

# Using Reverse Correlation to Infer the Representations Distinguishing Facial Gender, Affect, and Individuals.

Michael C. Mangini and Irving Biederman

University of Southern California

People reveal considerable expertise in the classification of a face in terms of gender, expression, and identity, yet the representation mediating such performance is often not available to conscious, explicit description. To specify these representations, observers classified faces appearing in sinusoidal noise as male/female, happy/unhappy, or Tom Cruise/John Travolta. Unbeknownst to the subjects, the underlying face stimulus was identical on every trial. Therefore, all variations in the stimuli—and the subjects' responses—could be attributed to the noise. The correlation of the subjects' responses with the noise was used to compute a "classification image" (Ahumada, 1996) that yielded clear exemplars of the classes. Reverse correlation may thus provide a method for making explicit otherwise ineffable perceptual representations.

Upon looking at a photo of an individual we can tell that person's approximate age, race, gender, attractiveness, and emotional expression in a single glance. We can do this whether we are familiar with the individual or not. A recent study (Mangini and Biederman, 2000) tested whether observers are more sensitive to changes in emotional expression, gender, or the identity of an individual. In that study subjects performed a match-to-sample task where the amount of contrast required to achieve 75% accuracy was the dependent variable. The results showed that observers are most sensitive to differences in emotional expression, and least sensitive to identity. This confirmed that the increase in subjects' performance on expression and gender classifications over individuation was due to observers' sensitivities, rather than differences in the amount of stimulus information or response uncertainty between the three tasks.

Although we can all distinguish a male face from a female face, or a happy face from an unhappy face, when asked to describe the differences most people are at a loss to verbalize just what the image information is that distinguishes these categories. For example image b) in Fig. 1 is a linear discriminant calculated from a principal component decomposition of 20 male and 20 female faces. When added to an androgynous face, image a), the linear discriminant produces a face that is clearly male or female. But, can you guess which gender, c) or d) is created by the addition? When provided with category differences subjects showed they were more sensitive to the information that corresponded to gender and expression changes than that which corresponded to changes in identity. But what information do subjects utilize when performing such a task? The

purpose of the present study was to derive an approximation of the representations mediating such discriminations. We employed a variant of Ahumada's (1996) reverse correlation technique to calculate *classification images* to derive approximations to our subjects' decision dimensions.

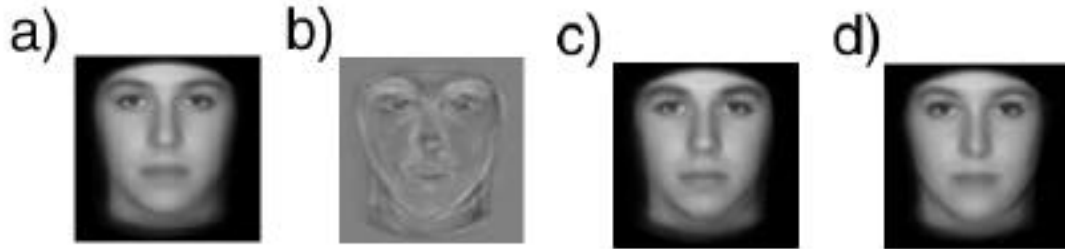


Figure 1. The differences between faces are not consciously accessible. a). An average of male and female faces. b). A linear discriminant that when added to face a) will transform the image. Try to guess whether the addition of b) results in face c) or d).

## Methods

Three experiments were conducted to analyze the information subjects utilized when performing three tasks: 1) determining the gender of a face, 2) determining whether a face was expressing positive emotion (happy) or negative emotion (anger, or sadness), and 3) determining whether the individual was John Travolta or Tom Cruise. In every case the faces appeared in high noise. Subjects responded with one of four confidence ratings: probably Travolta, possibly Travolta, possibly Cruise, or probably Cruise (and likewise with male/female and happy/unhappy). Unbeknownst to the subject the face stimulus appearing in the noise was identical on every trial. For the gender and expression tasks, the stimulus was the same consisting of a face that was the arithmetic mean of 200 images, which were the morphs of 10 male individuals and 10 female individuals neutral expressions. This face, termed the base face, appears neutral in both gender and expression. A morph of images of John Travolta and Tom Cruise was used as the base image.

The noise was composed by summing 4092 sinusoids. We used sinusoidal rather than white pixel noise because it more closely approximates the preferred stimulus for early visual areas and because the former converged more rapidly on an effective classification image. 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 per image), six orientations (0,30,60,90,120 and 150 degrees), and two phases (0 and  $\pi/2$ ) were summed to create one noise pattern (Figure 2). The amplitudes of the sinusoids were selected

randomly from a uniform distribution. Thirty-six subjects performed in each experiment, each subject performed 390 trials.

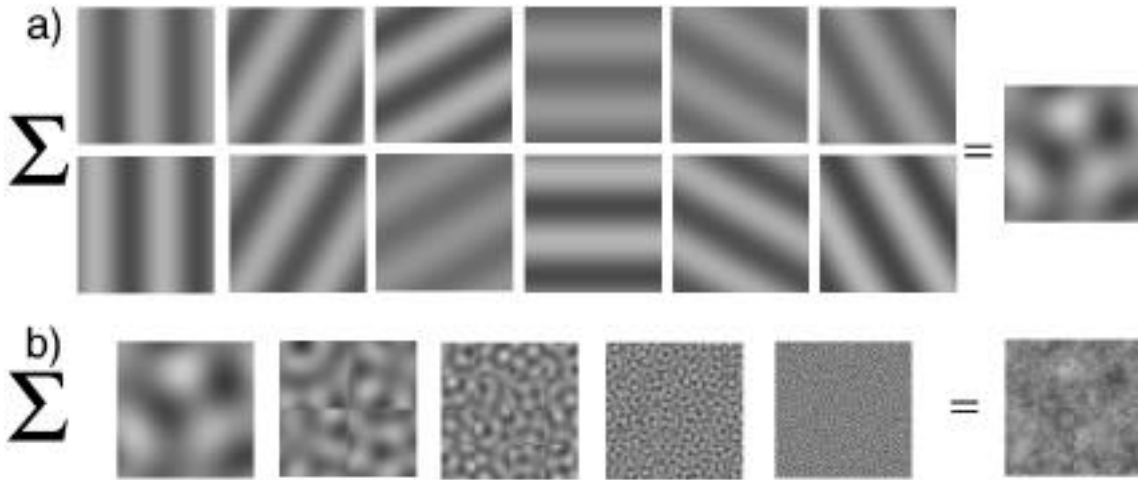


Figure 2. Creating the Noise: a). Six orientations and two phases of the sinusoids are summed to create the 2-cycles/image noise. b). All five octaves are summed to create the noise pattern.

## Results

Every noise pattern that received a confident “Probably” response was summed for the two classes separately and the classification image was computed by subtracting the average noise from one class (e.g. Male), termed a *category image*, from the category image of the other class (e.g. Female). The classification images were computed and their resultant class images for the three tasks are shown in Fig. 3.

The sinusoidal components were each tested for statistical significance. Repeated independent T-tests were conducted and Rom’s procedure (Rom, 1990) was used to control for experiment-wise Type I error. 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 classification image can be reconstructed using only those sinusoidal components that reached significance. The images in Fig. 4 reveal that, indeed, relatively few components, in the order of 100, adequately recreate the class differences.

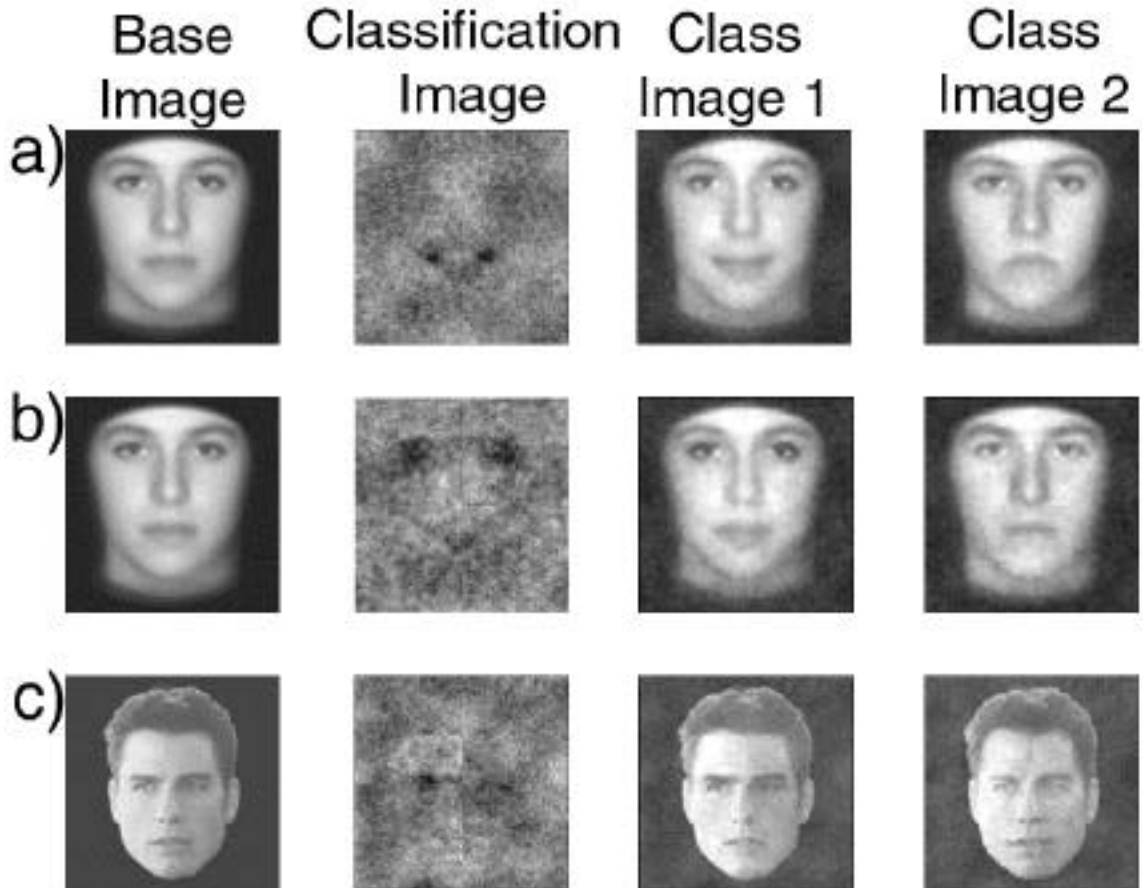


Figure 3. The results from the three experiments. The dark and light areas of the classification images indicate the areas that influenced the subjects' classifications. a). The results from happy/unhappy categorization. The addition of the classification image to the base face results in class image1, which appears happy. The subtraction of the classification image results in class image 2, which appears unhappy. b). The results from the gender classification task c). The results from the Travolta/Cruise task.

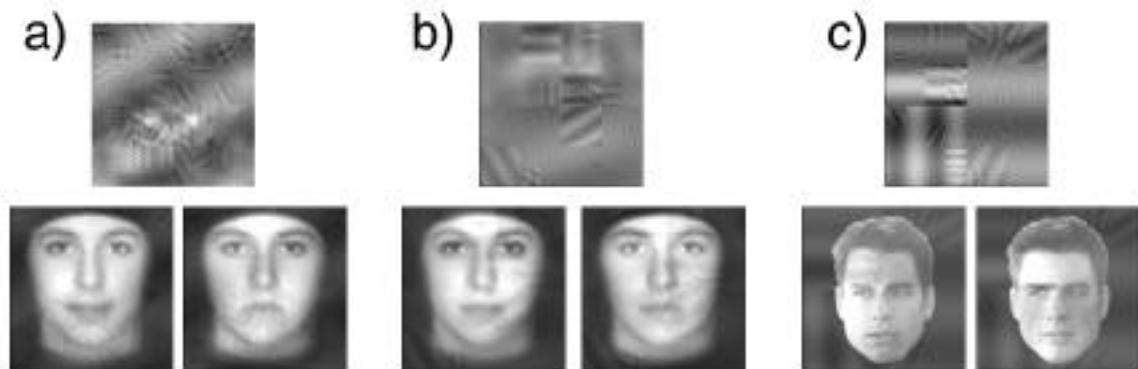


Figure 4. The significant components of the classification image. The sinusoidal nature of the classification images is now clearly visible. Class differences: a) expression, b) gender, and c) celebrity identity are adequately recreated by only a) 187, b) 85, and c) 52 components.

## **Conclusions and Discussion**

With only 390 trials per subject, classification images provide an effective method for deriving linear approximations to the representations mediating face classifications. This method has been utilized to measure the differences between normal and prosopagnosic observers (Mangini and Biederman, 2001), and the differences between humans and baboons in a human vs. baboon face classification task. (Martin-Malivel, Mangini, Fagot, & Biederman, unpublished). We are currently investigating Independent Component Analysis (Bell and Sejnowski, 1995) to include in the analysis the higher order statistics of the reverse correlation data set. Reverse correlation may thus provide a method for making explicit otherwise ineffable perceptual representations.