ELSEVIER

Cognitive Science 28 (2004) 209–226

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Making the ineffable explicit: estimating the information employed for face classifications

Michael C. Mangini*, Irving Biederman

Department of Psychology, University of Southern California, 3641 Watt Way, HNB 2520, Los Angeles, CA 90089-2520, USA

Received 14 October 2002; received in revised form 29 October 2003; accepted 4 November 2003

Abstract

When we look at a face, we readily perceive that person's gender, expression, identity, age, and attractiveness. Perceivers as well as scientists have hitherto had little success in articulating just what information we are employing to achieve these subjectively immediate and effortless classifications. We describe here a method that estimates that information. Observers classified faces in high levels of visual noise as male or female (in a gender task), happy/unhappy (in an expression task), or Tom Cruise/John Travolta (in an individuation task). They were unaware that the underlying face (which was midway between each of the classes) was identical throughout a task, with only the noise rendering it more like one category value or the other. The difference between the average of noise patterns for each classification decision provided a linear estimate of the information mediating these classifications. When the noise was combined with the underlying face, the resultant images appeared to be excellent prototypes of their respective classes. Other methods of estimating the information employed in complex classification have relied on judgments of exemplars of a class or tests of experimenter-defined hypotheses about the class information. Our method allows an estimate, however subtle, of what is in the subject's (rather than the experimenter's) head.
© 2004 Cognitive Science Society, Inc. All rights reserved.

Keywords: Face classification; Face recognition; Reverse correlation; Classification image; Response classification

* Corresponding author. Present address: Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Ave., NE 20-443, Cambridge, MA 02139, USA.

E-mail addresses: mangini@mit.edu (M.C. Mangini), bieder@usc.edu (I. Biederman).

0364-0213/$ – see front matter © 2004 Cognitive Science Society, Inc. All rights reserved.
doi:10.1016/j.cogsci.2003.11.004

1. Introduction and background

What is the information mediating face classifications? Although observers can articulate individual features which often, in fact, correlate with images of such classes, such as the down turned mouth in an unhappy face, these descriptions are often incomplete and fail to capture the subtleties. For example, people can readily recognize celebrities but if asked to describe the difference between Tom Cruise's and John Travolta's faces, they typically resort to naming individual features, such as Travolta's cleft chin, which fails to capture the rich differences in their appearance. Although readily perceived, such differences remain ineffable.

We addressed this problem of determining the information used by a perceiver of a face through an extension of the reverse-correlation, classification-image technique originally proposed for low-level psychophysics by Ahumada (Ahumada & Lovell, 1971). Early implementations of the technique investigated low-level visual processes, for example, Vernier acuity (Ahumada, 1996), letter discrimination (Watson, 1998), and Gabor detection (Ahumada & Beard, 1999; Solomon & Morgan, 1999). As evidenced by this current special issue there has recently been an increase in exploring what such methods can tell us about higher level visual processes. For example, Gold and colleagues have used the classification-image method to investigate illusory contours (Gold, Murray, Bennett, & Sekuler, 2000), and face identification (Sekuler, Gold, & Bennett, 2000).

In our method, participants classified images of noisy faces (in different blocks of trials) as male or female, happy or unhappy, or Tom Cruise or John Travolta. In all cases, the underlying (or base) image never changed and thus provided no information as to the classification. The addition of the noise (in the form of 4092 truncated sinusoids at five scales and six orientations whose coefficients varied randomly from trial to trial) sometimes produced a classification of the face as male on some trials and as female on other trials, in the gender classification task, for example. (The subject was forced to make a classification.) We averaged the noise on the trials where the image was classified as female separately from the trials where the image was classified as male. A "classification image" was computed by taking the difference between the two average noise composites thus yielding the noise components that served to distinguish the two classes. When the classification image was added to the base face it produced an image that appeared female (or happy or Tom Cruise) and when subtracted from the base face produced an image that appeared male (or unhappy or John Travolta). The major contribution of the method is that the classification image is produced solely by the way in which the subject classified the noise + (the constant) base image and thus provides an estimate of the information in the subject's *head* determining his or her response.

Why did we bother to estimate the information throughout the entire face using this reverse-correlation technique? Why not simply test specific hypotheses for the different classes by, for example, showing an upturned mouth on some trials and a down turned mouth on other trials to assess whether such a cue is relevant for expression? Although such specific tests can, in many cases, provide information that is, indeed, relevant to a particular classification, they assess only a tiny subset of the full information available in the face for a particular classification. Worse, one is limited to a specific hypothesis that the experimenter might have as to the difference (and the magnitude of the difference) between the two classes. Aside from Travolta's cleft chin, what other features should be manipulated when distinguishing him from Cruise? In general, our goal

is to find out what is in the subjects' heads, rather than the experimenters' heads—and we wish to understand this information in its full complexity, rather than limited to a few local features.

A number of computer vision scientists have studied the full complexity of faces, determining, for example, efficient representations of faces as in Turk and Pentland's (1991) Eigenface computation or in various networks trained to distinguish male from female faces (e.g., Valentin, Abdi, & O'Toole, 1994) or different expressions (e.g., Calder, Burton, Miller, Young, & Akamatsu, 2000; Dailey, Cottrell, Padgett, & Adolphs, 2002). Although such networks can discover the pixel values distinguishing one class from another, one doesn't know if this information corresponds to what people (or primates) actually use. The essential point here is that the information distinguishing the classes for these techniques is calculated from *pictures* rather than from perceptual classifications.

This distinction is also relevant to the difference between our technique and Gosselin and Schyns' (2001) Bubbles technique (presented in detail in the current issue), which were developed independently. Bubbles employs reverse correlation to estimate informative areas/scales used by subjects. A major distinction between the two methodologies is that with Bubbles the underlying face *does* contain information relevant to the category, as it is a partially revealed male or female face for a gender classification task, for example. The regions of the face are only revealed at the loci and scale of Gaussians. The averaging of the Bubbles that lead to criterion performance reveals the regions and scale for distinguishing among a particular set of male and female *pictures*. Our technique uses not informative instances of the category but, as noted earlier, a constant image midway between the two classes. There is thus no distinguishing information in the underlying (or "base" image) so all the distinguishing informations come from the subjects' heads, not the images. Gosselin and Schyns (2003) recognized this distinction terming such an approach "superstitious perceptions" and suggested that such methods provide a pure measure of internal classifiers.

2. Experiment

2.1. Methods

2.1.1. Participants

For each of the three tasks a group of 36 undergraduate students at the University of Southern California participated for extra-credit in Psychology courses. All subjects reported normal or corrected-to-normal vision and were naïve to the purpose of the experiment.

2.1.2. Stimuli and apparatus

All trials in each task consisted of a presentation of the *base image* in high noise. The stimuli were 8-bit grayscale, 128 × 128 pixel images presented on a gamma corrected Sony Trinitron 20 in. monitor at a distance of 1 m. At 640 × 480 resolution the stimuli subtended a visual angle of approximately 4.5°. To construct the base image for the gender and expression tasks, 20 digital photographs were taken of 10 males and 10 females each posed in a frontal view with a neutral expression under controlled lighting conditions with a digital camera. Models wore a bathing cap to conceal their hairlines and no jewelry, facial hair, or obvious makeup was

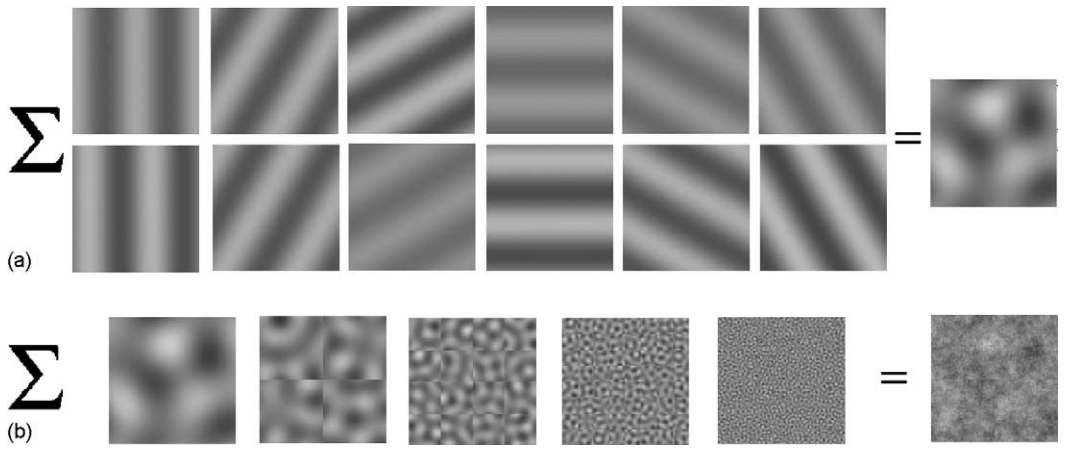

Fig. 1. Creating the noise. (a) Six orientations and two phases (cosine upper row; sine lower row) of random amplitude sinusoids are summed to create the 2 cycles/image noise. The same process is repeated four times for each "tile" of the 4 cycles/image noise, 16 times for the 8 cycles/image noise, etc. (b) All five octaves are summed to create the noise pattern. To the right are shown the 2, 4, 8, 16, and 32 cycles/image noise patterns of all orientations.

visible in photographs. Image morphing software (Gryphon Morph 2.5) was used to morph each model to, two arbitrarily chosen models of the same gender, resulting in a total of 200 images. Taking the mathematical average of these 200 faces resulted in the base image.[1] For the individuation task, a morph of images of John Travolta and Tom Cruise was used as the base image. The base images are shown in the first column of Fig. 3.

For each trial a random noise stimulus was generated. Although most experiments using the response classification technique have used Gaussian white pixel noise, our noise was composed of truncated sinusoids that are nominally localized in space, frequency, orientation, and phase.[2] Each sinusoid consisted of two cycles of a sine wave in a square envelope. Sinusoidal patches at five octave scales (2, 4, 8, 16, and 32 cycles/image), six orientations (0, 30, 60, 90, 120, and 150°), and two phases (0 and $\pi/2$) were summed to create one noise pattern (Fig. 1). The resulting noise pattern was described by 4096 parameters that corresponded to the amplitudes of the sinusoids. The amplitude parameters were selected randomly from a zero mean uniform distribution. The separate scales were multiplied by constant factors that insured that the range of each scale did not exceed the dynamic range available for the noise. These factors also contributed to the frequency profile presented in Fig. 5. A justification of why this noise was chosen is presented in Section 3.

Image creation and presentation was performed on a Macintosh G4 computer using Matlab with the Psychophysics and VideoToolbox extensions (Brainard, 1997; Pelli, 1997).

2.1.3. Procedure

Subjects were assigned to one of three discrimination tasks: gender, expression (happy versus unhappy), or identity (John Travolta versus Tom Cruise). At the beginning of the experiment the subjects were told that they would be making a simple discrimination but that it would be difficult to perform. For the Gender and Expression tasks subjects were told that a set of

faces were created that changed in subtle ways, and that the noise would make it difficult to discriminate between the categories. The subjects in the identity tasks were shown images of Steve Martin and Jay Leno that were warped to the same geometry (i.e., the eyes, nose, mouth, hairline, chin, etc. were in identical pixel coordinates and occupied identical areas). They were then instructed that they would be performing a task to discriminate between John Travolta and Tom Cruise, and that the images of these celebrities had been treated in a similar fashion. The subjects were not shown any images of Cruise or Travolta. The subjects were told that this warping manipulation in combination with the visual noise would make the task very difficult. In both cases these instructions were deceptive for, in fact there was no information in the base image that could inform the subject's decision.

The subjects were instructed to provide one of four responses on the computer keyboard for each trial: probably Travolta, possibly Travolta, possibly Cruise, or probably Cruise (and likewise for male/female and happy/unhappy for subjects in the gender and expression tasks). Reaction times were not collected.

After the brief instruction, noisy images were presented one at a time in the center of the screen for 1 s after which the screen cleared until the subject made his response. Fig. 2 provides a sample of noisy images that could have been presented during the gender and expression tasks. Each subject categorized 390 noisy face images. No feedback was provided, as there were no *correct* or *incorrect* responses. The subject could avail himself of two breaks, one after the 130th trial and the other after 260th trial. The entire experiment lasted approximately 35 min.

2.2. Results and discussion

Because the base image was identical in every presentation for a given subject, any systematic changes in subjects' classifications from trial to trial could be directly attributed to the noise. On each of the 390 trials the 4092 parameters that defined a noise stimulus for the trial was assigned to one of four categories based on the subject's response. The classification image was calculated by subtracting the mean of the parameters that elicited high confidence responses for one category from the mean of the parameters that elicited high confidence responses for the opposite category. High confidence patterns were those images to which subjects responded were "probably" in the category as opposed to just "possibly" in the category. Murray, Bennett, and Sekuler (2002) have shown that collecting confidence ratings and appropriately weighting the confidence intervals can increase the signal-to-noise ratio (SNR) of the resultant classification image. The maximum SNRs achieved were at points where the subject was highly confident about a false alarm response. Low confidence responses in their study had very low SNRs. We chose to perform all of the presented analyses utilizing only the high confidence responses.[3] Subjects responded with high confidence on 38, 41, and 39% of the trials in the expression, gender, and identity tasks, respectively.

The resultant classification images calculated on all the subjects' data are presented in Column 2 of Fig. 3. Adding or subtracting the classification image to or from the base image resulted in faces that appear to be good prototypes of their classes as shown in Columns 3 and 4. This appears to be true not only of the group data but the data for individual subjects as well, as shown in the rightmost two columns of Fig. 3 which shows the results for a median subject in terms of Euclidean pixel distance from the mean.

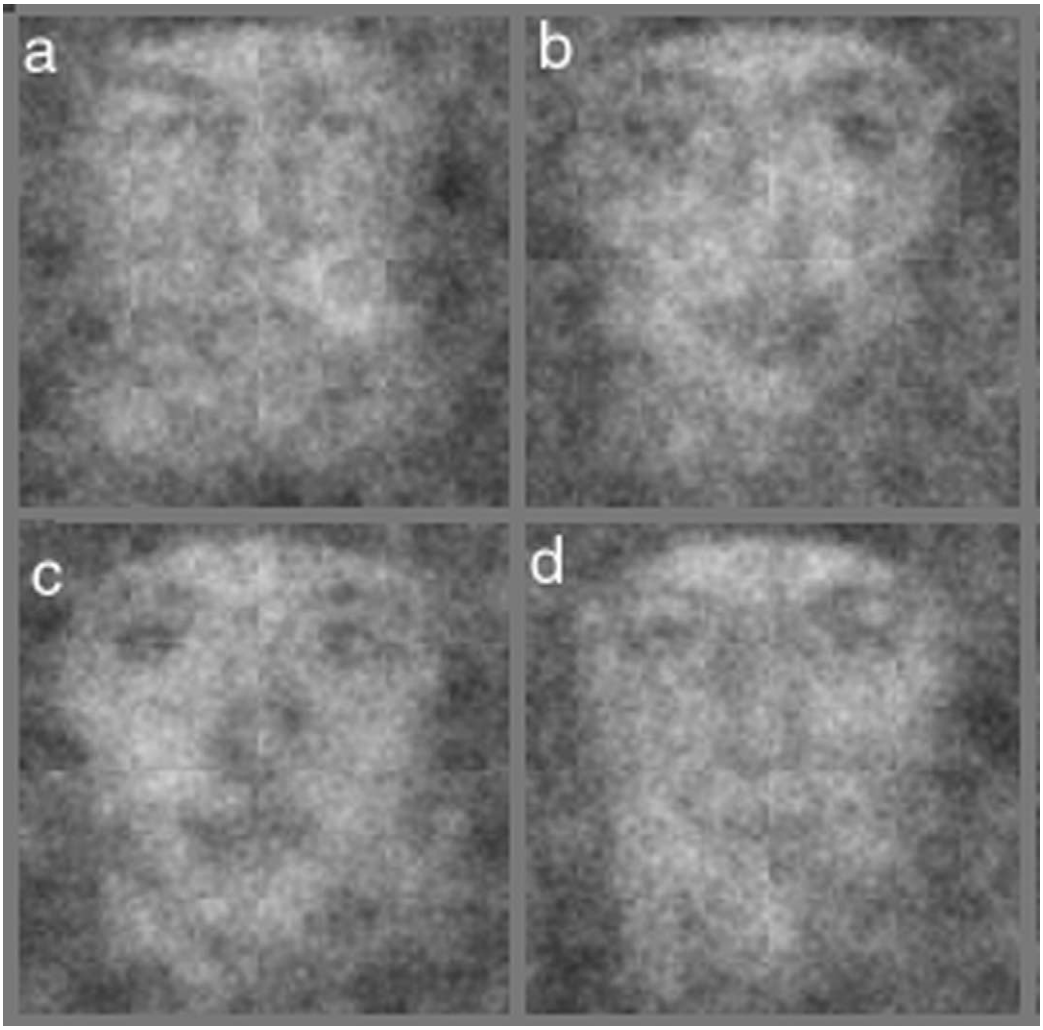

Fig. 2. Four noisy face images produced by superimposing sinusoidal noise over the identical base image. Different patterns of the noise can render the resultant image as looking male or female and happy or unhappy. A sample of observers judged panel a as an unhappy male, panels b and c as happy females, and panel d as a happy male. (In the actual experiment, an individual subject made only a single classification judgment, e.g., happy or unhappy.)

As noted previously, the Schyns, Bonnar, and Gosselin (2002) method localizes the regions that observers employ to discriminate one picture from another. We can compare our results to theirs by noting the regions of our classification images that are of high contrast. The results for the two methods are in good agreement on a coarse level for the expression and gender tasks in assigning high value to the regions around the eyes and mouth, respectively. For identity, a more distributed region over the face is employed. The classification-image technique need not be limited to simply localizing the areas utilized. We can also view the direction of changes that influence observer decisions. For example, by examining the reconstruction images for

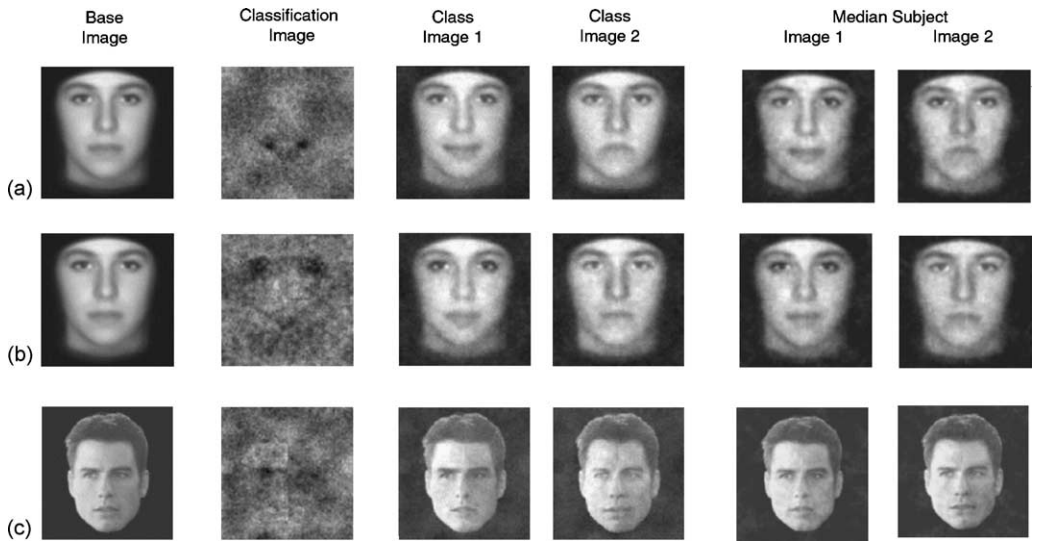

Fig. 3. The results from the three tasks based on the high confidence "probably" trials (a) Happy/unhappy, (b) male/female, (c) Tom Cruise/John Travolta. The base images were identical for the expression and gender tasks. The darkest and lightest areas of the classification images, which are from the data of all 36 subjects, indicate the areas that most influenced the subjects' classifications. The addition of the classification image to the base face results in Class Image 1, which appears happy. The subtraction of the classification image results in Class Image 2, which appears unhappy. The same addition and subtraction operations produce the class images for (b) female and male and (c) Cruise and Travolta, respectively. The rightmost two columns show the classification images for the median subject calculated in terms of Euclidean pixel distance for a given subject's classification image and the average classification image.

Gender in Columns 3 and 4, it becomes clear that while the eyes and mouth play a large role, the diffuse energy in the center of the face creates distinctive gender changes of the nose. This illustrates how this technique can discover the diffuse, subtle information employed by face perceivers.

A classification image can be constructed from only those sinusoidal components that differed significantly between the two categories on each task. Significance was tested with repeated independent t tests for each of the 4092 components with adjustment for multiple comparisons (Rom, 1990). For the Expression task, 187 components reached significance ($p < .0005$), for Gender, 85 components ($p < .001$), and for the celebrity identity task, 52 components ($p < .001$). The images in Fig. 4 reveal that, indeed, relatively few components, in the order of 100, are adequate to recreate the class differences.

Valentin et al. (1994) and Sergent (1989) have speculated, on the basis of statistical analysis of sets of faces, that whereas gender and expression can be conveyed by low frequency information, individuation is carried in higher frequency channels. Fig. 5 shows the class images for all sinusoids separately for each of the five scales. For all tasks, including individuation, a large portion of the information distinguishing the classes appears to be at 4 and 8 cycles/image. This shows that human observers categorizing faces by identity do chose to make use of low frequency information in performing their categorizations. Furthermore, this low frequency in-

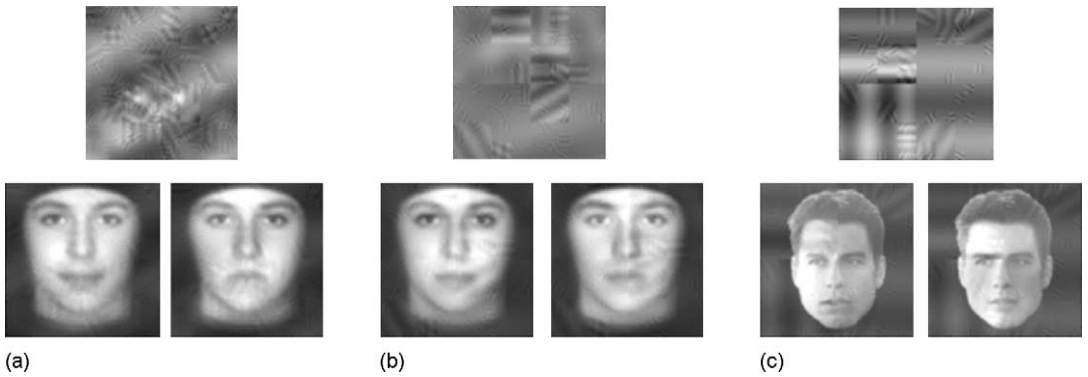

Fig. 4. (Upper row) The significant ($p < .001$) components of the classification image. The sinusoidal nature of the classification images is now clearly visible. (Lower row) Class differences: (a) expression, (b) gender, and (c) celebrity identity are adequately recreated by their (a) 187, (b) 85, and (c) 52 statistically significant components, respectively.

formation can lead to perceptually distinct categories. In agreement with these studies it appears from Fig. 5 that identity categorization utilizes the 16-cycle/image octave to a greater extent than does gender categorization. Because we were interested in including the hair line and hair for the identity task, the size of the internal face was 68.6% smaller (so that 16 cycles/image is approximately 8.75 cycles/face width for the identity condition and 12.5 cycles/face width in the gender and face conditions) and thus the apparent greater utilization of the 16 cycles/image scale in the identity task could simply reflect the appeal of lower frequency information over the face itself. In none of the tasks does the 32-cycle/image band appear to have captured a noticeable category difference. The reduced reliance on high frequency information is not simply a function of the greater energy in the low frequency sinusoids because Schyns et al. (2002) showed that with equal energy across scales subjects still preferentially employed low (3 cycles/face) frequencies when classifying the expression of a face. Our own inferences about the relative weight given to different scales have to be tempered by considerations of statistical power: Both the number of subjects that we could run and the length of the sessions were the known limiting factors.

These results appear consistent with noise masking and image filtering experiments, which have shown that discrimination of emotion is most efficient at information centered at 8 cycles/image (Schwartz, Bayer, & Pelli, 1998) and that face identification is supported by the information carried by frequencies between 8 and 16 cycles/image (Costen, Parker, & Craw, 1994). What is also apparent from Fig. 5 is the humans chose to utilize information across two or more octaves and that this choice appears to be task dependant.

The response classification technique, as described here, provides a useful method for deriving linear approximations to the information subjects used to classify faces. (We will discuss the limitations of linear approximations in Section 3.) As such, the results of this task have made explicit what would otherwise have remained ineffable. From the classification images obtained in Fig. 2 we see that human observers are flexible, in that they make use of information at different facial locations and spatial scales dependent on their task.

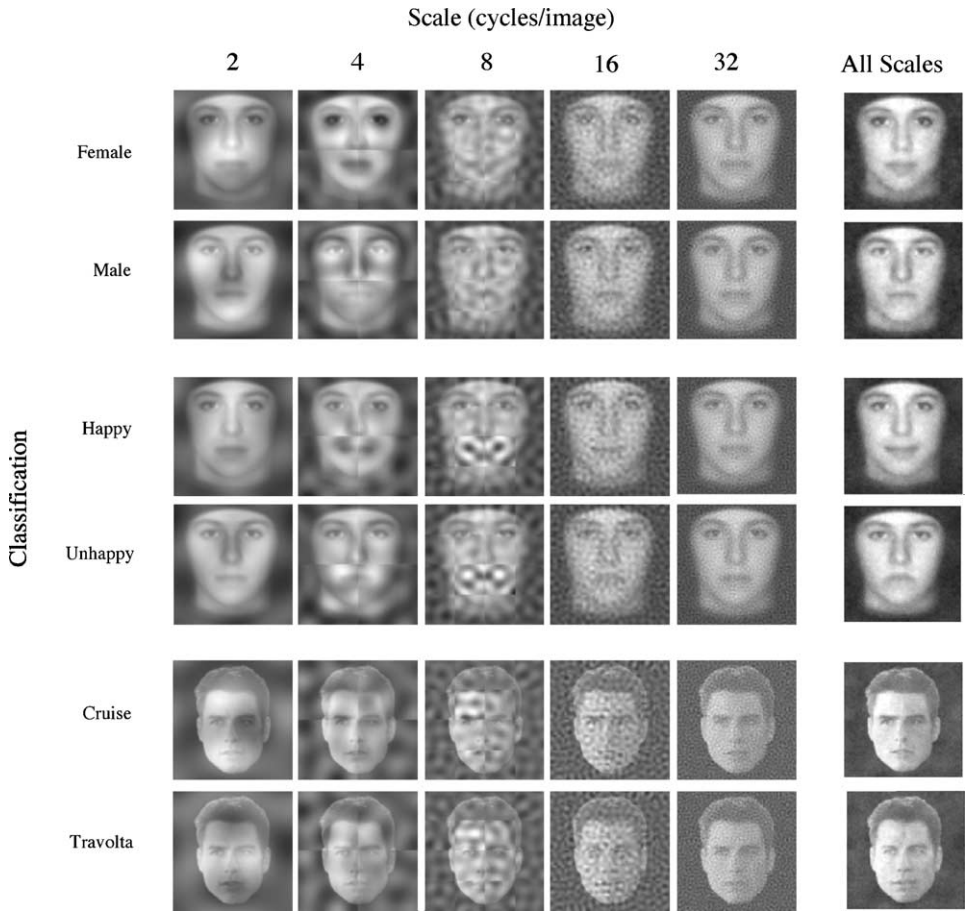

Fig. 5. Class images (the base image (+) or the classification image (−)) with the noise shown separately for each of the five scales. For example, the female at 4 cycles/image is the combination of the base image with the classification image recreated with only the 4 cycles/image sinusoids. The noise at 32 cycles/image appears to add little to the differentiation of the classes.

3. General discussion

We will first discuss how the addition of noise can be considered a method for sampling stimuli from a local region of feature space. We then discuss our choice to use truncated sinusoids rather than pixels (standard white noise) as features. Last, because a classification image is a linear approximation, we provide a general discussion of linear models in the context of face classification.

3.1. Comparison of our method with the standard response classification technique

The response classification technique used in the present experiment is based on that originally described by Ahumada (1996). Ahumada's approach follows from signal detection theory

(SDT) which emphasizes explicit parametric modeling of the relationship between the physical properties of a stimulus and the observer's decisions, where all decisions of an observer are considered to emanate from a probabilistic process that has some level of intrinsic uncertainty (Green & Swets, 1966). The response classification method was developed to approximate the linear template observers used in performing a two alternative forced choice test (Ahumada, 2002).

The current experiment, while it used the response classification technique, diverged substantially from the SDT approach. This was necessary because our goal was not to model the discrimination of a particular known signal with a linear template, as is done by many users of the response classification technique (Ahumada, 1996; Ahumada & Beard, 1999; Solomon & Morgan, 1999; Watson, 1998). For example, to determine why the efficiency of human observers is almost always markedly lower than that of ideal observers, these investigators would be interested in how an observer's classification image differs from the ideal template. Alternatively, some investigators have studied how a particular variable, for example, attentional cueing (Eckstein, Shimozaki, & Abbey, 2002), or perceptual learning (Gold et al., 2004), alters the observer's template.

If we had adopted the SDT approach in performing a response classification experiment for gender, for example, we might have used a male and a female face (or several male and several female faces). However, we would not have known the extent to which our particular selection of faces could have determined the resultant classification image. Such an experiment would have likely produced a classification image that closely approximated the differences between the stimuli for the two categories, so we would have run an experiment just to acquire a noisy version of the differences between the particular faces we happened to select.

Instead, we sought to estimate the information subjects would normally utilize when classifying faces along some dimension, for example, male versus female. Our solution was to present a single stimulus that was uninformative for the task but which placed subjects near the category boundary for the classification they were to perform.

This single stimulus can be viewed as a single point in a high dimensional feature space. For example, a 64×64 pixel image would be located in pixel feature space at location $(x_1, x_2, \ldots, x_{4096})$ where each coordinate is the brightness of one pixel in the image. The addition of noise in this view is equivalent to displacing this point randomly in the feature space. Over many trials the various noisy stimuli constitute a cloud of points distributed about the central point.

This view of the task allows one to consider the addition of noise not as a means of increasing uncertainty between two responses, but as an unbiased method for sampling from a local area of feature space. The behavioral (or even electrophysiological, or neuroimaging) responses can then be used to compute a regression function that maps from stimulus features to behavior. We discuss the feature space we chose, sinusoids, and the limitations of linear regression for face classification in the next two sections.

3.2. Sinusoidal versus white noise

Although information estimation by reverse correlation has typically been done with white pixel noise, we chose to use truncated sinusoids because (a) sinusoids more closely approx-

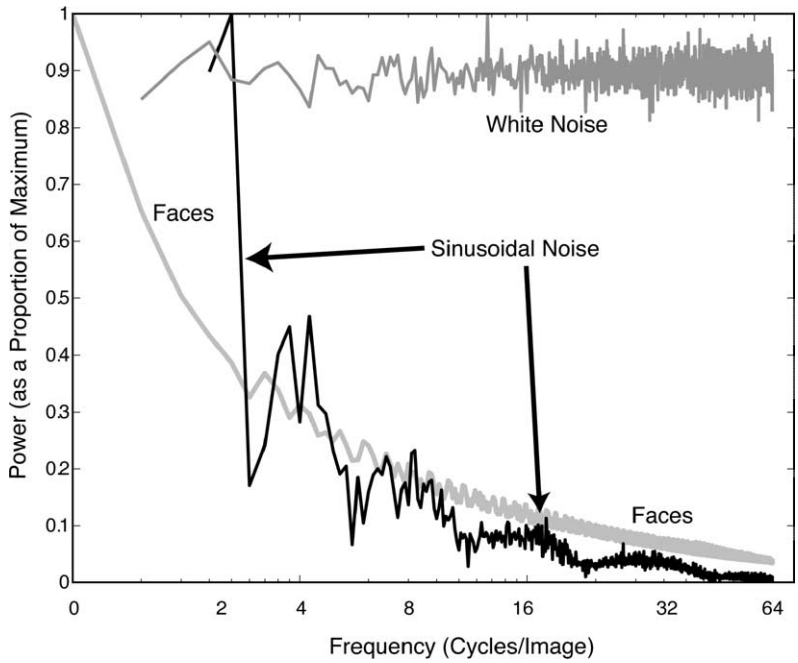

Fig. 6. The average amplitude spectra of 100 sinusoidal noise patterns, 100 white noise patterns, and 40 faces.

imate the preferred stimulus for early cortical visual areas (De Valois & De Valois, 1990), (b) some theorists have posited that the representation of a face retains essential aspects of the initial cortical spatial representation, albeit with scale and translation invariance (Biederman & Kalocsai, 1997; Lades et al., 1993; Wiskott, Fellous, Krüger, & von der Malsburg, 1997),[4] and (c) our sinusoidal noise patterns produced a power spectrum which is similar to that measured in faces (Fig. 6). With the noise spectrum similar to the signal distribution, an observer is more likely to interpret the noise as a stimulus change. While such a power spectrum could have been achieved by filtering white noise, the sinusoidal noise requires four times fewer parameters than the white noise.

Importantly, in simulations with a theoretical observer performing a modest numbers of trials, for example, less than 1000, sinusoidal noise converged to a more accurate estimate than did white noise. The theoretical observer analysis was performed utilizing templates that were the least squares linear approximations of the pixel differences between (a) a set of male versus female faces, (b) a set of happy versus unhappy faces, and (c) between two androgynous morphed faces (an identity classification). For each of the tasks, on each trial, the observers correlated one random sinusoidal noise pattern and one white noise pattern with the template to make a category judgment. Classification images were computed as they were in the subjects' tasks. The correlation between the pixels in the classification image and the pixels of the stored templates was used as a measure of convergence. Fig. 7 shows the results of the simulations. For all three tasks, sinusoidal noise produced better estimates of the template (higher correlations) within the first 1000 trials. Although the pixel noise produced an estimate that was equivalent to

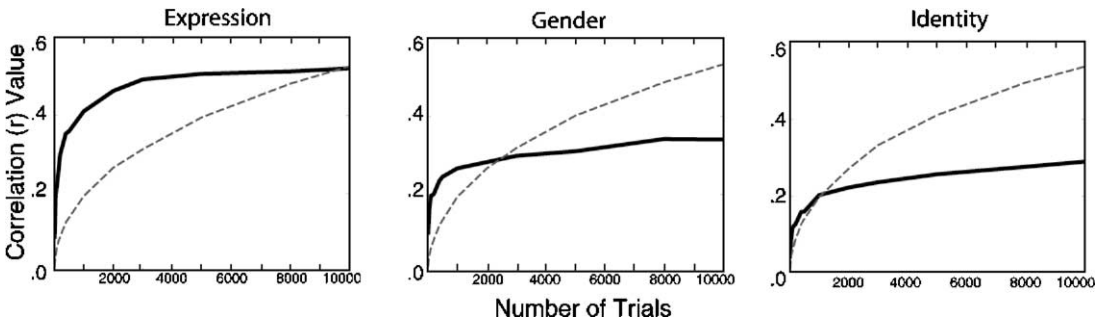

Fig. 7. Degree of convergence of two noise types the sinusoidal (thick black line) and the white noise (gray dashed line) to the linear templates of three theoretical observers performing the expression, gender, and identity tasks. In all three cases the sinusoidal noise provides a better estimate early in testing (less than 1000 trials). However, after 10,000, 4000, and 1000 trials the white noise condition yields higher correlations than the sinusoidal condition for expression, gender, and identity tasks, respectively.

the sinusoidal noise by 1000 trials for the individuation task, the pixel noise did not produce an equivalent estimate until 4000 trials for the gender tasks, and 100,000 trials for the expression task. As can be seen in Fig. 6, these differences among the tasks in the relative goodness of the estimates derived from pixel versus sinusoidal noise was almost completely a function of the sinusoidal noise insofar as the growth in the correlation value for the pixel noise was consistent between tasks. Because the advantage of the sinusoidal noise would primarily be in the low frequency range, most likely the advantage for the sinusoids in the expression task is because there is relatively less energy in the high frequencies of the theoretical observer's identity template compared with the individuation template.

3.3. Linearity in face classification

Modeling a classification as a *linear* process implies several strong constraints that may not be obvious to those not familiar with the concept. To elucidate linear classification, we will consider three ways in which a linear model may fail to capture the psychological reality of human performance in face categorization tasks. Linear classification (1) may over generalize; (2) may provide a poor metric for the underlying mental process; and (3) is insensitive to interactions that may be important in face recognition.

Points 1 and 2 arise because many data sets are not well fit by a line. In real world applications with high dimensional data the fit would be for a hyperplane. It is obvious that a U-shaped relationship between variables would not correspond well to a linear fit. To help illustrate more subtle problems, an imaginary data set is plotted as a set of points in Fig. 8. In this imaginary plot, squares represent male faces and circles represent female faces. The *x*-dimension in the plot represents the age of the individual. The *y*-axis is an imaginary feature that allows separation of the genders. Inspection of the points allows us to see a general trend in which at the youngest ages both male and female faces are closer to the mean of the female faces and are less distinguishable in this *y*-dimension. At the oldest ages both male and female faces are closer to the mean of the male faces.

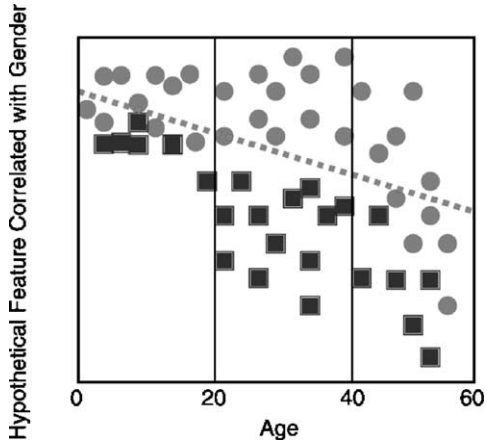

Fig. 8. Hypothetical scatter plot of age and a feature correlated with gender illustrating potential problems with linear boundaries. Circles = females. Squares = males. If only the middle third of the age range was used, the linear boundary (solid black line) between genders would not capture the phenomenon that young males resemble females but that older females resemble males.

Using a linear boundary to separate these classes is not disastrous; the dashed line would be 77% correct in separating males from females. However, if only the middle third of the age range had been used to determine the classification boundary, the category boundary would have been perfectly horizontal. In this hypothetical instance our limited data set would miss the interaction between our gender feature and age resulting in systematic misclassification of the youngest and oldest thirds of our population. Likewise, we could consider other dimensions, for example, race, that may also interact with gender classification. Using more inclusive data sets should improve generalization, but because the solution is global (a single linear/hyperplanar boundary) the overall accuracy will decrease as the same simple linear boundary is used to approximate more and more complex category boundaries.

For the same reasons that a linear category boundary may over-generalize the separation between categories, it may also be a poor metric for the differences within the category. For example, in Fig. 9 below, we are using a linear template to create two sets of stimuli. Row A displays four Gabor patches; each Gabor patch contains four times the contrast energy of the previous patch. A number of psychophysical experiments (Pelli, 1990) have shown that a linear template appears successful at predicting threshold detection for the Gabor stimuli such that the same increment in external noise energy must be added to each successive patch to keep the observer at threshold performance. The faces in Row B were calculated by performing a linear regression on the pixels of a set of images of male and female faces. Starting with a face only slightly more feminine than the mean of male and female faces, each face to the right contains four times the energy of the *female* signal as the face to the left. Although Face 2 appears more female than Face 1, Face 4 does not appear to be a "super female." Rather it appears to be grotesque. A linear metric thus appears to fail at extrapolating beyond small steps when it comes to qualitative femininity.

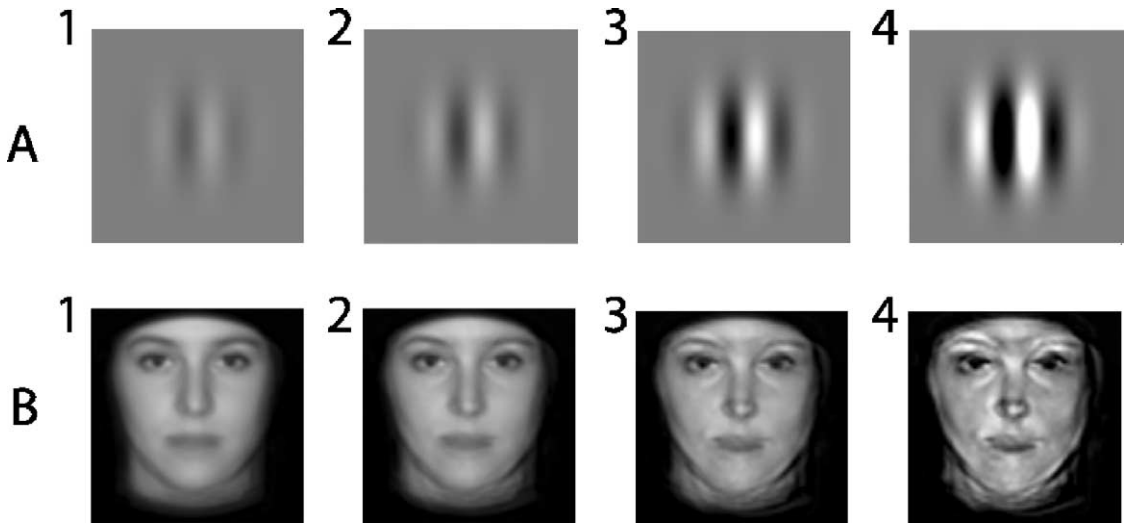

Fig. 9. Threshold detection of a Gabor patch in external noise has been successfully modeled utilizing a linear template. Every step in Row A from 1 to 2, from 2 to 3, and from 3 to 4 will require the same increase in external noise energy to remain at the same level of accuracy. Row B: Face 1 is only slightly more feminine than the mean of male and female faces; each face to the right contains four times the energy of the *female* signal as the face to the left. The linear metric appears to fail at extrapolating beyond small steps when it comes to the "signal" of femininity. Face 4 does not appear to be a "super female"; rather it appears to be a grotesque.

A third limitation of linear template models is that they do not allow for some aspects of the interaction between features. A trivial example of this would be the inability of a linear template to determine whether an individual was winking. This is a face recognition version of the XOR problem as described by Minsky and Papert (1969). In attempting to use the response classification technique to derive a classification image for a winking face, averaging across left and right eye winks, would yield a face in which both left and right eyes are "half closed." Because the average of all trials perceived as non-winks, either both eyes open or both eyes closed, is also a face with both eyes "half closed," the difference of the two classes yields nothing. While this is certainly a special case in which the linear model fails completely, it is not unreasonable to expect that interactions do play a role in face processing and such interactions will not be captured by the linear model.

There is experimental evidence that suggests face recognition by humans may rely on such interactions. Tanaka and Farah (1993) showed that recognition of individual features were more accurate in the context of faces. A linear template model would show no difference in the amount of information available in the feature alone versus the feature in its context conditions. The interactive effects between features can be demonstrated perceptually as discovered by Pelli and his colleagues (Schwartz et al., 1998). In the interests of control, the investigators used a photograph of a face and inserted either a mouth that was smiling or one that was frowning. As can be seen in Fig. 10 altering the mouth alone changes the perception of the entire face including the upper part of the face so that the eyes appear to "light up" with the smile.

Fig. 10. An illustration of Schwartz et al.'s (1998) demonstration that the insertion of a mouth that is either smiling or frowning exerts effects at other locations in the face so that the eyes appear to "light up" with the smiling face.

Appeal has often been made to configural information as a signature of face processing. Because it is unclear what "configural" information in a face image means, it is difficult to determine whether our method captures the configural changes in faces. What is clear is that under subjective inspection, the Class Images in the third and fourth columns of Fig. 3 do not appear to lack any essential information characteristic of their classes. Inasmuch as some may argue that we have not captured configural information, then it could be said that configural differences are not necessary to produce clear categorical differences for gender, expression, or identity. Perhaps we should consider what "configural" information of a face could possibly mean. The only neurocomputational account of configural face information is that proposed by Biederman and Kalocsai (1997). They argued that many of the phenomena associated with face recognition could be derived by assuming that faces are represented by the original spatial (Gabor) filter activation values, with allowance for translation and scale invariance, in a manner proposed by von der Malsburg and his associates (Lades et al., 1993; Wiskott et al., 1997). Medium and low spatial filters would cover broad regions of a face so local variation in an image would have effects on filters whose receptive fields were centered at considerable distances from that variation. Rather than a (highly implausible) process that would make explicit measurements between, for example, different face parts, the configural effects as well as all the shape informations are implicit in the activation of the Gabor filters. While an investigator can use clearly defined operations to produce "featural" versus "configural changes," a Gabor representation does not distinguish between these kinds of face information. Either class of changes will produce variation in the thousands of Gabor coefficients (in the model; likely many millions in the brain). This may be a major reason why people find it so difficult to describe the differences between two similar faces and why the eyes might appear to light up from the presence of an upturned mouth.

Why then would one use a linear model for human face categorization? One benefit that Eckstein and Ahumada (2002) noted is that linear representations can be inspected visually in

the same dimensions as the stimulus, an especially appealing characteristic to vision scientists. Most important, linear models are simple and can be estimated with moderate amounts of data collection. Acquiring enough data to characterize interactions with high confidence is normally prohibitively expensive. For example, a 64×64 pixel image has 4096 pixels. Computing just the linear template in this pixel *space* requires estimating 4096 parameters in a regression function. A measure of the two-way interactions in this space would require the estimation of 4096×4096, more than 16 million parameters. Last, linear estimates are often good approximations under small steps. So, while our linear template fails in extrapolating out to *super female* faces in Fig. 9, it does provides a useful local metric for the attribute of femininity under sufficiently small steps.

3.4. The classification image is not necessarily a representation

Some might be tempted to regard the significant sinusoids shown in Fig. 4 as *representations* of (the differences between) the various categories. We advise caution on this point. Our initial choice of feature space, truncated sinusoids, and our choice of regression function, linear, completely determined the form of what normally constitutes a *representation*. These experimental variables were chosen before any data were collected. For this reason the experimental data have only been discussed in terms of the information subjects have utilized. This said, in fact the sinusoids (or Gabors) are not only well matched to early cortical stage tuning, they are well matched to the representations assumed by a number of theorists. The von der Malsburg face Recognition System (Lades et al., 1993; Wiskott et al., 1997) posits an array of columns of sinusoidal-like (actually Gabor) filters, with each column termed a *Gabor jet*. This system won a U.S. national competition for face recognition performance (Okada et al., 1998). Biederman and Kalocsai (1997) have argued that many of the distinctive phenomena of face recognition can be derived from a representation defined by the activation values of Gabor jets such as that embodied in the von der Malsburg system.

4. Summary

Classification images, when obtained in the manner described in this report, provide an efficient and effective method for deriving linear estimates of the information employed by human observers when making face classifications. Other methods of estimating the information employed in complex classification have relied on judgments of exemplars of a class or tests of experimenter-defined hypotheses about the class information. Our method allows an estimate, however subtle, of what is in the subject's (rather than the experimenter's) head. The method also allows reconstruction of an image based on that estimate. The classification image technique thus provides a method for making explicit otherwise ineffable perceptual representations.

Notes

1. This base image was created for use in a previous experiment in which the morphing step provided added benefits. For the purposes of this experiment the authors place no

importance on the morphing process. The average of the original 20 images would have produced a very similar image.

2. We say *nominally* localized because within the square window of any particular sinusoid the frequencies, orientations, and phases are localized. However, if a Fourier decomposition was performed over the space of the entire image the envelope of the higher frequencies would create Gibbs effects which may overlap in lower frequencies. The tiling process also induces high frequency artifacts visible at the boarders of the tiles.
3. This weighting is not optimal but as there are no hits or false alarms we could not enforce a payoff matrix so there is no available method for calculating an optimal weighting.
4. To be precise, these authors have argued for a Gabor (a Gaussian damped sinusoid) representation, which closely resembles a truncated sinusoid. Unfortunately, reconstructions are difficult with Gabors as they are not orthogonal, a problem not encountered with truncated sinusoids.

Acknowledgments

We thank Zhong-Lin Lu, Bosco Tjan, and Frédéric Gosselin for their helpful inputs to this work. Supported by ARO DAAG55-98-1-0293, NSF EEC-9529152, HFSP 99-44, JSMF No. 96-44.

References

Ahumada, A. J. (1996). Perceptual classification images from Vernier acuity masked by noise [Abstract]. *Perception*, *26*(Suppl. 18), 18.

Ahumada, A. J. (2002). Classification image weights and internal noise level estimation. *Journal of Vision*, *2*(1), 121–131, http://journalofvision.org/2/1/8/, doi 10.1167/2.1.8.

Ahumada, A. J., & Beard, B. L. (1999). Classification images for detection [Abstract]. *Investigative Ophthalmology and Visual Science*, *40*, 3015.

Ahumada, A. J., & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustic Society of America*, *49*(6), 1751–1756.

Biederman, I., & Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society London: Biological Sciences*, *352*, 1203–1219.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2000). A principal component analysis of facial expressions. *Vision Research*, *41*, 1179–1208.

Costen, N. P., Parker, D. M., & Craw, I. (1994). Spatial content and spatial quantisation effects in face recognition. *Perception*, *23*(2), 129–146.

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). EMPATH: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience*, *14*, 1158–1173.

De Valois, R. L. & De Valois, K. K. (1990). *Spatial vision*. New York: Oxford University Press.

Eckstein, M. P., & Ahumada, A. J. (2002). Classification images: A tool to analyze visual strategies. *Journal of Vision*, *2*(1), 1–2, http://journalofvision.org/2/1/introduction.html, doi 10:1167/2.1.1x.

Eckstein, M. P., Shimozaki, S. S., & Abbey, C. K. (2002). The footprints of visual attention in the Posner cueing paradigm revealed by classification images. *Journal of Vision*, *2*(1), 25–45, http://journalofvision.org/2/1/3/, doi 10.1167/2.1.3.

Gold, J. M., Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology*, *10*, 663–666.

Gold J. M., Sekuler A. B., Bennett P. J., (2004). Characterizing perceptual learning with external noise. *Cognitive Science* doi: 10.1016/j.cogsci.2003.10.005.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research*, *41*, 2261–2271.

Gosselin, F., & Schyns, P. G. (2003). Superstitious perceptions reveal properties of memory representations. *Psychological Science*, *14*, 505–509.

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

Lades, M., Vortbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., et al. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, *42*, 300–311.

Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2002). Optimal methods for calculating classification images: Weighted sums. *Journal of Vision*, *2*, 79–104, http://journalofvision.org/2/1/6/, doi 10.1167/2.1.6.

Okada, K., Steffens, J., Maurer, T., Hong, H., Elagin, E., Neven, H., et al. (1998). The Bochum/USC face recognition system and how it fared in the FERET phase III test. In H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, & T. Huang (Eds.), *Face recognition: From theory to applications* (NATO ASI Series F). Berlin: Springer.

Pelli, D. G. (1990). The quantum efficiency of vision. In C. B. Blakemore (Ed.), *Vision: Coding and efficiency*. Cambridge, UK: Cambridge University Press.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, *77*, 663–666.

Schwartz, O., Bayer, H., & Pelli, D. (1998). Features, frequencies, and facial expressions [Abstract]. *Investigative Ophthalmology and Visual Science*, *39*, 173.

Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological Science*, *13*, 402–409.

Sekuler, A. B., Gold, J. M., Bennett, & P. J. (2000). *The efficiency of face recognition: Effects of inversion and contrast reversal*. Poster presented at the annual Psychonomic Society.

Sergent, J. (1989). Microgenesis of face processing. In H. D. Ellis, M. A. Jeeves, F. Newcombe, & A. Young (Eds.), *Aspects of face processing*. Dortrecht: Martinus Nijhoff.

Solomon, J. A., & Morgan, M. J. (1999). Reverse correlation reveals psychophysical receptive fields [Abstract]. *Investigative Ophthalmology and Visual Science*, *40*, 3013.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology*, *46A*, 225–245.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *The Journal of Cognitive Neuroscience*, *3*, 71–86.

Valentin, D., Abdi, H., & O'Toole, A. J. (1994). Categorization and identification of human face images by neural networks: A review of the linear autoassociative and principal component approaches. *Journal of Biological Systems*, *2*, 413–429.

Watson, A. B. (1998). Multi-category classification: Template models and classification images [Abstract]. *Investigative Ophthalmology and Visual Science*, *39*, 1109.

Wiskott, L., Fellous, J.-M., Krüger, N., & von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, *19*, 775–779.