Supplementary material for: Characterisation of field loss based on microperimetry is predictive of face recognition difficulties

Thomas S A Wallis *, †1,3,4, Christopher Patrick Taylor †1, Jennifer Wallis †2, Mary Lou Jackson2, and Peter J Bex1

1Schepens Eye Research Institute, Harvard Medical School
2Massachusetts Eye and Ear Infirmary, Harvard Medical School
3School of Psychology, The University of Western Australia
4Present address: Universität Tübingen, Germany

November 22, 2013

1 Supplementary materials

1.1 Ideal observer

Performance in our task can be compared to the performance of an ideal observer that performs the task optimally given the information available in the stimulus. The optimal

*Corresponding author: thomas.wallis@uni-tuebingen.de
†These authors contributed equally
decision rule for this task, assuming Gaussian contrast noise, is to compute the 2D (pixel wise) cross-correlation between each possible response face and the test face, then to select the response with the maximum cross correlation (Gold et al., 1999; Tjan et al., 1995). We estimated the performance of the ideal observer as a function of the signal-to-noise ratio (SNR; for Gaussian pixel noise added to the stimulus) by computer simulation (see Figure 1). For the stimulus conditions shown to observers (dashed vertical line), the ideal observer never made mistakes. That is, humans perform the task sub-optimally.
Figure 1: Face recognition performance for the ideal observer as a function of the ratio of signal contrast to Gaussian luminance noise contrast (SNR), for the unblurred stimulus and high frequency cutoffs of 32 and 16 cy / image. The performance of the ideal observer (open circles) was evaluated via Monte Carlo simulation. Test faces were presented with the same three blur levels as shown to human observers, at the same white noise contrast, and at 13 additional white noise contrasts in log-spaced steps. The ideal observer provided 1000 simulated trials at each combination of blur and noise contrast for a total of 14,000 trials per blur level. Solid lines represent the fits of a Weibull GLM (Knoblauch and Maloney, 2012). The grey vertical line marks the SNR shown to subjects. Average performance in each blur condition for human subjects are shown as filled points (x position is offset to aid readability). The error bar on these points shows the s.e.m between subjects in each group.

The ratio of ideal-to-human performance (“absolute efficiency”) is usually used to quantify this sub-optimality. We cannot compute this measure here since we did not measure full psychometric functions for human performance as a function of the SNR; we arbitrarily
selected an SNR to present to the human observers. That is, in Figure 1, the horizontal location of the human data is arbitrary and the slope of human performance as a function of contrast noise level is unknown.

However, some sense of the discrepancy between human and ideal can be gained from the ideal observer’s psychometric functions in Figure 1, as the ideal observer requires $\approx 2$ log units more noise contrast than humans to even begin to make mistakes, and around 4 log units to reach equivalent performance to humans. This is consistent with previous studies that have found that face recognition efficiency is extremely poor even in normally-sighted observers (Gold et al., 1999; Mäkelä et al., 2001; Melmoth et al., 2000; Näsinen, 1999). In contrast, relatively good efficiencies have been shown for other tasks, such as simple detection (Kersten, 1987; Taylor et al., 2009) and for extracting information about a data set from graphs (Legge et al., 1989). This comparison suggests that human face recognition in tasks such as ours must be limited mainly by constraints inherent to the visual system rather than by a lack of information in the stimulus.

By measuring the full psychometric functions for face recognition in patients with low vision as a function of SNR, future studies could determine the theoretical limit on performance improvements expected from image modification techniques (such as contrast enhancement) for a particular task. If humans are primarily limited by internal noise rather than by noise in the stimulus, it is possible that no amount of stimulus enhancement could effect performance improvement. Potentially, knowing the limits of internal noise could help direct low vision research to effectively target visual tasks in which improvements could be expected given particular stimulus changes or training.
1.2 Multilevel model and Bayesian inference

We fit the data with a multilevel model with three levels of parameters: subject, group and population. The trial-by-trial binary correct or incorrect response data $y$ for each subject $i$ in group $j$ were assumed to arise from a Bernoulli distribution with expected value $p$:

$$y_{ij} \sim \text{Bernoulli}(p_{ij})$$

where the $\sim$ symbol denotes “is distributed as”. The expected value $p$ was modelled as a logistic function with a lower asymptote $\gamma$, fixed at the chance performance level of 0.1:

$$p_{ij} = \gamma + (1 - \gamma) \times \text{logit}^{-1}(\eta_{ij})$$

where the inverse logit function is $1/(1 + \exp(-\eta))$. The linear predictor $\eta_{ij}$ for subject $i$ was given by a linear function of the intercept $a$ plus a slope coefficient of log high frequency cutoff $b$ plus a slope with block $c$ (to capture any linear effect of learning):

$$\eta_{ij} = a_{ij} + b_{ij} \log_{10}(\text{HFC}) + c_{ij}\text{block}$$

The three coefficients from each subject were assumed to arise from a group-level distribution:

$$a_{ij} \sim \text{Normal}(\mu_j, \sigma_j)$$

Since the same structure is used for all coefficients we simplify the notation here to only refer to one, but the model contained parameters $\mu_j$ and $\sigma_j$ for all coefficients. The group level mean and standard deviations are themselves conditional upon population-level

---

1 R code for reproducing this analysis is available from the first author’s Github page: https://github.com/tomwallis/microperimetry_faces
distributions:

\[
\begin{align*}
\mu_j & \sim \text{Normal}(\mu_\mu, \sigma_\mu) \\
\sigma_j & \sim \text{Gamma}(\alpha_\sigma, \beta_\sigma)
\end{align*}
\]

(5)

Here we use the subscripted letter to denote the group-level parameter that is conditional on the higher-level parameter. Gamma denotes a gamma distribution with shape \(\alpha_\sigma\) and scale \(\beta_\sigma\). A gamma distribution is used for the standard deviation parameter \(\sigma_j\) since standard deviation cannot be negative and the gamma distribution is only defined for non-negative values.

To estimate the posterior distribution over model parameters using Bayes’ rule, it is necessary to specify prior distributions on the model parameters. In this respect our entire model structure can be considered a “prior” (see Gelman and Hill, 2007): we believe it is logical to structure the model with individual subjects nested in groups, and groups nested in populations. Having defined this model structure in which low level parameters are conditional on higher-level parameters, we now must specify priors for the top-level hyperparameters (i.e. at the population level). The credible values for the lower levels of the model are then calculated based on the top-level priors and the data. We take the approach of using vague prior distributions (i.e. with large variance) reflecting a relatively agnostic stance on the model parameters. Since these distributions reflect large uncertainty about the true parameter values, they are quickly overwhelmed by the data causing them to have little influence on the posterior distribution (Kruschke, 2013). The model results presented in the paper are fit with the following priors on the population level:
\[ \mu_{\mu} \sim \text{Normal}(0, 3) \]
\[ \sigma_{\mu} \sim \text{Uniform}(0, 10) \]
\[ m_{\sigma} \sim \text{Uniform}(0, 10) \]
\[ sd_{\sigma} \sim \text{Uniform}(0, 10) \]

(6)

We specify priors on the mean \( m_{\sigma} \) and standard deviation \( sd_{\sigma} \) of the Gamma distribution for \( \sigma_j \) since it is more intuitive to place priors over the mean and deviation than over shape and scale (see for example Kruschke (2011, , Figure 9.8, page 170). These parameters are then reparameterised into shape and scale for the gamma distribution as:

\[ \alpha_{\sigma} = \frac{m_{\sigma}^2}{sd_{\sigma}^2} \]
\[ \beta_{\sigma} = \frac{m_{\sigma}}{sd_{\sigma}^2} \]

(7)

To summarise the model using the blur slope coefficient \( b \) as an example, the parameter \( b \) for subject \( i \) in group \( j \) is hypothesised to arise from a group normal distribution with mean \( \mu_j \) and standard deviation \( \sigma_j \). The mean of each group’s blur slope is conditional upon a single population normal distribution with mean \( \mu_{\mu} \) and standard deviation \( \sigma_{\mu} \), and the standard deviation of each group’s blur slope is conditional upon a single population gamma distribution with shape \( \alpha_{\sigma} \) and scale \( \beta_{\sigma} \). These population-level distributions are given weakly-informative priors. For MCMC sampling the model was parameterised slightly differently, with each model coefficient described as a shifted-and-scaled distribution of errors to increase sampling efficiency (Stan Development Team, 2013). For more details than can be described here, please see the analysis code on the first author’s Github page.

To understand why these priors represent vague values we must consider the scale of the model space. We tested three levels of high frequency cutoff (186, 32, and 16), and the model is fit to the logarithm of these (2.27, 1.50 and 1.20). The value of \( p \) (Equation 2) can range from 0.1 to 1. The value of \( \eta \) (given by Equation 3) is unbounded (i.e. from
negative to positive infinity), but for practical purposes resides in the range of -2.5 to 2.5 (which give expected values of 0.106 and 0.994). Therefore, for any given pairing of intercept $a$, blur slope $b$ and block slope $c$, the value of Equation 3 becomes relatively meaningless outside of the $+/- 2.5$ range, so the values of these parameters must produce an $\eta$ around this range. For example, the prior on $\mu$, for all parameters has a mean of zero, together corresponding to an $\eta$ of 0 and an expected performance level of 0.55 (including $\gamma$, the lower bound correction for chance performance). The standard deviation of 3 spreads most of the prior certainty over a range of -6 to +6 (i.e. 2 standard deviations around the mean of zero), which can be seen to over-span the useful range of the parameter space and therefore represents a relatively flat prior centred on zero.

This prior can be compared to the group-level posteriors in Figure 2. The posterior distributions for all groups fall within the range of -0.5 to 2, and the central tendency for three of the groups lie between values of 0.5 and 1. That is, the data has caused the posterior densities to become more certain (less variable) than the weak prior and also to shift away from the prior mean of zero for three of the groups, to the extent that the data suggest that the true mean is different. The uniform prior over variance parameter $\sigma$, in the range of 0 to 10 indicates that we are uncertain about the true variance between the groups on intercept $a$ and slope $b$. Group variance hyperpriors $m$ and $sd$ are given uniform priors over the range of 0 to 10. This selection of priors allows the variance between individuals in the group to take on a wide range, from individuals being similar to individuals being dissimilar.
Figure 2: Samples from the posterior distribution for intercept, slope of high frequency cutoff (blur), and slope of block (learning) parameters at the group level. Samples from the final MCMC chain are shown as coloured points, with mean and 95% credible intervals (HDIs) shown as black points and error bars.

Multilevel models have the desirable property of creating “shrinkage” between individual subject estimates. Since subjects are assumed to be drawn from a common population, estimates of their effects “share strength” which has the effect of reducing the influence of outliers on inference from the model fits (Gelman, 2006). Estimates of individual subject effects therefore lie between the completely pooled population estimates (fitting a single model to all data across subjects, or equivalently fixing $\sigma_\mu$ to be zero) and the within-subjects effects (fitting independent models to each subject). The degree of shrinkage is dictated by the data and the prior, since we estimate $\sigma_\mu$ from data: if subjects or groups show highly variable performance, the population-level parameter estimates will be very uncertain and there will be little shrinkage. If many subjects show similar estimates except for one, the estimate of that subject’s true mean will be strongly pulled towards the (otherwise highly certain) population estimate. This reduces the influence of outliers on inferences made on the population scores in a principled way (see Gelman et al., 2004; Kruschke, 2013, 2011, for
further examples).

Like any choice of statistical model, the hierarchy we hypothesise here along with the specific prior distributions we use can be regarded as “priors” (Gelman et al., 2004). How robust are our conclusions to the modelling choices we make? We have tested several alternative models and discuss one here. We examined a similar model to the one reported in the paper, but without the third (population) level hyperparameters. That is, the fits to each group are completely independent, and no shrinkage will occur between group estimates. The results of this fitting showed greater uncertainty for the group-level parameters compared to the 3-level model as expected, but the key results of the paper did not change substantively. Since we wish to draw inferences about human performances on our task, we believe the use of the 3-level model in the paper is justified. The top-level hyperparameter helps the data constrain credible values for model parameters in groups with less data, as well as drawing all groups towards the grand mean where less data are available, theoretically producing more robust predictions. We believe this choice should be acceptable to a skeptical reader, since it is far from trivially presuming our result by specifying it in our prior: the prior assumes that all the groups are the same, with large uncertainty; the group differences we observe in the posterior are purely driven by the data.

1.2.1 Bayesian correlation coefficients

To compute Bayesian correlation coefficients we estimate the posterior over the correlation matrix between the five predictor variables. We first compute the rank orders of the data matrix, then normalise these ranks (z-score) to compute the Spearman rank order correlation coefficient. These data are considered to arise from a multivariate normal distribution with a mean of zero and a covariance matrix that is estimated from the data and the prior. To encourage conservative inference, we place weak a priori credibility on correlations of zero using the following correlation matrix prior distribution (see Stan Development Team, 2013,
where $\alpha$ is a shape parameter and $\Sigma$ is a positive definite symmetric matrix with unit diagonal (i.e. a correlation matrix). When $\alpha$ is one, the prior has uniform density (i.e. all correlation matrices are equally plausible). As $\alpha$ approaches infinity, the prior becomes the identity matrix (i.e. all correlations are zero). We set $\alpha$ to be 2, which is a weak prior that represents the belief that zero correlations are more plausible than strong correlations, but with high uncertainty.

Estimating the correlation coefficients in this way did not produce drastically different conclusions from traditional statistical estimation. In a bootstrapping regime we found that visual acuity and fixation stability were significantly associated with performance (Spearman’s $\rho$), consistent with the conclusions of our Bayesian approach. In addition, the association between fixation stability and acuity was also significant, where it did overlap zero in our method. We believe our method is the more conservative approach, since bootstrapping estimates can be unstable with such few data.

### 1.3 No differences evident for blur slope and block slope

There were no credible differences between the groups in either blur slope or in learning over blocks (see Figure 3).
Figure 3: Mean and HDIs of the distributions of difference scores between group level parameters for blur slope and learning slope. No credible differences are observed (all distributions overlap zero).

1.4 Correlation between acuity and fixation stability

It is possible that the correlation between acuity and fixation stability we observe is driven by the use of diverse patient populations, and that within each group there is little evidence for a correlation.

Below is a scatterplot of these data. It is extremely difficult to draw conclusions about this due to the limited group size, however it is entirely possible that little correlation exists within the LV group. Note that LV:F patients were selected partly on the basis of their good fixation stability, and so the lack of relationship here is by design.
Figure 4: Scatterplot of fixation stability versus letter acuity. The LV group is shown as white squares, the LV:F group as grey diamonds.
References


