observers have the same bottom up frame of reference for eye movements, but over time observers use different top down strategies to complete the memory task imposed in this experiment.

The effect of observers' strategies upon the selection of locations for fixation in complex scenes was demonstrated in a classic study by Yarbus (1967). Yarbus showed that fixation locations varied greatly within individual observers when viewing the same painting but with different instructions prior to viewing. This classic experiment demonstrated that top down strategies can have a large influence on the locations saccaded to. We can use Yarbus' finding to explain the observed decrease in consistency between participants in our experiments within the proposed *strategic divergence* framework. Early consistency would reflect similar strategies being selected by observers immediately following stimulus onset. There are presumably many different top down strategies that could be employed to complete a memory task such as was required in our experiment and it may be that there was time for several different strategic approaches to be employed by an observer over the course of several seconds of viewing. Given Yarbus' demonstration that different strategies can produce large differences in the locations that observers fixate, our results could be explained by strategy switching over time, with different observers choosing different strategic interpretations of the task over time.

Given the importance of high-level strategies, it is important to consider the nature of the visual representation that a high-level system would require. Using raw image intensities would be problematic because most tasks do not uniquely define a target in terms of its raw physical properties. However, an intermediate representation, invariant to differences in the world such as illumination, but which allows discrimination of informative from non-informative locations, could be of great utility. Simply because a system is driven largely by top down mechanisms, need not imply that selection cannot be defined in terms of a relatively low-level invariant saliency based representation. Hence the search template used in high-level tasks (e.g. face recognition) may not be encoded in terms of raw image intensities, but may instead be encoded in terms of illumination invariant informative characteristics of the image.

The two features that were most strongly implicated in targeting, particularly for the high frequencies, were contrast and edge-content. These two features are relatively invariant to illumination variation. However, the other two features, luminance and chromaticity, are more prone to illumination confounds. Local luminance will be dominated by illumination and failures of colour constancy will invalidate any system based on colour.

In contrast to previous measures, our signal detection measure unconfounds the significance and the magnitude of effects. Although we replicate the finding that fixated and non-fixated locations are highly significantly different in terms of their salience, the magnitude of the effect is small. Maximum discriminability between fixated and non-fixated positions in the image is 63% for contrast and edge-content and 57% for luminance and chromaticity. This is far from the perfect discrimination (100%) that would be expected if saccades were driven entirely by image statistics. We argue that this demonstrates that the involvement of visual features is perhaps weaker than has been implied in recent salience-based models (e.g. Itti & Koch, 2000; Parkhurst et al., 2002). Rather the magnitude of involvement of features appears consistent with our proposed intermediate invariant representation. It should be noted that our study does not offer an exhaustive survey of image features and also does not consider the interactions between features. It could be that another feature or a complex interaction between features is more discriminatory than those investigated here.

By comparing the performance of salience models at different spatial scales, we are able to assess the relative contribution of different spatial frequencies. It is clear that high spatial frequencies are far more discriminatory than low spatial frequencies for contrast, edge-content and chromaticity. Parkhurst and Niebur (2003) extracted contrast, spatial correlation and spatial frequency content at fixation using sampling patches of different sizes, and compared the statistics of these patches to non-fixated patches. They found greater differences, and hence a stronger involvement of the image features, for small patches, suggesting that finer scale information was more strongly implicated in selection than the coarser scales.

For low spatial frequencies discrimination between fixated and non-fixated image locations was near chance for contrast, edge-content and chromaticity, suggesting that these low scales are not significantly involved in saccade targeting. The results for brightness deviate slightly from those for the above three features. At high frequencies the results are similar; locations selected for fixation are brighter or darker than expected if selection was random. However, at scales lower than 1.35 cycles per degree, ROC area values fall below 0.5, suggesting that the eye is avoiding especially bright or dark low frequency information in the image. This is consistent with a selection strategy or representation that avoids regions of sky or shadow in the images.

Preference for high spatial frequencies in target selection may arise from the task instructions. Given that bottom up guidance can only account for up to 63% of the fixation position data at best, top down modulation is likely be involved. Hence the instructions given to observers may have influenced the selection of positions for fixation during viewing. In our experiments, observers essentially undertook a memory task; they were asked questions about objects in the scenes after viewing each image. The types of objects, their locations and their sizes were varied, as were the types of questions asked about the objects. This was done in order to keep viewing as general as possible. However, the emphasis upon objects in the questioning might produce a viewing strategy biased toward the higher spatial frequencies. This possibility is currently under investigation.

Our suggestion that saccadic guidance involves an intermediate invariant representation can be seen to be consistent with the *strategic divergence* framework for saccade target selection. The suggestion that the low-level component does not change over time would be consistent with the existence of a constant intermediate representation. It is the interpretation and inspection of this intermediate representation that changes over time and with varying task demands. This variation in interpretation of the representation corresponds to the suggested top down variation in the *strategic divergence* model. Our findings open up an exciting new area of study. Can we characterise more extensively the features that are used to construct the intermediate representation upon which selection operates? When carefully controlled for, can we investigate the influence of differing higher level task demands on the interpretation and utility of the intermediate representation and hence influence what the oculomotor system prioritises in the representation as salient for selecting fixation targets? Clearly, there is much to be done in this area, but our proposition of an intermediate invariant representation and of *strategic divergence* in the higher level interpretation of this representation offers a framework within which further exploration can be structured.

Acknowledgments

We thank two unnamed reviewers for their helpful comments and insights on earlier versions of this manuscript. This research was supported by a Biotechnology and Biological Sciences Research Council grant (S15868) to BWT.

References

Baddeley, R. (1996). Searching for filters with 'interesting' output distributions: An uninteresting direction to explore? *Network-Computation in Neural Systems, 7*(2), 409–421.

Ballard, D. H., Hayhoe, M. M., Li, F., Whitehead, S. D., Frisby, J. P., Taylor, J. G., et al. (1992). Hand eye coordination during sequential tasks. *Philosophical Transactions of the Royal Society of London Series B—Biological Sciences, 337*(1281), 331–339.

Braun, J., & Sagi, D. (1990). Vision outside the focus of attention. *Perception and Psychophysics, 48*(1), 45–58.

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

Canosa, R. L., Pelz, J. B., Mennie, N. R., & Peak, J. (2003). High-level aspects of oculomotor control during viewing of natural-task images. In B. E. Rogowitz & T. N. Pappas (Eds.), *Proceedings IS & T/SPIE 15th Annual Symposium on Electronic Imaging: Human Vision and Electronic Imaging VIII* (Vol. 5007, pp. 240–251).

Croner, L. J., & Kaplan, E. (1994). Receptive-fields of P-ganglion and M-ganglion cells across the primate retina. *Vision Research, 35*(1), 7–24.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial-frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A—Optics Image Science and Vision, 2*(7), 1160–1169.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Field, D. J. (1987). Relations between the statistics of natural images and the response profiles of cortical cells. *Journal of the Optical Society of America A—Optics Image Science and Vision, 4*, 2379–2394.

Glassner, A. S. (1995). *Principles of digital image synthesis*. San Francisco: Morgan Kaufmann.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley.

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.

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

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research, 35*(13), 1897–1916.

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

Land, M. F., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception, 28*(11), 1311–1328.

Lee, B. B., Kremers, J., & Yeh, T. (1998). Receptive fields of primate retinal ganglion cells studied with a novel technique. *Visual Neuroscience, 15*(1), 161–175.

Linsenmeier, R. A., Frishman, L. J., Jakiela, H. G., & Enrothcugell, C. (1982). Receptive-field properties of X-cells and Y-cells in the cat retina derived from contrast sensitivity measurements. *Vision Research, 22*(9), 1173–1183.

MacLeod, D. I., & Boynton, R. M. (1979). Chromaticity diagram showing cone excitation by stimuli of equal luminance. *Journal of the Optical Society of America, 69*(8), 1183–1186.

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

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision, 10*(3), 165–188.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997). Fixation sequences made during visual examination of briefly presented 2D images. *Spatial Vision, 11*(2), 157–178.

Melcher, D., & Kowler, E. (2001). Visual scene memory and the guidance of saccadic eye movements. *Vision Research, 41*(25–26), 3597–3611.

Nakayama, K., & Mackeben, M. (1989). Sustained and transient components of focal visual-attention. *Vision Research, 29*(11), 1631–1647.

Niebur, E., & Koch, C. (1996). Control of selective visual attention: Modelling the where pathway. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.). *Neural information processing systems* (Vol. 8, pp. 802–808). Cambridge, MA: MIT Press.

Olshausen, B. A., Anderson, C. H., & Vanessen, D. C. (1993). A neurobiological model of visual-attention and invariant pattern-recognition based on dynamic routing of information. *Journal of Neuroscience, 13*(11), 4700–4719.

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.

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

Pelz, J. B., & Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research, 41*(25–26), 3587–3596.

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

Shinoda, H., Hayhoe, M. M., & Shrivastava, A. (2001). What controls attention in natural environments?. *Vision Research, 41*(25–26), 3535–3545.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network-Computation in Neural Systems, 14*(3), 391–412.

Treisman, A. (1988). Features and objects—the 14th Bartlett Memorial lecture. *Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 40*(2), 201–237.

Valeton, J. M., & Vannorren, D. (1983). Light adaptation of primate cones—an analysis based on extracellular data. *Vision Research, 23*(12), 1539–1547.

Vorobyev, M., & Osorio, D. (1998). Receptor noise as a determinant of colour thresholds. *Proceedings of the Royal Society of London Series B—Biological Sciences, 265*(1394), 351–358.

Wolfe, J. M., & Gancarz, G. (1996). Guided search 3.0: A model of visual search catches up with Jay Enoch 40 years later. In V. Lakshminarayanan (Ed.), *Basic and clinical applications of vision science* (pp. 189–192). Dordrecht, The Netherlands: Kluwer Academic.

Wyszecki, G., & Stiles, W. S. (2000). *Color science: Concepts and methods, quantitative data and formulae* (2nd ed.). New York: John Wiley.

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