Integration of Facial Information is Sub-Optimal

Jason M. Gold (jgold@indiana.edu)
Departments of Psychological and Brain Sciences and Cognitive Science, Indiana University, 1101 East 10th Street
Bloomington, IN 47405 USA

Bosco S. Tjan (btjan@usc.edu)
Department of Psychology, University of Southern California, SGM 501
Los Angeles, CA 90089 USA

Megan Shotts (mshotts@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University, 1101 East 10th Street
Bloomington, IN 47405 USA

Abstract

How efficiently do we combine information across facial features when recognizing a face? Previous studies have suggested that the perception of a face is not simply the result of an independent analysis of individual facial features, but instead involves a coding of the relationships amongst features. This additional coding of the relationships amongst features is thought to enhance our ability to recognize a face. In our experiments, we tested whether an observer's ability to recognize a face is in fact better than what one would expect from their ability to recognize the individual facial features in isolation. We tested this by using a psychophysical summation-at-threshold technique that has been used extensively to measure how efficiently observers integrate information across spatial locations and spatial frequencies. Surprisingly, we found that observers integrated information across facial features less efficiently than would be predicted by their ability to recognize the individual parts.

Keywords: Face Recognition; Ideal Observer; Information Integration.

Introduction

The ability to accurately recognize human faces is vitally important to human social interactions. As such, there has been a great deal of interest in exploring the psychological and neurophysiological mechanisms that mediate human face recognition. Much of this research has focused on determining whether the individual elements in a face (e.g., eyes, nose, mouth) are processed independently or if the relationships amongst the elements are also encoded in the facial representation.

Several lines of evidence are consistent with the idea that the spatial relationships amongst facial features play a crucial role in face identification. Some of this evidence draws upon the "face inversion effect": the finding that, unlike most other objects, faces tend to be much more difficult to identify when they are inverted than when they are upright (Maurer, Grand, & Mondloch, 2002; Valentine, 1988; Yin, 1969). This effect is typically accounted for by positing that upright faces are processed as single units with the spatial relationships amongst elements encoded in the representation, whereas the elements of inverted faces are processed independently. As a result, the extra information that is encoded for upright faces allows an observer to identify an upright face more quickly and accurately than an inverted face. Other experiments (Tanaka & Farah, 1993) have found that observers are more accurate at identifying facial features (e.g., a nose) within the context of a normal face than either in isolation or in the context of a face whose features have been spatially scrambled.

The results of these experiments suggest that observers benefit from the spatial arrangement of features within a face in a way that would not be predicted by their ability to recognize the individual features in isolation. In our experiments, we wished to directly test this idea by measuring how efficiently observers integrate information across features in a face identification task. The technique we used is based on a summation-at-threshold method developed previously to measure the efficiency of information integration across spatial and spatial frequency tuned analyzers (Graham, 1989; Graham, Robson, & Nachmias, 1978; Nandy & Tjan, 2008). In our experiments, we measured observers' contrast sensitivities (i.e., the reciprocal of contrast threshold) for identifying facial features (i.e., left eye, right eye, nose, mouth) either in isolation or all together in combination. Optimal information integration predicts that an observer's squared contrast sensitivity to features when shown in combination should be the same as the sum of their squared contrast sensitivities to the individual features when shown in isolation (i.e., the ratio of the combined squared contrast sensitivity to the sum of the individual squared contrast sensitivities should be equal to one). Sub-optimal integration predicts this ratio should be less than one, and super-optimal integration predicts this ratio should be greater than one (Nandy & Tjan, 2008).

Based on the results of previous experiments with faces, we would expect to see super-optimal integration, because observers are thought to derive an additional benefit from the use of the relationships amongst features when they are shown together in combination as opposed to when they are shown in isolation. Contrary to this prediction, we found that most observers integrated information sub-optimally, in

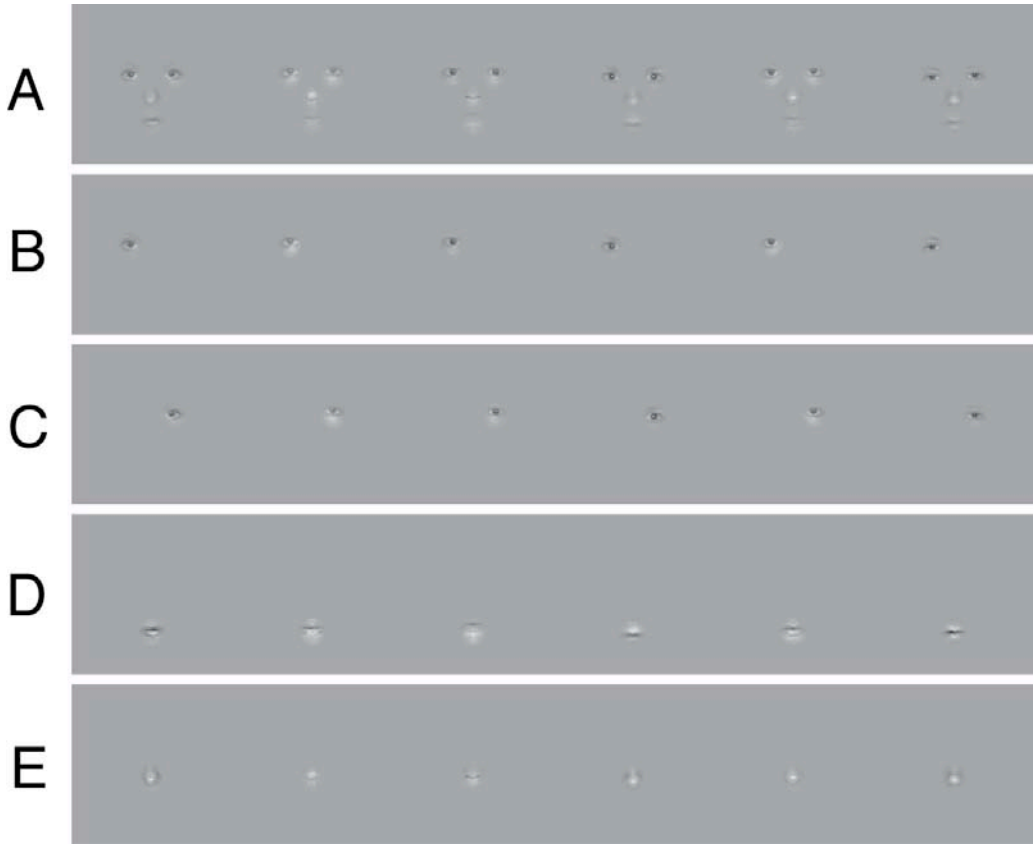

Figure 1: Stimuli used in the face identification experiments. All stimuli were based on the six combined faces shown in row A. Row B shows the left eye stimuli, row C the right eye stimuli, row D the mouth stimuli and row E the nose stimuli.

a fashion more in line with basing their decision on the individual feature to which they are most sensitive.

Methods

Subjects

Six subjects (three women and three men, aged 18-37) participated in this study. All subjects except for two authors (MS and JMG) were naive to the purposes of the experiment. All had normal or corrected-to-normal visual acuity. Each naive subject was paid for his or her participation. All subjects had given their informed consent.

Stimuli

Stimuli were modified from a set used in previous experiments on human face recognition (Gold, Bennett, & Sekuler, 1999a, 1999b). Six grayscale faces were used (three male, three female). Each face was 256 x 256 pixels in size (2.5° x 2.5°, from a viewing distance of 130 cm), and was multiplied by a set of four Gaussian windows (σ = 9 pixels), each centered on a different facial feature (left eye, right eye, nose and mouth). These windows were used to isolate the facial features. The locations of the four windows were the same across all the faces, and they were chosen to insure that a given feature fell within a given window for each face (see Figure 1a).

The value of each pixel in each image was expressed in terms of contrast, where contrast is defined as $(L_{pix} - L_{bg})/L_{bg}$, where L_{pix} is the luminance of a given pixel and L_{bg} is the background luminance. The regions of the faces not falling within the Gaussian windows were set to zero contrast (i.e., L_{bg}).

Four additional sets of six face images were generated from this first set of images. One set contained only the left eyes of each face (Figure 1b); a second set only the right eyes (Figure 1c); a third set only the noses (Figure 1d); and a fourth set only the mouths (Figure 1e). In each of the images that contained only a single feature, the regions where the other features appeared in the original images were set to zero contrast.

White Gaussian contrast noise was added to each pixel of the image that was shown on each trial (μ = 0; σ = 0.1). A unique sample of noise was generated for each pixel on each trial.

Procedure

A one-of-six identification task was used to estimate identification thresholds for each feature condition (left eye, right eye, nose, mouth, combined). The contrast of the

images was manipulated across trials according to a 1-down 1-up staircase in each condition to obtain an observer's 50% correct identification threshold (chance performance was ~16% correct). The staircases were randomly interleaved during each experimental session, which meant that the stimulus types were also randomly mixed within each session. On each trial, the observer saw a noisy stimulus and was presented with the set of six noise-free images from which the noisy image had been chosen (e.g., if a left eye had been shown, the six possible left eye images were shown in a selection screen for the observer to choose from).

On each trial, a box appeared around the region where the stimulus was going to appear. The observer started the trial with a mouse click, and the noisy image was shown for ~500 ms, after which the image was replaced with a selection window. The observer used the mouse to click on the image they thought had appeared in the stimulus interval. Accuracy feedback was given in the form of a high or low beep.

Each observer participated in five sessions of 500 trials. The first two sessions were not included in the analyses to remove any initial learning effects from the data. A Weibull psychometric function was fit to the staircase data (i.e., a fit to percent correct as a function of stimulus contrast) in each condition and the 50% correct identification threshold was estimated from each fit. Bootstrap simulations were used to estimated confidence intervals for the thresholds.

Integration Index

Following Nandy & Tjan (2008), an integration index Φ was defined as follows:

$$\Phi = \frac{CS^2_{\textit{left eye + right eye + mouth + nose}}}{CS^2_{\textit{left eye}} + CS^2_{\textit{right eye}} + CS^2_{\textit{mouth}} + CS^2_{\textit{nose}}}$$

where c is an observer's contrast threshold and CS, an observer's sensitivity, is equal to $1/c$. Nandy and Tjan (2008) have shown that the integration index for a statistically optimal observer in this task will be equal to 1. Sub-optimal integration will yield an index less than 1, and super-optimal integration will yield an integration index greater than 1. Note that only a sub-ideal observer can actually achieve super-optimal integration: an integration index greater than 1 implies additional information is used when identifying the composite that is *not* used when identifying the individual elements in isolation.

Results

Figure 2a shows contrast sensitivities in each condition for the ideal observer[1] (dashed line) and three human observers[2] (solid lines with symbols). Figure 2b shows the corresponding integration index for each observer. Figure 2b also shows the predictions of the "best-feature" model (where each feature is analyzed independently and the decision is based on the feature to which the observer is most sensitive[3]), plotted as a dashed line with triangle symbols. The best-feature model is intended to provide a lower-bound for performance.

There are several interesting things to note about these data. First, the pattern of sensitivities across conditions is similar for all the observers, including the ideal observer. The fact that the ideal observer shows a similar pattern of performance to the human observers is interesting, because it indicates that the variations in human performance across conditions can largely be accounted for by the amount of physically available information in each set of stimuli. In this case, it shows that performance was worse for mouths and noses in isolation largely because there was simply less information physically present in those conditions (i.e., the stimuli were more physically similar to each other).

Second, the integration index for two of the human observers was significantly less than 1, indicating they were integrating information sub-optimally in the combined condition. In fact, these two observers were closer to the predictions of the best-feature model than optimal or super-optimal integration. This result is the opposite of what one would predict if observers were using the relationships amongst features to improve their performance when facial features are shown together rather than in isolation. Such a result is surprising, given that previous experiments on face recognition have suggested observers greatly benefit from using the relationships amongst facial features when recognizing faces. The one exception was observer VMD, who integrated information super-optimally. Apparently, this observer did in fact derive an additional benefit from viewing the facial features together rather than in isolation.

It is possible that the sub-optimal integration we found for two of the three human observers was somehow related to their conditional uncertainty in the experiment (recall that they did not know which condition would appear on each trial). For example, the conditional uncertainty may have

[1] The ideal decision rule can be derived using Bayes' rule (Green & Swets, 1966). For our task and stimuli, it is equivalent to choosing the noise-free signal that produces the highest cross-correlation with the noisy stimulus (Tjan, Braje, Legge, & Kersten, 1995).

[2] One additional human observer was excluded due to an inability to perform the task at a level sufficiently above chance.

[3] The "best-feature" model is closely related to the model of probability summation for signal detection (Graham, 1989). The integration index of the best-feature model is computed as the expected value of the maximum of the observer's squared contrast sensitivities to the individual face features to that of the sum of the squared contrast sensitivities to all of the face features (Nandy & Tjan, 2008), i.e.

$$\Phi_{\textit{best-feature}} = \frac{\left\langle \max[CS^2_{\textit{left eye}}, CS^2_{\textit{right eye}}, CS^2_{\textit{mouth}}, CS^2_{\textit{nose}}] \right\rangle}{CS^2_{\textit{left eye}} + CS^2_{\textit{right eye}} + CS^2_{\textit{mouth}} + CS^2_{\textit{nose}}}$$

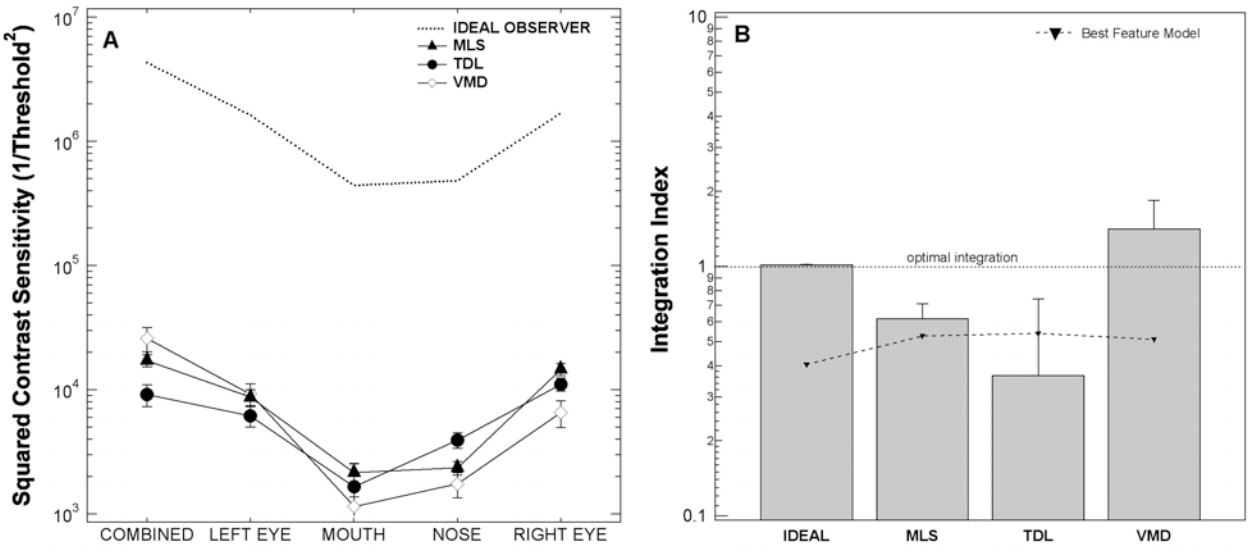

Figure 2: (A) Squared contrast sensitivities and (B) integration indexes for the ideal observer and three human observers. Error bars correspond to ±1 s.e., estimated by bootstrap simulations (Efron & Tibshirani, 1993).

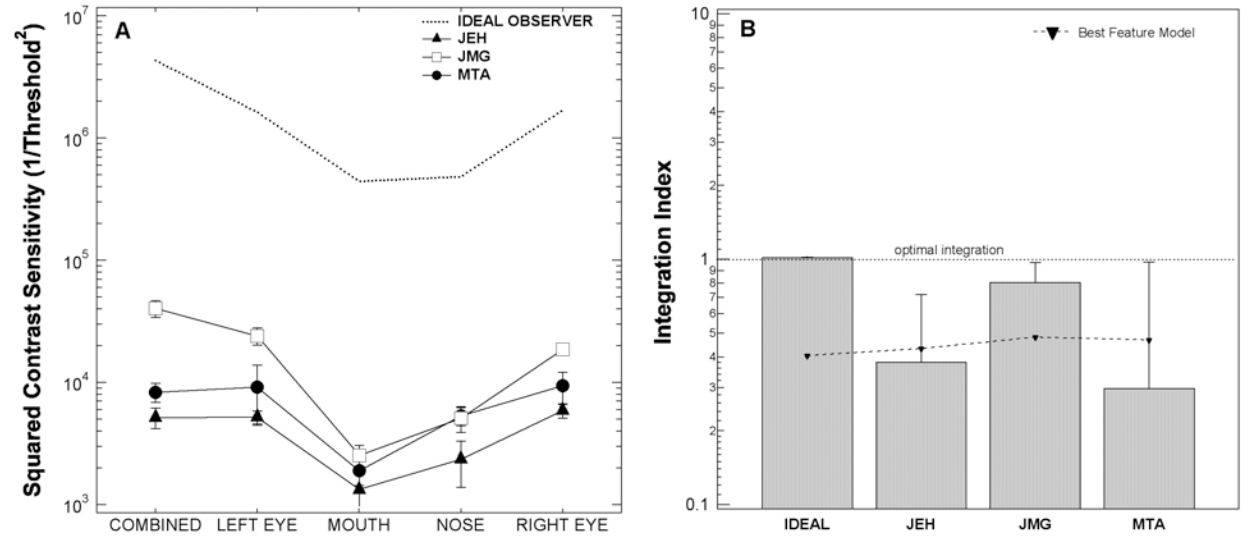

Figure 3: (A) Squared contrast sensitivities and (B) integration indexes for the ideal observer and three human observers. The conditions in this experiment were presented in blocks of 50 trials rather than randomly permuted as in the first experiment (Figure 2). Error bars correspond to ± 1 s.e.

induced observers to use an individual feature-oriented strategy on all trials, including those where the entire set of features was present (i.e., the combined condition). We tested this possibility by running a second set of observers through the same experiment, but with the conditions blocked rather than randomized. Specifically, each session contained a series of blocks, where the same condition was tested within each block for 50 consecutive trials. The order of conditions was randomized across blocks within each session and there were two blocks tested for each condition (a total 100 trials per condition within each session, for 5 sessions). Importantly, the observers were told at the beginning of each new block which condition they would be tested on for the next 50 trials.

The results of this experiment are shown in Figure 3. Figure 3a shows contrast sensitivities in each condition for

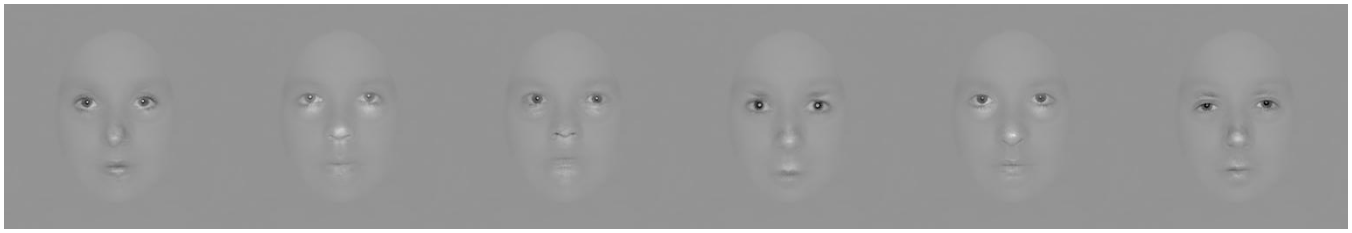

Figure 4: Face images from taken from Figure 1A, but with a constant background image added to each face. The area surrounding the features (the background image) is an average computed from the six original faces.

the ideal observer and three human observers. Figure 3b shows the corresponding integration index for each observer and the predictions of the best-feature model. Contrary to the idea that conditional uncertainty was responsible for the sub-optimal integration found in the first experiment, these data show that observers were generally *less* efficient at integrating information when the conditions were blocked rather than randomized. Two observers (JEH, MTA) were actually numerically *worse* than the lower bound set by the best-feature model[4].

Discussion

Our experiments were designed to test the prediction that observers make use of information from facial features when recognizing a complete face in a manner that is better than one would predict from their ability to detect the individual features in isolation. Contrary to this prediction, our results are more consistent with the idea that observers analyze each feature independently and base their decision on the single feature to which they are most sensitive.

Previous experiments with much simpler tasks and stimuli, such as the detection of sinusoidal gratings across space, have yielded results similar to our own (and have referred to this as "probability summation") (Graham, 1989). Pelli, Farell and Moore (2003) have also found that observers combine information sub-optimally across letters when recognizing English words. Taken together, these results suggest that the sub-optimal integration of information across facial features that we observed in our experiments may reflect a more general inefficiency in visual spatial information integration. One possible account for this effect could be that such tasks push against an upper limit on the processing capacity of visual spatial attention (Driver, 2001). Such a limitation could reduce the amount of information an observer is able to use at any given feature location when they are forced to simultaneously attend to more than one spatial feature at a time (as in the case of a composite face). If so, it is possible that observers do make use of relational properties when recognizing faces, but that the limitations imposed by spatial attention reduce the processing efficiency of individual features more than is gained by the use of relational or other second-order coding strategies.

One way to directly address the issue of limited spatial attention would be measure 'classification images' for each of the conditions in our experiments (Ahumada, 2002; Murray, Bennett, & Sekuler, 2002). A classification image is a spatial map that describes the relative weight given to each image location by an observer over the course of an experiment. Classification images are measured by correlating random pixel noise with an observer's decisions across trials. The efficiency of an observer's classification image can be measured by comparing their classification image with that of an ideal observer (Murray, Bennett, & Sekuler, 2005). Measurement of human and ideal classification images for individual and combined facial features would allow us to a) directly determine the efficiency of a human observer's weighting when the features are shown in combination vs. in isolation; and b) reveal the specific nature of any differences in weighting when the features are shown in combination vs. in isolation.

A second less direct way to address the issue of limited spatial attention would be to carry out our experiment with inverted facial features, where the attentional bottleneck would be identical to our original task. If observers rely on relational codes when identifying normal faces but not inverted faces (as previous experiments would suggest), we would expect integration efficiency to be higher for normal than inverted faces..

It is worth noting that inefficiencies similar to those we obtained with faces are also found with respect to the integration of information across spatial frequencies with simple compound grating detection tasks (Graham, 1989). However, recent experiments by Nandy and Tjan (2008) have found that observers optimally integrate information across spatial frequencies when identifying English letters. One obvious difference between spatial frequency integration and spatial integration is that stimuli filtered with respect to spatial frequency will occupy the same region of space. It would be interesting to see if, like letters, spatial frequency information in faces is integrated in an optimal fashion.

[4] The other observer in this second experiment (JMG) was an author and, unlike all the other observers, was very familiar with original face stimuli. The integration index for this observer was much higher than the other two observers in the second experiment, albeit still sub-optimal. This suggests the possibility that information may be integrated less efficiently when recognizing unfamiliar faces, and that training may serve to increase information integration efficiency.

One additional factor that may have contributed to the inefficient processing of composite faces that we observed in our experiments is the relatively unnatural viewing of faces through a set of Gaussian windows. That is, viewing features through a set of Gaussian windows may have disrupted any relational processing that normally takes place when recognizing a face. One way to address this issue would be to place the features within a fixed 'average' background image, as shown in Figure 4. In this figure, the Gaussian windowed faces from Figure 1A have been added to an image that was generated by averaging the regions surrounding the Gaussian windows in each of the six original face images. If the sub-optimal integration across features in our experiments was due to a lack of facial 'context' around the Gaussian windows, we would expect the faces shown in Figure 4 to greatly increase integration efficiency.

Conclusions

In this paper, we have reported the results of two experiments that indicate human observers are inefficient at integrating information across facial features. We have suggested several possible accounts for our results, including limits on spatial attention and the use of unnatural face stimuli. We are currently conducting experiments to test each of these possibilities.

Acknowledgments

This research was funded by National Institute of Health Grants EY019265 to J.M.G., and EY016093, EY017707 to B.S.T.

References

Ahumada, A. J., Jr. (2002). Classification image weights and internal noise level estimation. *J Vis, 2*(1), 121-131.

Driver, J. (2001). A selective review of selective attention research from the past century. *Br J Psychol, 92*(Pt 1), 53-78.

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

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999a). Identification of band-pass filtered letters and faces by human and ideal observers. *Vision Res, 39*(21), 3537-3560.

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999b). Signal but not noise changes with perceptual learning. *Nature, 402*(6758), 176-178.

Graham, N. (1989). *Visual Pattern Analyzers*. New York: Oxford University Press.

Graham, N., Robson, J. G., & Nachmias, J. (1978). Grating summation in fovea and periphery. *Vision Res, 18*(7), 815-825.

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

Maurer, D., Grand, R. L., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends Cogn Sci, 6*(6), 255-260.

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2002). Optimal methods for calculating classification images: weighted sums. *J Vis, 2*(1), 79-104.

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2005). Classification images predict absolute efficiency. *J Vis, 5*(2), 139-149.

Nandy, A. S., & Tjan, B. S. (2008). Efficient integration across spatial frequencies for letter identification in foveal and peripheral vision. *J Vis, 8*(13), 3 1-20.

Pelli, D. G., Farell, B., & Moore, D. C. (2003). The remarkable inefficiency of word recognition. *Nature, 423*(6941), 752-756.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Q J Exp Psychol A, 46*(2), 225-245.

Tjan, B. S., Braje, W. L., Legge, G. E., & Kersten, D. (1995). Human efficiency for recognizing 3-D objects in luminance noise. *Vision Res, 35*(21), 3053-3069.

Valentine, T. (1988). Upside-down faces: a review of the effect of inversion upon face recognition. *Br J Psychol, 79 (Pt 4)*, 471-491.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*, 141-145.