Vision Research 102 (2014) 41–51

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

# Deriving an appropriate baseline for describing fixation behaviour

Alasdair D.F. Clarke [a,b], Benjamin W. Tatler [c,*]

[a] Institute of Language, Cognition and Computation, School of Informatics, University of Edinburgh, UK
[b] School of Psychology, University of Aberdeen, UK
[c] School of Psychology, University of Dundee, UK

ARTICLE INFO

*Article history:*
Received 17 February 2014
Received in revised form 16 May 2014
Available online 29 July 2014

*Keywords:*
Eye tracking
Central tendency
Salience
Fixation location

ABSTRACT

Humans display image-independent viewing biases when inspecting complex scenes. One of the strongest such bias is the central tendency in scene viewing: observers favour making fixations towards the centre of an image, irrespective of its content. Characterising these biases accurately is important for three reasons: (1) they provide a necessary baseline for quantifying the association between visual features in scenes and fixation selection; (2) they provide a benchmark for evaluating models of fixation behaviour when viewing scenes; and (3) they can be included as a component of generative models of eye guidance. In the present study we compare four commonly used approaches to describing image-independent biases and report their ability to describe observed data and correctly classify fixations across 10 eye movement datasets. We propose an anisotropic Gaussian function that can serve as an effective and appropriate baseline for describing image-independent biases without the need to fit functions to individual datasets or subjects.



When we view complex scenes, where we look is influenced by a combination of low-level scene statistics (Itti & Koch, 2000), higher-level interpretation of the scene (Ehinger et al., 2009; Einhäuser, Spain and Perona, 2008), task goals (Buswell, 1935; Yarbus, 1967) and behavioural biases (Tatler & Vincent, 2009). If we are to understand the relative contributions of these different sources of guidance in scene viewing then techniques are required for quantifying the extent to which decisions about where to look can be attributed to each source.

At present, existing techniques can be categorised broadly into two approaches. First, the statistical properties at the centre of gaze can be quantified in order to measure how strongly a particular feature is associated with where gaze is directed (e.g., Pomplun, 2006; Reinagel & Zador, 1999). Second, locations that are likely to be fixated can be predicted based upon the distribution of statistical properties across an image and then the correspondence between the distribution of human fixation locations and the regions predicted as likely to be fixated from the statistical distribution can be assessed (e.g., Torralba, Oliva, Castelhano, & Henderson, 2006).

Both approaches can be used to assess the potential correspondence between a variety of low- or high-level features and fixation selection: provided that the feature under investigation can be quantified at each location in the scene, it is possible to quantify the strength of that feature at fixation or its distribution over the image. However, both approaches require a baseline measure in order to consider whether the association between the feature under test and fixation is greater than that expected by chance. Typically, a randomly generated set of locations is used to sample either the strength of the feature or the probability of selecting locations that fall within the regions predicted as likely to be fixated on the basis of the feature. The extent to which the control locations and the fixated locations correspond with the feature under test can then be used to assess whether any association between the feature and fixation is greater than would be expected by chance. A powerful and commonly used approach for making this judgment is to use the signal detection theoretic measure of the area under the receiver-operating-characteristics curve (see Green & Swets, 1966). The manner in which the random locations used as the baseline for such assessments are generated has important implications for the manner in which findings can be interpreted and indeed can significantly impact on the results (Henderson, Brockmole & Castelhano, 2007; Tatler, Baddeley & Gilchrist, 2005).

One approach is to use a uniform distribution for selecting control locations (e.g., Einhäuser, Spain & Perona, 2008; Parkhurst, Law & Niebur, 2002; Reinagel & Zador, 1999). Using such an approach means that any association between fixation and the feature under test that is beyond that found in the baseline comparison can be interpreted as suggesting that the feature is selected more

* Corresponding author. Fax: +44 (0) 1382 229993.
*E-mail address:* b.w.tatler@dundee.ac.uk (B.W. Tatler).

http://dx.doi.org/10.1016/j.visres.2014.06.016
0042-6989/

than would be expected if the eyes were directed randomly around a scene.

However, the existence of behavioural biases in how we view scenes (Tatler, 2007; Tatler & Vincent, 2009) suggests that a uniform random baseline may misrepresent selection with respect to features. That is, if the baseline comparison uses a uniform random distribution for generating control locations, any association found between fixation and features that extends beyond that in the baseline condition is likely to reflect a combination of selection based on image properties and image-independent biases in fixation behaviour. A more appropriate baseline for evaluating the association between an image feature and fixation placement is to select control locations from a distribution that reflects any image-independent biases in viewing behaviour. The most prominent and well-characterised image-independent bias in scene viewing is the central fixation bias: humans preferentially fixate the centre of the scene in a manner that is almost independent of the scene displayed to observers (Tatler, 2007; Tseng et al., 2009). As a result, control fixations can be drawn from distributions that reflect this central bias (see Tatler, 2007; Tatler, Baddeley & Gilchrist, 2005, for discussion of this issue).

There exist a number of ways that are typically used to construct a centrally-weighted distribution used in the baseline condition. One approach is to use a centred Gaussian to approximate the central bias and this may be fitted to the overall distribution of fixation locations in a dataset (Zhao & Koch, 2011), or scaled to the aspect ratio of the images presented (Judd, Durand & Torralba, 2012). Alternatively, these control distributions may be generated in ways that are aimed to maximise the chance of capturing any individual viewing biases that participants display when viewing scenes. There exist two main ways of attempting to capture individual viewing biases in baseline samples of features. First, the *(x,y)* locations of fixations on the test image can be used to sample features at the same locations in another (randomly selected) image (Parkhurst & Niebur, 2003). Second, $(x,y)$ locations of fixations made by the same participant but when viewing different images can be used to sample features on the test image (e.g., Tatler, Baddeley & Gilchrist, 2005; Tatler & Vincent, 2009).

At present, it is unclear whether and how these different approaches to creating a baseline distribution vary in their suitability. The present study compares distributions of fixations across multiple existing datasets of eye movements in order to consider whether a single common distribution might be an appropriate baseline across studies and individuals or whether it is necessary to tailor the baseline distribution to each study and individual.

Being able to capture the statistics of the baseline condition appropriately is necessary for three reasons. First, if we wish to consider the relative importance of any feature in decisions about where to look, it is desirable to be able to quantify the unique variance associated with the particular feature after removal of variance associated with other factors that may contribute to decisions about where to look. In this way, any assessment of the importance of visual information (low- or high-level) to fixation selection should partial out variance that is associated with any image-independent biases in looking behaviour. Thus, if we compare the feature of interest to an appropriate baseline that accounts for image-independent biases, then we are better able to characterise associations between that feature and fixation behaviour. This principle extends beyond simply evaluating low-level salience models to any domain in which it is desirable to be able to characterise the contribution of a particular source of information to inspection behaviour. For example, in visual search paradigms, it is also useful to be able to remove any component of the behaviour that is driven by looking biases that are unrelated to the stimuli displayed.

Second, we can use this baseline as a benchmark for evaluating models of eye movement behaviour in scene viewing, as employed by Judd, Durand and Torralba (2012). Models should at least be able to outperform a baseline model based on image-independent biases such as looking at the centre of the screen. In their extensive comparison of a range of different salience models, Judd, Durand and Torralba (2012) found that only two models managed to outperform an image independent central bias baseline constructed using an aspect ratio-scaled Gaussian distribution. As there appears to be no empirical basis for this exact baseline, this may in fact underestimate the amount of variance that can be explained, and hence over-estimates the performance of the salience models.

Third, we can treat any image-independent bias as a factor in eye movement control itself. Thus, if we can computationally model these biases and derive appropriate characterisations of these biases we can use these as a component of models of fixation selection. That is, we can produce models with modules for low-level information, high-level information and image-independent biases. Given the strength of the central bias and its ability to predict human fixations, it is surprising that it is not more commonly incorporated into computational models. Indeed in their review, Judd, Durand and Torralba (2012) found only three studies that explicitly included a central bias in their model: Parkhurst and Niebur (2003) use the "shuffle" method; Zhao and Koch (2011) fitted Gaussians to their data, but restricted their baseline to an isotropic central bias, i.e., they fitted a covariance matrix with equal horizontal and vertical variance; and Judd et al. (2009) used an isotropic Gaussian fall-off that was stretched to match the aspect ratio of the image. Other examples in the literature include Clarke, Coco and Keller (2013) who used Euclidean distance from the centre of the image, and Spain and Perona (2011) who used a wide range of distance functions based on the Euclidean metric. Appropriate characterisation of image-independent biases therefore will allow appropriate and effective additions to existing models of fixation selection.

In the present study we evaluated different approaches to characterising baselines for understanding fixation behaviour when viewing scenes. Using ten eye movements datasets, we compared four ways of characterising image-independent biases in fixation selection: (1) fitting an isotropic Gaussian to the data (as in Zhao & Koch, 2011), (2) fitting a Gaussian scaled to the aspect ratio of the images (as in Judd, Durand & Torralba, 2012), (3) anisotropic Gaussians where the vertical and horizontal variances were fitted to each dataset, and (4) anisotropic Gaussians where the vertical and horizontal variances were fitted to each participant within each dataset. The final two approaches attempt to capture any experiment-specific (approach 3) or subject-specific (approach 4) differences in image-independent biases and as such conform to the recommendations made in previous discussions of this issue (Borji, Sihite & Itti, 2013a, 2013b; Tatler, Baddeley & Gilchrist, 2005). By comparing across these four approaches we were able to consider the relative ability of each approach for describing the data effectively and also the impact that each approach has upon our ability to classify fixated and control locations using each approach. One potential problem with the subject-level fitting (approach 4) is that this is likely to be sensitive to the sample size of eye movements used to construct the baseline distributions. This is a particular issue in studies with small numbers of trials or short presentations times (hence few fixations per image). As a result we considered how these approaches for describing the baseline are influenced by small $n$. In all of these approaches an empirical fit of the data is required to produce the baseline. We considered whether this is really necessary or whether a general purpose function can be employed that can be used irrespective of the subject or experiment under investigation. Here we used

the average vertical and horizontal scaling parameters from our dataset-fitting approach (approach 3) as a general purpose baseline. We evaluated the ability of this baseline to explain the observed data and to classify fixated vs. control locations. From these comparisons we are able to make a recommendation for best practice when evaluating feature selection and model performance or when constructing models of fixation selection in scene viewing.

## 1. Method

### 1.1. Datasets

In the present study, we considered a collection of ten datasets collected over the previous decade. A number of different tasks are represented, including free-viewing, visual search, memory and scene description. Table 1 provides a summary of the number of subjects and images in each dataset together with the task and display durations. Table 2 shows details of the experimental setups in each of the 10 datasets analysed in the present study.

The images in seven of the ten datasets had an aspect ratio of 4:3. The images used by Yun et al. (2013) covered a range of aspect ratios but 4:3 was by far the most common and so we restricted our analysis to these images. The only other aspect ratio represented was 5:4 (Asher et al., 2013). The photographs used by Einhäuser, Spain and Perona (2008) are of mixed aspect ratio, but the images have had large black borders added which bring their aspect ratio up to 4:3.

For our analyses we excluded any fixations that fell outside the borders of the images. The very first fixation in each trial was excluded because it began prior to scene onset and its location was determined by the location of the pre-trial fixation target rather than any content of the scene that followed. For all remaining fixations, the $x$ and $y$ coordinates were normalised by half of the width of the image, i.e., the centre of the image corresponded to $(0,0)$ and fixations were points in the space $[-1,1]\times[-a,a]$, where $a$ is the aspect ratio used in the dataset (typically, $a=0.75$). An example of the distribution of fixations in a dataset is shown in Fig. 1.

### 1.2. Modelling

#### 1.2.1. Fitting empirical data

Previous implementations of the central bias are generally based on either the Euclidean distance-to-centre (Clarke, Coco & Keller, 2013) or a multivariate Gaussian probability density function. This Gaussian is sometimes isotropic (Zhao & Koch, 2011), and sometimes set so that the ratio of horizontal to vertical variance is set to the aspect ratio of the image (Judd, Durand & Torralba, 2012). From the form of the distribution in Fig. 1 it would appear that both Euclidean and Gaussian fall-offs are likely to provide a good fit with the data. However, we favour using a Gaussian as it has the desirable characteristic of assigning a positive, non-zero probability of fixation to all image locations. More specifically, we use a two-dimensional Gaussian pdf with zero mean and covariance matrix given by:

$$\begin{pmatrix} \sigma^2 & 0 \\ 0 & v\sigma^2 \end{pmatrix} \tag{1}$$

where $\sigma^2$ is the horizontal variance of the fixations. We then calculate the likelihood of the data for distributions with various $v$.

We evaluate five different methods for producing a centrally-weighted Gaussian baseline:

- *Isotropic:* $\sigma^2$ is fitted to data, $v=1$.
- *Aspect ratio:* $\sigma^2$ is fitted to data, $v=0.75$ (=0.8 for Asher, Tolhurst, Troscianko, & Gilchrist (2013)).
- *Experiment-fitted:* $\sigma^2$ and $v$ fitted to whole dataset for each experiment.
- *Subject-fitted:* $\sigma^2$ and $v$ fitted to each participant's data.
- *Proposed baseline:* $\sigma^2$ and $v$ fixed across subjects and datasets, set to average values from the *experiment-fitted* fits.

In the first four cases fits were optimised to explain the observed fixation distributions (maximising likelihood). For the *proposed baseline* parameters were set to the average $\sigma^2$ and $v$ derived from the *experiment-fitted* parameter estimates.

#### 1.2.2. Classification performance of baseline models

In order to demonstrate that setting $v<1$ leads to a significant improved description of the data, we evaluated the ability of each of the five baseline models to classify the empirical data from an equal number of uniformly distributed samples. We did this by training a logistic classifier and reporting the area under the ROC curve (AUC). AUC values are reported as the mean of 1000-bootstrapped samples with range and interquartile range shown in box-and-whisker plots of the data in order to assess the relative classification abilities of the five methods.

#### 1.2.3. Sensitivity of models to varying n

A key issue in evaluating the suitability of our various baseline models is how robust these approaches are to variation in the amount of data over which baseline fits are fitted. Our fourth proposed approach – *subject-fitting* – is particularly at risk of requiring fits over small numbers of fixation locations. Small sample sizes may result from small numbers of trials $n_t$ or from short presentation times (therefore small numbers of fixations per trial, $n_f$). In the present study, we explored the effect of sample size on the *subject-fitted* baseline as this is both the most commonly used approach and is the one most at risk from small sample sizes. We considered the effect of varying number of trials, for two datasets with large numbers of trials (Clarke, Coco & Keller, 2013; Judd, Durand & Torralba, 2012) per participant by randomly selecting subsets of trials varying in size from $1{:}n_t$. For this analysis we used 10-fold cross validation to calculate performance of fits based upon differ-

**Table 1**
Summary of the 10 datasets used throughout this study.

| | Observers | Images | Task | Display duration |
|---|---|---|---|---|
| Clarke, Coco and Keller (2013) | 24 | 100 | Object naming | 5000 ms |
| Yun et al. (2013) – SUN | 8 | 104 | Image description | 5000 ms |
| Tatler, Baddeley and Gilchrist (2005) | 14 | 48 | Memory | Variable |
| Einhäuser, Spain and Perona (2008) | 8 | 93 | Object naming | 3000 ms |
| Tatler (2007) – free | 22 | 120 | Free viewing | 5000 ms |
| Judd et al. (2009) | 15 | 1003 | Free viewing | 3000 ms |
| Yun et al. (2013) – PASCAL | 3 | 1000 | Free viewing | 3000 ms |
| Ehinger et al. (2009) | 14 | 912 | Visual search | Variable |
| Tatler (2007) – search | 30 | 120 | Visual search | 5000 ms |
| Asher et al. (2013) | 25 | 120 | Visual search | Variable |

**Table 2**
Details of the experimental setups in each of the 10 datasets analysed in the present study. We provide only information reported in the original articles. Question marks indicate information not reported in the original article.

| | Eye tracker | Viewing distance | Screen size | Image size | Viewing angle | Chin head rest |
|---|---|---|---|---|---|---|
| Clarke, Coco and Keller (2013) | EyeLink II | 50 cm | 21″ | 800 × 600 | 31 × 25° | No |
| Yun et al. (2013) – SUN | EyeLink 1000 | ? | ? | ? | ? | ? |
| Tatler, Baddeley and Gilchrist (2005) | EyeLink I | 60 cm | 17″ | 800 × 600 | 30 × 22° | No |
| Einhäuser, Spain and Perona (2008) | EyeLink 1000 | 80 cm | 20″ | 1024 × 768 | 29 × 22° | Yes |
| Tatler (2007) – free | EyeLink II | 60 cm | 21″ | 1600 × 1200 | 40 × 30° | No |
| Judd et al. (2009) | ? | 2 feet | 19″ | 1024 × 768[a] | ? | Yes |
| Yun et al. (2013) – PASCAL | EyeLink 1000 | ? | ? | ? | ? | ? |
| Ehinger et al. (2009) | ISCAN RK-464 | 75 cm | 21″ | 800 × 600 | 23.5 × 17.7° | Yes |
| Tatler (2007) – search | EyeLink II | 60 cm | 21″ | 1600 × 1200 | 40 × 30° | No |
| Asher et al. (2013) | EyeLink 1000 | 55 cm | ? | 1024 × 1280 | 37.6 × 30.5° | Yes |

[a] For the Judd et al. dataset images varied in pixel dimensions but the majority were at 1024 × 768.

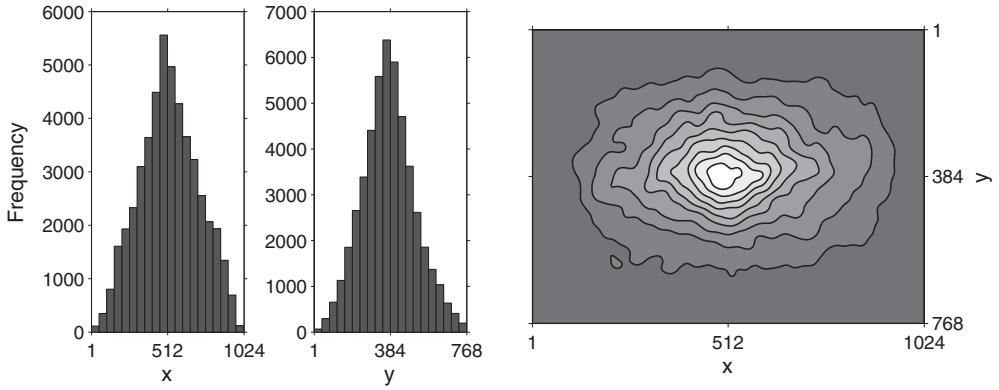


**Fig. 1.** Distribution of fixations from the Judd et al. (2009) dataset for each of the *x* and *y* dimensions alone, together with a contour plot of the *xy* joint distribution. In the contour plot, lines delineate deciles in the dataset.

ent sized (varying $n_t$) subsets from 90% of the data for classifying the remaining 10% of the data. The plotted AUC values are means across 10 bootstrapped samples in each of the 10 folds.

Having a small number of fixations per trial is a common problem but realistically only arises as a result of short presentation times. In such situations, the distributions of fixations and hence the baseline estimates will be influenced not only by the limited availability of fixations, but also by the known variation in image-independent biases over the first few fixations after scene onset. In particular the image-independent bias to look near the centre of scenes is known to vary in strength as viewing progresses, with a more pronounced bias soon after scene onset than later in viewing (Tatler, 2007). We therefore took a second approach to modelling the influence of small numbers of fixations per trial by considering the suitability of our *proposed baseline* as a function of the number of fixations in a trial by fitting $\sigma^2$ and $v$ to only the first $n$ fixations in each dataset. We also fitted functions to describe how $\sigma^2$ and $v$ vary with the number of fixations collected. While fitting the first $n_f$ fixations gives a realistic impression of the suitability of our *proposed baseline* and the reliability of *subject-fitted* baselines for datasets comprising trials of varying duration, it does not allow us to describe the ability to fit image-independent biases at any given moment in viewing. In order to consider this issue we fit data for the $n$th fixation in each dataset. Taken together our fits of the first $n_f$ fixations and the $n$th fixation in viewing allow us to characterise not only how well different baseline approaches described the data for varying trial durations, but also how the baseline fits varied over the course of viewing. Any change in $\sigma^2$ and $v$ reflect how the distribution of fixations changed over fixation number, with larger $\sigma^2$ indicating greater spread and larger $v$ indicating a greater horizontal to vertical ratio.

## 2. Results

### 2.1. Fitting empirical data

In the evaluated datasets the means of the fixation locations were indeed at the centre of the image (Fig. 2). This suggests that there were no large systematic biases towards fixating one region over any other. For example, a bias towards fixating the lower half of the image would not have been surprising as, generally, there is more informative image content below the mid-line horizon than above: the upper half of images is more likely to contain sky (outdoor scenes) or walls/ceilings (indoor scenes) and as such is less likely to contain informative scene content. However, like Tatler (2007) we found no such vertical shift, with all distributions centred around the vertical and horizontal centre of the scene. Therefore, in the analyses that follow we used a Gaussian centred at the scene centre.

For the *isotropic* and *aspect ratio* baseline models we fitted a diagonal covariance matrix to each dataset, allowing $\sigma^2$ to vary, but setting $v$ to 1 in the case of the *isotropic* model and 0.75 (or 0.8 for the Asher et al. (2013) dataset) for the *aspect ratio* model. For the *experiment-fitted* and *subject -fitted* baseline models we fitted a diagonal covariance matrix to each dataset, allowing both $\sigma^2$ and $v$ to vary.

The distributions of $\sigma^2$ and $v$ for the *experiment-fitted* baseline model are given in Fig. 3. For all of the evaluated datasets $v$, the ratio of vertical to horizontal variance, was not only less than 1 (i.e., vertical variance was less than horizontal variance), but was also less than would be expected from the aspect ratio of the images (typically 0.75). The mean value, $v = 0.45$ suggests that the vertical variance is less than half the horizontal variance.

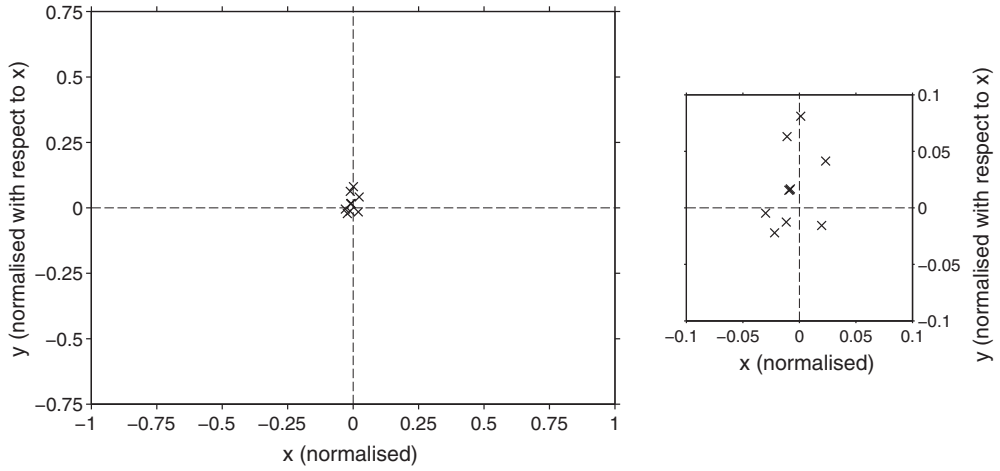


**Fig. 2.** Mean fixation positions for the seven non-search datasets. Each × shows the mean fixation location over all fixations (pooled over subjects and trials) in the dataset. Right, zoomed plot of the central region of the scenes to show dispersion of mean fixation locations in each dataset around the central point in the screen.

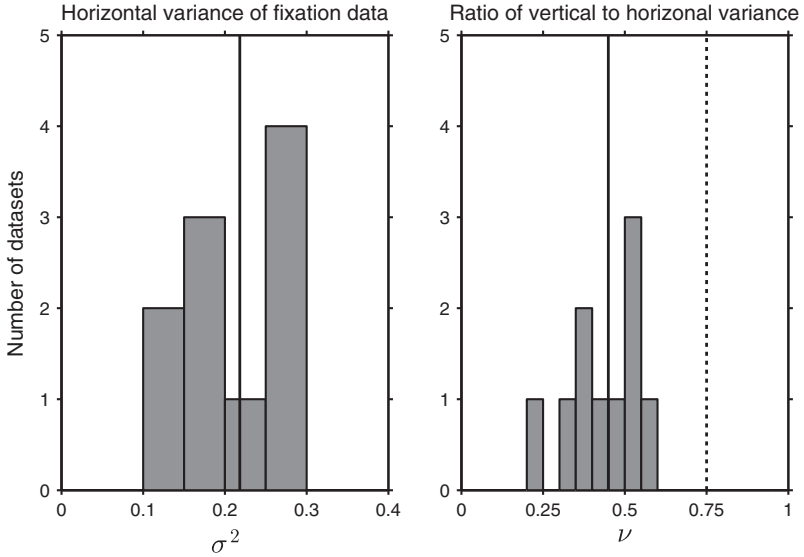


**Fig. 3.** Distribution of $\sigma^2$ and $v$ in the 10 datasets considered in the present study. The solid vertical line shows the mean over datasets, while the dotted vertical line in the plot of $v$ indicates 0.75, the aspect ratio of the majority of images the 10 datasets considered in our study.

For the *proposed baseline* model we used the mean $\sigma^2$ and $v$ across datasets calculated using the *experiment-fitted* baseline fits. As such, in our *proposed baseline* model $\sigma^2 = 0.22$ and $v = 0.45$.

Fig. 4 shows how varying $v$ affected the likelihood of the fixations from each dataset. The results consistently showed that the horizontal variation is larger than the vertical variation, and furthermore, that setting $v$ to the aspect ratio of the image does not capture all of this effect. We also found that the difference between fitting each dataset individually, and just using $\sigma^2 = 0.22$, $v = 0.45$ was comparatively minor. Interestingly the Ehinger et al. (2009) dataset appears to be an outlier. Presumably this is due to the nature of the images: when searching for pedestrians in photographs of street scenes, it is unsurprising there are more fixations located along a horizontal band and less variance in the vertical direction, and indeed the authors use a horizontal band as a contextual prior in their study.

### 2.2. Classification performance of baseline models

We assessed the ability of each of our five baseline models to distinguish the empirical fixations from a set of uniformly distributed points using logistic regression. The results are shown in Table 3 and Fig. 5. For all ten datasets, the isotropic baseline performed the worst. The differences between our *proposed baseline* and both the *experiment-fitted* and *subject-fitted* baseline models were relatively minor.

### 2.3. Sensitivity of models to varying n

If we estimate the central bias from small sample sizes, the estimate is likely to be a poorer fit and thus less well able to explain empirical fixation distributions. Small sample size for estimates of baselines is a particular problem for *subject-fitted* baselines, especially if either the number of trials is small or the presentation times for images are short. We simulated the problem of small numbers of trials on the estimates of the central fixation bias by randomly sampling $n_t$ trials (Fig. 6) for two datasets and considering how well baselines fitted to these limited samples explained data on other (test) data from the same subject. For reference the performance of our *proposed baseline* is also plotted alongside these fits. It is clear that when the size of the dataset is limited by having few trials, *subject-fitted* Gaussians were poor estimates of the

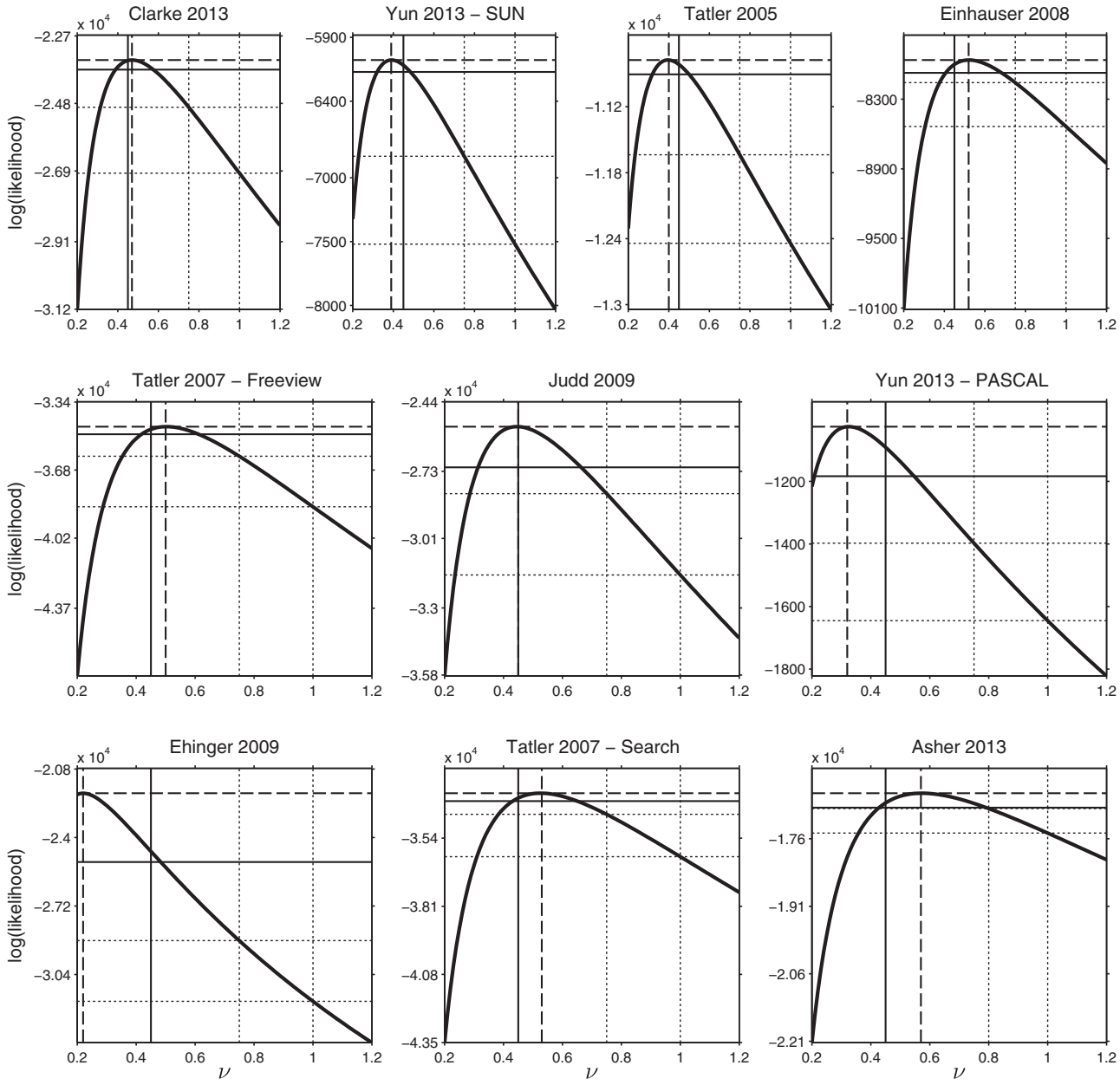


**Fig. 4.** The influence of varying $v$ on the likelihood of the fixations from each dataset. In each plot, the curve shows the effect of varying $v$ on the likelihood for Gaussians with $\sigma^2$ fitted to each dataset individually. Dashed crosshairs show the value of $v$ that offers the best description of the observed data. Dotted crosshairs show the likelihoods for Gaussians used in two previous approaches for describing baseline distributions: setting $v$ to the aspect ratio of the images, $v = 0.75$, or using isotropic, $v = 1$, Gaussians to describe central tendencies. The solid black crosshairs show the likelihood for each dataset using the baseline model proposed in the present study, $\sigma^2 = 0.22$ and $v = 0.45$.

**Table 3**
Area under ROC for each of the five baseline models evaluated for each of the 10 datasets.

| | Isotropic | Aspect ratio | Experiment fitted | Subject fitted | Proposed baseline |
|---|---|---|---|---|---|
| Clarke, Coco and Keller (2013) | 0.728 | 0.736 | 0.742 | 0.741 | 0.742 |
| Yun et al. (2013) – SUN | 0.738 | 0.733 | 0.734 | 0.731 | 0.74 |
| Tatler, Baddeley and Gilchrist (2005) | 0.631 | 0.642 | 0.66 | 0.661 | 0.658 |
| Einhäuser, Spain and Perona (2008) | 0.751 | 0.759 | 0.767 | 0.766 | 0.769 |
| Tatler (2007) – free | 0.714 | 0.72 | 0.724 | 0.724 | 0.724 |
| Judd et al. (2009) | 0.780 | 0.788 | 0.795 | 0.799 | 0.795 |
| Yun et al. (2013) – PASCAL | 0.796 | 0.807 | 0.823 | 0.824 | 0.820 |
| Ehinger et al. (2009) | 0.646 | 0.668 | 0.729 | 0.732 | 0.703 |
| Tatler (2007) – search | 0.619 | 0.624 | 0.628 | 0.630 | 0.628 |
| Asher et al. (2013) | 0.590 | 0.594 | 0.597 | 0.601 | 0.597 |
| Improvement over isotropic | – | 0.010 | 0.024 | 0.025 | 0.020 |

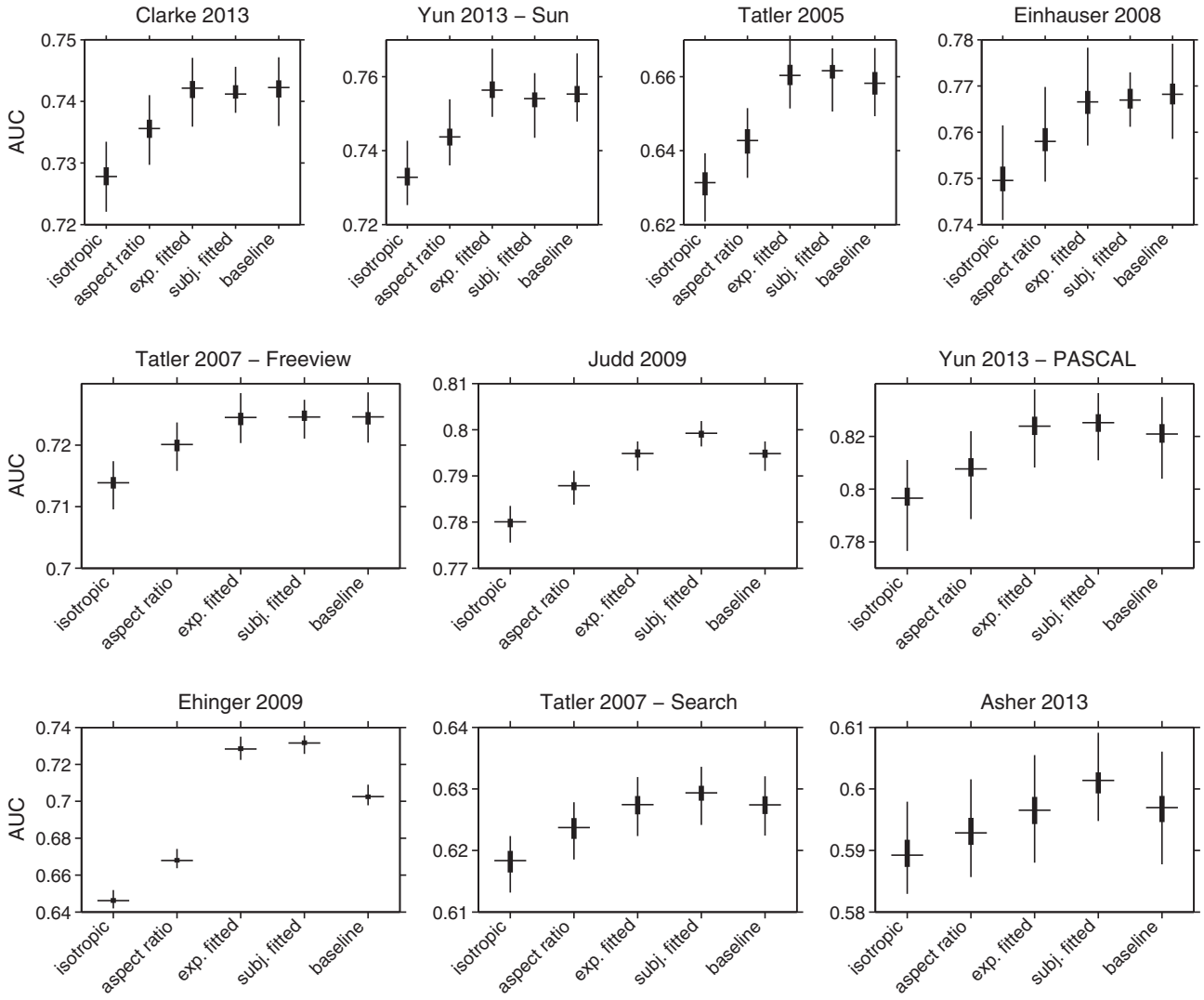


**Fig. 5.** Classification performance for the five baseline models for each of the 10 datasets. Classification performance was assessed by training a logistic classifier and testing its ability to distinguish fixations from uniformly distributed samples, for which we report the area under the ROC curve (AUC). In these box-and-whisker plots the horizontal line shows the mean AUC from 1000 bootstrapped samples, the filled box indicates the interquartile range and the whiskers indicate the full range in the data.

underlying central bias in fixation behaviour. As such, subject-specific fits based on small $n$ will be a less reliable baseline than a baseline with our proposed fixed vertical and horizontal scaling (Fig. 6).

Fig. 7 shows how $\sigma^2$ and $v$ varied when fitting only the $n$th fixation in each trial or the first $n$ fixations of each trial in each dataset. The patterns are similar for $\sigma^2$ and $v$ for the $n$th or first $n$ fixations in each trial: with lower $\sigma^2$ and – to a lesser extent – higher $v$ early in the trial than later in the trial. These changing values show that the first few fixations after scene onset were distributed differently from later fixations, with less horizontal spread and greater vertical spread. We fitted functions to describe the change in mean $\sigma^2$ and $v$ over the first $n$ fixations in a trial as follows: $\sigma^2 = 0.23 - 0.29/n$ (fit $r^2 = 0.99$), $v = 0.43 - 0.09/e^n$ (fit $r^2 = 0.97$) where $n$ = the number of fixations collected per trial. It should be noted that $v$ asymptotes very early in viewing, at around 3–4 fixations. As such, no modification of our *proposed baseline* $v$ of 0.45 is necessary provided presentations times allow at least 3–4 fixations per trial in any experiment. For $\sigma^2$, some modification of our *proposed baseline* $\sigma^2$ of 0.22 is necessary for experiments where fewer than 10 fixations are collected per trial, and for this we recommend using the function above.

## 3. Discussion

Characterising image-independent biases in eye movements, such as the tendency to look at the centre of a scene, is important for understanding eye guidance in scene viewing for at least three reasons. First, image-independent biases are an appropriate and necessary baseline for quantifying the association between visual features in scenes and fixation selection (e.g., see Tatler, Baddeley & Gilchrist, 2005). Second, the overall performance of models of fixation selection is often measured by comparing the model's performance to that from a reference model based on image-independent biases such as the central fixation tendency (e.g., Judd, Durand & Torralba, 2012). Therefore, an appropriate description of the image-independent biases in scene viewing is essential for making such evaluations. Third, we can incorporate the description of image-independent biases into models of fixation selection in order to improve the ability of models to generate human-like fixation behaviour (see Clarke, Coco & Keller (2013), Judd, Ehinger, Durand, & Torralba (2009), Parkhurst & Niebur (2003), Spain & Perona (2011) and Zhao & Koch (2011), for examples of incorporating central tendencies into models).

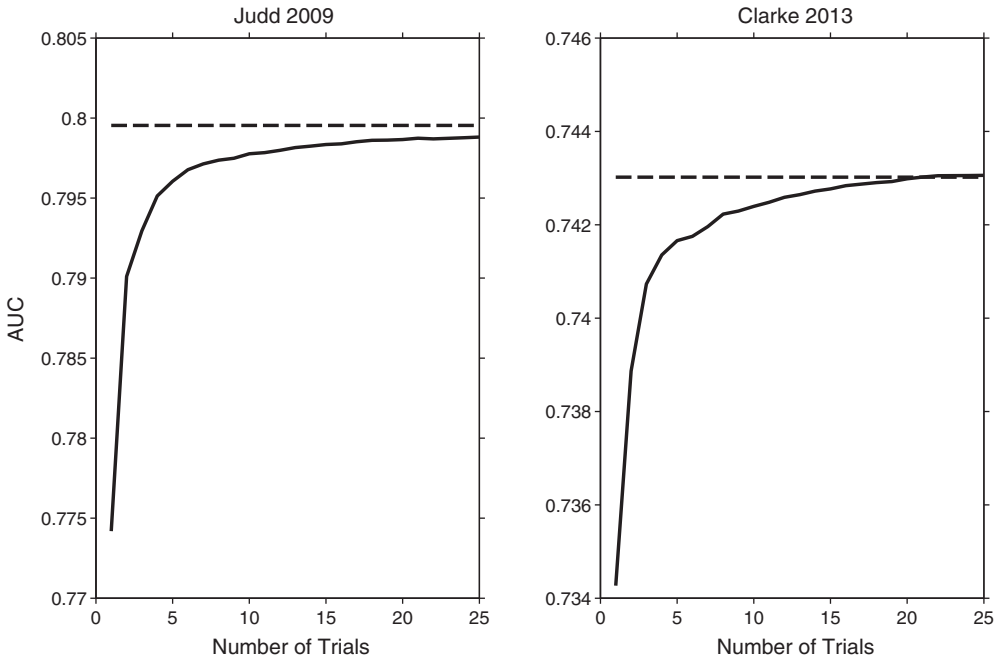


**Fig. 6.** Classification performances for baseline estimates based on $n_t$ trials for classifying unseen test data. Performances are based on 10-fold cross validations and plots are of average AUC values across 10-bootstrapped samples within each of the 10 folds. In each plot the dotted line shows the classification performance of our recommended baseline function. Our proposed baseline offers considerably better classification performance than subject-fitted baselines when datasets are limited by small numbers of trials.

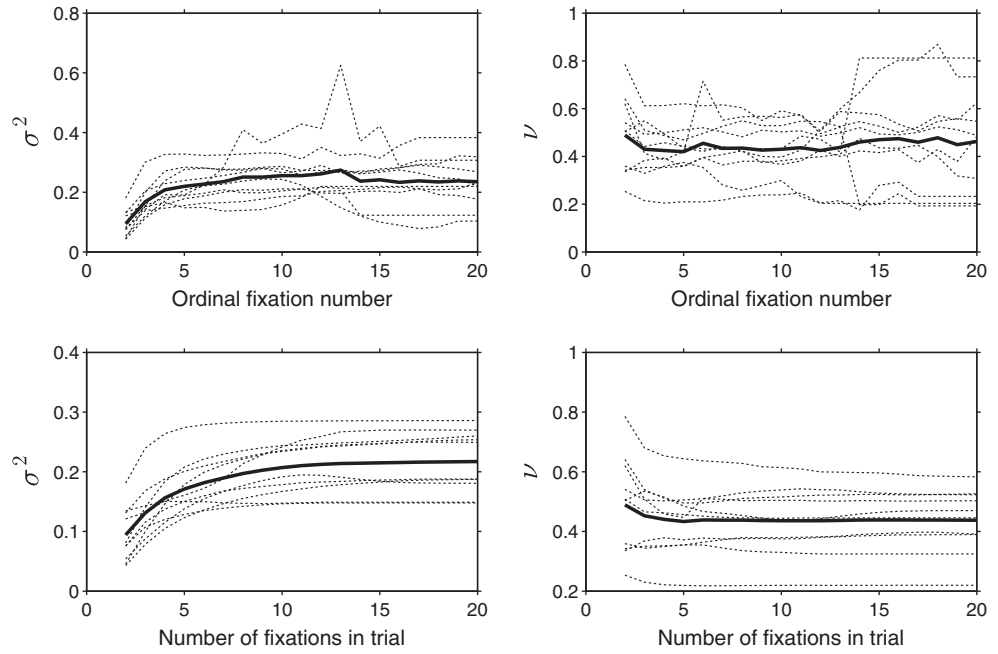


**Fig. 7.** How $\sigma^2$ and $v$ vary with the number of fixations collected after scene onset. The solid black line shows the average across datasets. Each black dotted line shows an individual dataset.

Previous authors have argued for the need to compare feature content at fixated locations to that at control locations, with control locations drawn from a distribution that reflects any image-independent biases in inspection behaviour (e.g., Borji & Itti, 2013; Judd, Durand & Torralba, 2012; Tatler, Baddeley & Gilchrist, 2005). However, different conventions exist for constructing the distribution for this baseline comparison dataset (Borji, Sihite & Itti, 2013a, 2013b). Some authors have used a uniform distribution for generating baseline locations (e.g., Einhäuser, Spain & Perona, 2008). However, other authors have preferred measures that sample non-uniformly for their baseline samples in order to capture aspects of the typical, non-uniform inspection behaviour that is ubiquitous in scene viewing. A popular approach for constructing a baseline distribution is to use the fixation locations of the same individual on other images (e.g., Tatler, Baddeley & Gilchrist, 2005). Using uniform random sampling or

sampling that reflects image-independent biases can produce very different findings, and it has been argued that uniform sampling for baseline comparisons can mis-represent the association between low-level features and fixation placement (Henderson, Brockmole & Castelhano, 2007; Tatler, Baddeley & Gilchrist, 2005).

Using baselines that capture viewing biases such as the tendency to look at the centre of the screen allows evaluations to essentially account for between- or within-individual image independent biases in inspection behaviour. However, creating a baseline in this way is problematic if (1) the number of images in a study is small or (2) presentation times are short. In both cases, the number of samples that are used to create the baseline set of locations will be small and thus estimates of any effect may be noisy. It is also not clear from previous studies that have recommended such approaches whether this degree of description is necessary to provide an appropriate baseline: that is, is it necessary to construct baseline samples that reflect individual- and experiment-specific biases, or can a function that describes image-independent biases across individuals and tasks be employed? Such a function would remove the issues associated with small datasets sizes when gathering baseline samples.

In the present study we found that fitting an anisotropic Gaussian to each dataset (thus to each experiment) produced a better description of viewing behaviour than an isotropic Gaussian fitted to each dataset. Thus, a baseline distribution that reflects a greater spread of fixations horizontally than vertically provides a better description of fixation behaviour than an isotropic distribution. If the anisotropy was scaled to the aspect ratio of the image, the baseline model both described the empirical observations better and classified fixation data more reliably than the *isotropic* baseline model. Descriptive power and classification accuracy were better still if the horizontal and vertical scaling of the Gaussian was fitted to each dataset (*experiment-fitted* baseline). In this case we found that for all datasets the best vertical scaling was less than that of the aspect ratio – that is, vertical spread was less than would be expected from the aspect ratio of the images alone. Fitting the Gaussian separately to each individual (*subject-fitted* baseline) produced a baseline that was comparable to the *experiment-fitted* baseline in nine of the ten datasets we explored: only for the Judd et al. (2009) dataset was the *subject-fitted* baseline noticeably better than the *experiment-fitted* baseline model for classifying fixations. This interestingly suggests that there may be little statistical advantage to constructing baselines separately for each participant in an experiment and that fitting across participants at the experiment level provides a baseline that is equally appropriate for meaningful statistical comparisons.

One issue with creating a baseline by fitting to individual subjects or experimental datasets is that the estimate of the underlying biases will become noisy for smaller dataset sizes. In an attempt to circumvent this potential issue we evaluated a baseline model constructed by taking the average vertical and horizontal scaling from the *experiment-fitted* Gaussians. The rationale is that it would be of benefit to be able to use a fixed Gaussian function for any dataset of fixations gathered during scene viewing experiments. We found that the mean classification performance of our *proposed baseline* model was surprisingly close to that for the *experiment-fitted* and *subject-fitted* baseline models. Indeed in nine of the ten datasets the interquartile ranges for these three baseline models overlapped considerably. Only for the Ehinger et al. (2009) dataset was our *proposed baseline* model noticeably inferior to the *experiment-fitted* and *subject-fitted* baseline models. It may be that the combination of task (find people) and image set (street scenes) resulted in a distribution of fixation behaviour that was unlike that found in the other datasets we evaluated. Indeed, we found that the effect of varying the vertical scaling for the fitted Gaussians was very different for this dataset than for the other nine datasets. The authors themselves found that a horizontal band across the extent of a scene offered a good explanation of the data (providing a contextual prior for searching for people in street scenes). Because removing any influence of image-independent biases such as the central fixation bias from evaluations of other factors in models of scene viewing is advantageous, we would argue that it is advantageous to include our proposed baseline model even in datasets like that collected by Ehinger et al. where our proposed baseline offers a poor overall fit to the observed data. In doing so we isolate any fraction of inspection behaviour that is attributable to these biases and so obtain a potentially better and fairer estimate of the contribution of other factors to eye guidance.

The good performance of our *proposed baseline* model across a variety of experimental tasks and subjects suggests that there is no need to fit experiment-level or subject-level differences in inspection biases. We therefore propose from the datasets examined in the present report that an appropriate baseline distribution for experiments using images with aspect ratios around 4:3 is a Gaussian probability density function with zero mean and covariance matrix $[\sigma^2, 0; 0; v\sigma^2]$ where $\sigma^2 = 0.23$ and $v = 0.45$. This recommended baseline avoids the risk of failing to estimate the influence of image-independent biases when fits are based on small numbers of observations: we found that the *subject-fitted* baseline offered poor descriptions of fixation distributions for small numbers of trials. Indeed, the *subject-fitted* baseline and was poorer than our *proposed baseline* if the number of trials in the experiment was less than around 15. Given that the strength of the central bias in scene viewing is higher early in viewing than later on Tatler (2007), we considered the influence of small numbers of fixations by modelling how $\sigma^2$ and $v$ in our *proposed baseline* change depending upon either (1) whether we modelled only the $n$th fixation in a trial or (2) how many fixations are collected after scene onset (thus modelling only the first $n$ fixations per trial). We found $v$ to be relatively unaffected by the number of collected fixations, with little change over fixations if only modelling the $n$th fixation and only a small increase when modelling fewer than the first 3–4 fixations per trial. We therefore suggest that $v$ of 0.45 is likely to be appropriate irrespective of presentation time in an experiment, especially if trial durations allow at least 3–4 fixations to be collected. For $\sigma^2$, it may be necessary to change the value used for trials in which fewer than 10 fixations are collected per trial (although little change is seen beyond the first five fixations), or if modelling any individual fixation up to around the 5th–7th in viewing. When modelling data collected with trial durations that result in fewer than 10 fixations per trial we therefore recommend using a $\sigma^2 = 0.23 - 0.29/n_f$.

There now exists a number of models of salience in scenes. These models use feature-level descriptions of scenes, typically describing the extent to which particular pixels or groups of pixels differ from their immediate surroundings or the scene as a whole (see Borji & Itti, 2013; Judd, Durand & Torralba, 2012). These feature-level descriptions are then compared to human fixation behaviour in order to evaluate whether they offer good descriptions of how scenes are viewed. Recently, Judd, Durand and Torralba (2012) suggested that an appropriate benchmark for testing the performance of salience models is to compare their ability to account for human fixation locations to the ability of a centre-bias baseline model to account for the same fixation locations. In their evaluation, only two models outperformed their central bias baseline model: GBVS (Harel, Koch & Perona, 2006) and the authors' proposed model. Moreover, the difference in AUC between these two models and the baseline was small: 0.018 for GVBS and 0.028 for the authors' proposed model. For their baseline model, Judd et al. used a Gaussian that reflected the aspect ratio of the viewed scenes. We found that fitting the Gaussians to the individual datasets, individual subjects or using our proposed baseline

settings provided a better description of the centre bias than our aspect ratio model and that this resulted in an increase in AUC of 0.014, 0.015 and 0.010 respectively over the aspect ratio model. Thus, using any of these descriptions of the image-independent biases narrows the gap further between the best performing models of salience and a simple image-independent centre bias model. It is therefore vital that any evaluation of the performance of a computational model of salience should employ the most appropriate description of image-independent biases as a baseline condition. It remains to be seen whether existing models of salience can outperform more appropriate descriptions of the centre bias in scene viewing.

If the goal of modelling viewing behaviour is to produce a model that generates and predicts fixation behaviour rather than describes it, then factors that contribute to fixation selection processes should be accurately described and incorporated into models. As a result, such models increasingly include a component engineered to reflect image-independent biases to fixate the centre of the scene (Clarke, Coco & Keller, 2013; Judd et al., 2009; Parkhurst & Niebur, 2003; Spain & Perona, 2011; Zhao & Koch, 2011). Our proposed baseline offers a parameter-free component that can be incorporated into models of fixation behaviour, which describes centre biases across databases robustly, consistently outperforming baselines based on an isotropic central bias, or scaling by the aspect ratio. This baseline offers impressive explanatory power for describing human fixation distributions, with high performance for classifying fixations in the 10 datasets analysed here and therefore offers an important component of any model of eye movement behaviour. Incorporating this module in models of scene viewing should produce models that generate fixation behaviour that is more like that generated by human observers.

It should be noted that while our recommended baseline is tested across ten datasets, drawn from a number of different tasks including free viewing, search, memorisation and scene description, it is important to validate this baseline against a wider variety of datasets in the future. There was some variation between the experimental setups across the 10 datasets, with differences in viewing distance, screen size, image resolution, image viewing angle and the use of chin/forehead stabilization. It is not clear whether these factors may themselves influence the nature of the image-independent biases. While there is variation between our datasets there is not sufficient variation to permit an exploration of this issue, but our *proposed baseline* offers a description of image-independent biases that works well over the range of setups analysed here. We expect that images with aspect ratios substantially different from 4:3 will require a different covariance matrix, as will images with non-canonical views of scenes. Indeed, viewing behaviour differs for 4:3 aspect ratio images presented with content shown at different orientations (Foulsham, Kingstone & Underwood, 2008) and placing natural scenes within a circular aperture reduces the prevalence of horizontal eye movements and increases the prevalence of vertical eye movements (Foulsham & Kingstone, 2010). Similarly, using dynamic scenes may reduce the influence of the screen centre on viewing (Cristino and Baddeley, 2009; 't Hart et al., 2009) and centre biases may not be a feature of viewing real world scenes ('t Hart et al., 2009; Tatler et al., 2011). However, it was not our goal to describe a baseline suitable to all experimental situations, but instead one that is suitable for the experimental setups most commonly used in the field: where images are displayed on computer monitors, with relatively small variations in angular extent, often in 4:3 aspect ratio or similar and most commonly using free viewing or search tasks. In the present work we have shown that in such situations a baseline that is not tailored to individual datasets (thus different sets of images and different tasks) or individual subjects performs as well as baselines that are fitted to each dataset or subject. This suggests that our recommended baseline is unlikely to be parochial to any particular image sets, individuals or tasks and so is likely to generalise to new datasets, and can serve as a suitable and easy to implement baseline for many experimental scene viewing datasets.

## Acknowledgments

The authors would like to thank Matthew Asher and Wolfgang Einhäuser for sharing their datasets. The support of the European Research Council under Award number 203427 Synchronous Linguistic and Visual Processing is gratefully acknowledged. We would like to thank Marc Pomplun and an anonymous reviewer for their comments and suggestions on a previous version of this manuscript. Finally, we thank the researchers who have made their eye movement datasets available publicly.

## References

Asher, M. F., Tolhurst, D. J., Troscianko, T., & Gilchrist, I. D. (2013). Regional effects of clutter on human target detection performance. *Journal of Vision, 13*(5), 25:1–15.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207.

Borji, A., Sihite, D. N., & Itti, L. (2013a). Objects do not predict fixations better than early saliency: A re-analysis of einhauser et al.'s data. *Journal of Vision, 13*(10), 18:1–4.

Borji, A., Sihite, D. N., & Itti, L. (2013b). analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing, 22*(1), 55–69.

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

Clarke, A. D. F., Coco, M. I., & Keller, F. (2013). The impact of attentional, linguistic, and visual features during object naming. *Frontiers in Psychology, 4*, 927.

Cristino, F., & Baddeley, R. (2009). The nature of the visual representations involved in eye movements when walking down the street. *Visual Cognition, 17*(6–7), 880–903.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition, 17*(6–7), 945–978.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8*(14), 18:1–26.

Foulsham, T., & Kingstone, A. (2010). Asymmetries in the direction of saccades during perception of scenes and fractals: Effects of image type and image features. *Vision Research, 50*(8), 779–795.

Foulsham, T., Kingstone, A., & Underwood, G. (2008). Turning the world around: Patterns in saccade direction vary with picture orientation. *Vision Research, 48*(17), 1777–1790.

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

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in Neural Information Processing Systems*, 545–552.

Henderson, J. M., Brockmole, J. R., & Castelhano, M. S. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Oxford, UK: Elsevier.

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

Judd, T., Durand, F., & Torralba, A. (2012). *A benchmark of computational models of saliency to predict human fixations* (Tech. Rep. MIT-CSAIL-TR-2012-001). Cambridge, MA: MIT Computer Science and Artificial Intelligence Laboratory.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision* (pp. 2106–2113).

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.

Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research, 46*(12), 1886–1900.

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

Spain, M., & Perona, P. (2011). Measuring and predicting object importance. *International Journal of Computer Vision, 91*(1), 59–76.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14), 4:1–17.

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

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision, 11*(5), 5:1–23.

Tatler, B. W., & Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition, 17*(6–7), 1029–1054.

't Hart, B. M., Vockeroth, J., Schumann, F., Bartl, K., Schneider, E., Konig, P., et al. (2009). Gaze allocation in natural stimuli: Comparing free exploration to head-fixed viewing conditions. *Visual Cognition, 17*(6–7), 1132–1158.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review, 113*(4), 766–786.

Tseng, P. H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*(7), 4:1–16.

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

Yun, K., Peng, Y., Samaras, D., Zelinsky, G., & Berg, T. (2013). Studying relationships between human gaze, description, and computer vision. In *2013 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 739–746), doi:http://dx.doi.org/10.1109/CVPR.2013.101.

Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision, 11*(3), 9:1–15.