Are Faces Special?

A Visual Object Recognition Study: Faces vs. Letters

Qiong Wu

58-11 205 St.
Bayside, NY 11364

Stuyvesant High School

345 Chambers St.
New York, NY 10282

Q. Wu (2001) *Are faces special? A visual object recognition study: faces vs. letters.* Intel Science Talent Search. http://psych.nyu.edu/pelli/#intel

Summary

Visual object recognition is something that people do easily, yet no one knows how we do it. Research in visual cognition suggests that there is a wide spectrum of kinds of objects, with faces at one end, processed holistically, and letters and words at the other, processed by parts (Farah, 1994). It has recently been discovered that the visual channel used to identify letters is different than that used to detect a sinusoidal grating pattern with the same nominal spatial frequency as the letter. We used visual noise masking to find the channel used to identify a face, and found, surprisingly, that it is identical to the channel used to identify letters, even though these objects come from opposite ends of the continuum.

Introduction

The things we see every day can be broken down into objects. Our visual system picks up the image of these scenes and somehow recognizes the individual objects. In each part of the visual field there is a best stimulus for each neuron to respond. Different neurons answer to different stimuli.

Fletcher (1940) measured thresholds for a pure tone in the presence of noise of many frequencies, and found that threshold was greatly elevated by noise frequencies close to that of the signal, and little or unaffected by distant frequencies. The graph of threshold versus frequency is a tuning curve for that auditory “channel”. The visual channel is analogous, selective to the spatial frequency of patterns in an image, instead of the temporal frequencies of sound. Similarly, this experiment measures noise frequencies that are most effective in elevating the visual threshold, in order to learn what the visual channel’s best frequency is.

Sine waves (see Fig. 1) are the simplest type of stimulus. Since the channel is only detecting the stimulus, the frequency at which the channel uses to detect should be the same as the stimulus frequency, thereby having a slope of 1.

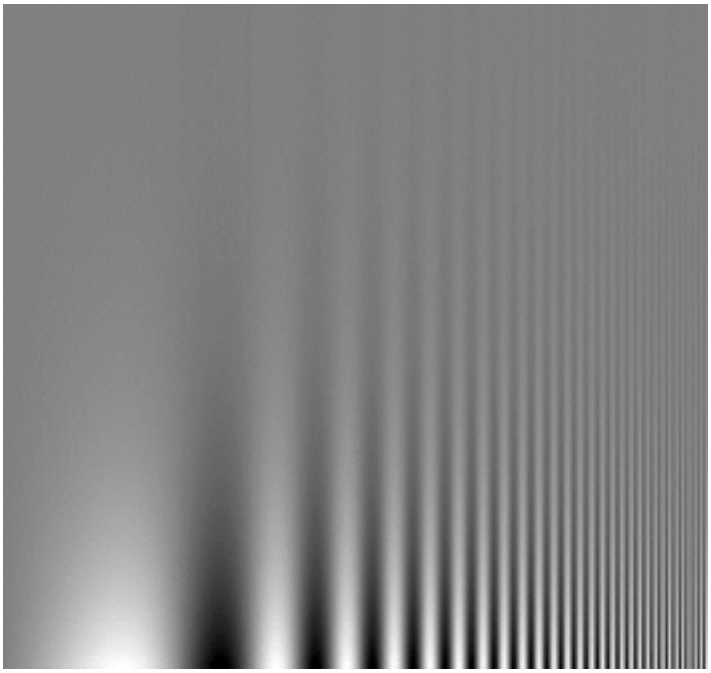


Fig. 1 – Sine waves are narrowband and very simple. They consist of black and white strips that softly blend into each other.

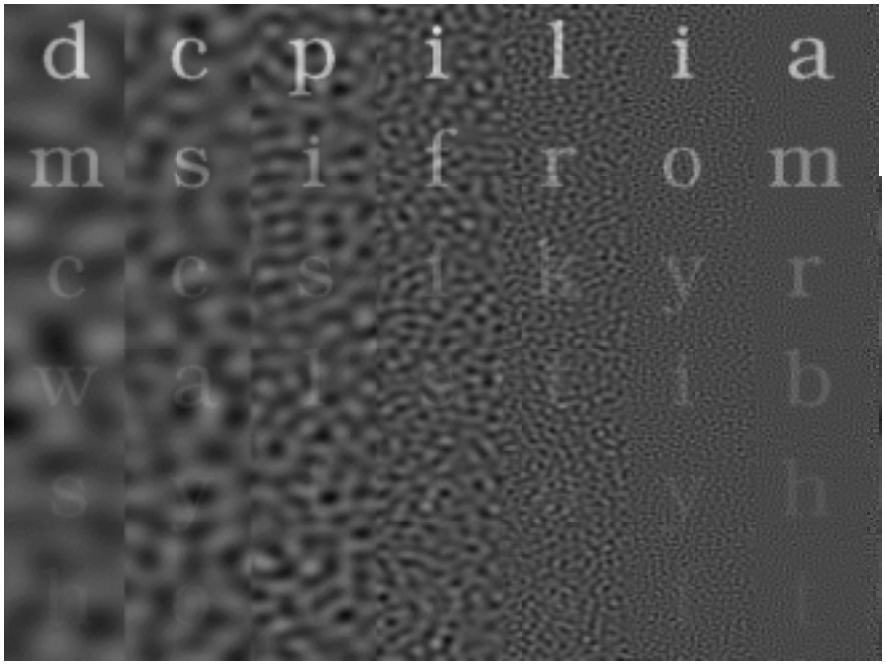


Fig. 2 – Letters in noise. Noise is added at different cut-off frequencies. This demonstrates that the noise in the middle column (or 3 cycles/degree) masks identification of the object more effectively than at either lower or higher frequencies.

As for objects, many people have suggested that faces and letters are of two different ends of the object spectrum. Letters, unlike gratings, are spatially compact but spectrally broad, which implies their identification to be derived by multiple channels, or at least a broader channel than the one used for grating detection (Solomon and Pelli, 1994).

An important question raised in facial expression recognition is whether faces are processed on the basis of their individual features or, more holistically, of their overall shape. It has been found that the visual system relies on holistic representation of the face when recognizing an individual's identity (Tanaka and Farah, 1993). If this is the case, then when the face is viewed as a whole, the contours and curves, as well as any shadows generated by facial features (eyes, nose, etc.) will resemble a sine wave rather than an object, like a letter. It'll be very interesting to either prove or disprove this expectation.

Another question raised is whether the image is viewed top-down or bottom-up. Top-down means that the observer determines which channel to use to process the image after the retina picks it up. Bottom-down, on the other hand, means that the stimulus itself controls the channel frequency used, and that our visual system lacks the power to choose a most efficient channel frequency to avoid the noise masked onto an image.

One of the most effective ways of analyzing how an observer performs on vision recognition tasks is to compare contrast thresholds. Contrast threshold is how much contrast must be present in a stimulus for it to still be identified accurately. We used visual noise masking to measure which noise efficiency was most effective in raising threshold.

Methods

Stimuli:

Two sets of stimuli were used in this experiment. The first is a set of eight stimuli consisting of Paul Ekman's seven basic emotions (happiness, sadness, fear,

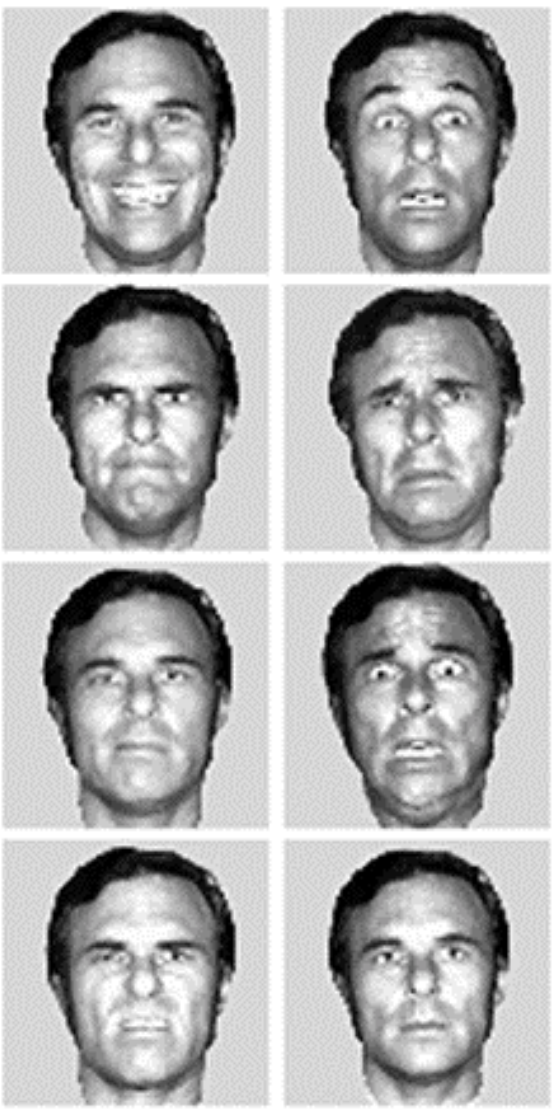


Fig. 3a – Set of Paul Ekman's faces displaying the seven basic emotions and the neutral face.

Fig. 3b – set of Odelia's faces with very subtle changes.

surprise, disgust, anger and contempt) and the neutral face (Ekman, 1992). The second set of stimuli consisted of eight faces of a woman named Odelia. She was photographed making twenty-one funny expressions. We selected eight out of the set of 21 that had the subtlest differences amongst them. The two salient differences between the two sets of stimuli are that Ekman's hair is very dark and he is expressing real emotions whereas Odelia's hair is very light-colored and she is just making faces.

The stimuli were masked with varying frequencies of random white noise, from 0 to infinity, inclusive. Noise is like the snow that we see on a television when nothing is being broadcasted. The faces were tested at a wide range of sizes, or nominal frequencies. The observer(s) adjusted the distance that they sat away from the computer screen so that the angular size of the image would not change.

Observers:

There were two observers for this experiment. One was the author herself, and the other was a college student from Professor Pelli's lab. The two observers carried out the experiment by the same rules and methods.

Procedures:

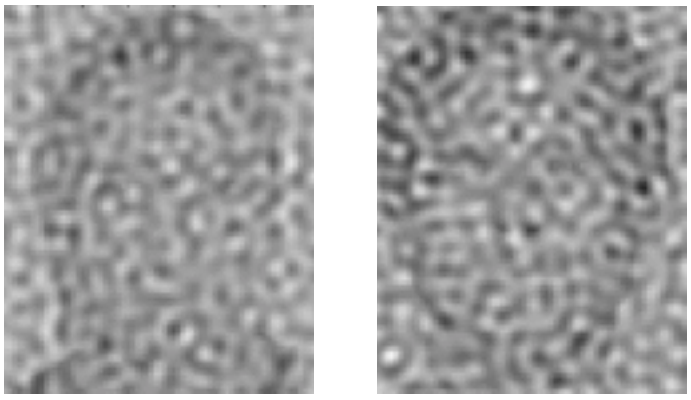


Fig. 4 – A noise mask is placed at the critical frequency making the facial expression hard to identify. (Paul, *right;* Odelia, *left*)

The observer fixates on a small black square at the center of the computer screen. Upon clicking the mouse, the fixation point disappears and is replaced with the signal (a face showing an expression), randomly chosen out of a specific set of stimuli by the computer. The signal can be an image with no noise, or an image masked with noise, either fine or coarse, or an image that is completely masked with white noise, depending on the noise added. It appears for 200 ms, and then disappears. After another 200 ms, the entire set of faces comes up on the screen, clearly visible, without noise. The subject uses the mouse to click on the face in the set that matches the one that popped up 200 ms earlier. This response screen is fixed and does not change in size, contrast, or location. Correct responses are rewarded by a short beep. This cycle constitutes a trial, and forty such trials make a run. The computer changes the frequency of the masking noise added to the image after each run.

Efficiency:

Before executing the tasks, the subjects first learned the faces to maximize their performance. The stimulus was presented in two conditions: no noise or cloaked in noise. The observers ran approximately 2000 trials for each face set before stabilizing efficiency, which average about 3 – 5%. Efficiency was determined by the comparison of the energies of the image with noise, image without noise, and an ideal's energy when run by the computer (Pelli and Farell, 1999).

$$\text{Efficiency} = E/N_{ideal} \,/\, ((E_{white\ noise} - E_{no\ noise}) / N)$$

Analysis:

Simply put, we suggest that the reader imagine that we applied one noise frequency at a time, and measured which was most effective in raising threshold. With

that intuition one may safely skip ahead to the results. In fact there are technical and theoretical advantages to applying a wide band of noise frequencies and doing a bit of analysis to extract what amounts to the same information, but not confounded by problems with dynamic range and off-frequency looking (Solomon and Pelli, 1994).

In this experiment, we varied the cut-off frequency by using noise from 0 (low-pass) frequency to infinity (high-pass). There are two curves, each increasing as the cut-off frequency moves in the appropriate direction to increase the noise bandwidth. We look for the place where the curves rise most steeply, as this indicates where the noise is most effective. The second panel plots the steepness, showing that both ways of doing the experiment yield very similar estimates of best channel frequency.

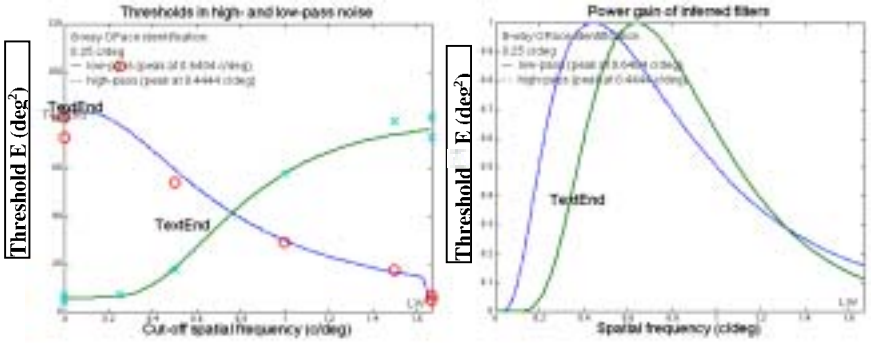
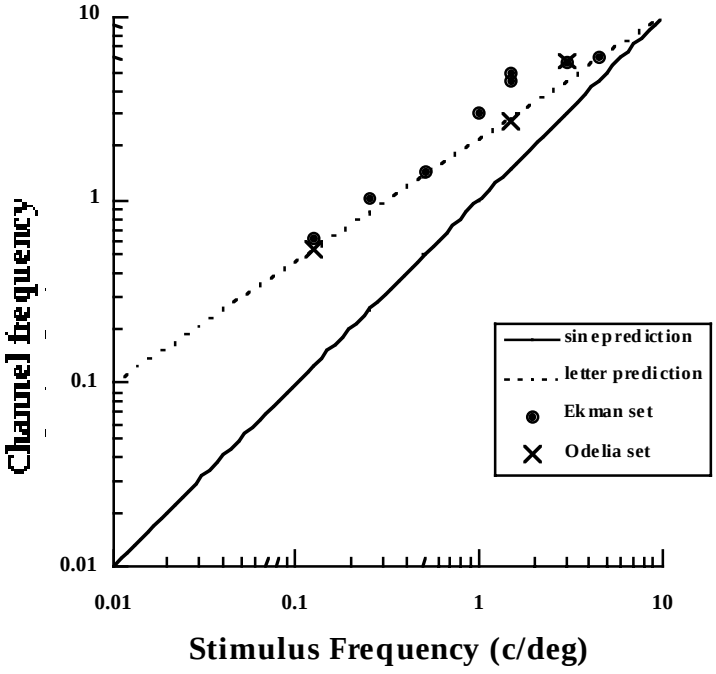


Fig. 5– Graphs of the contrast thresholds of Odelia's faces at 0.25 c/deg. A spatial frequency of 0.6 c/deg is used in detecting the face signal from the low-pass and high-pass. This is purely scientific and the reader need not comprehend this in order to understand the paper itself.

Results



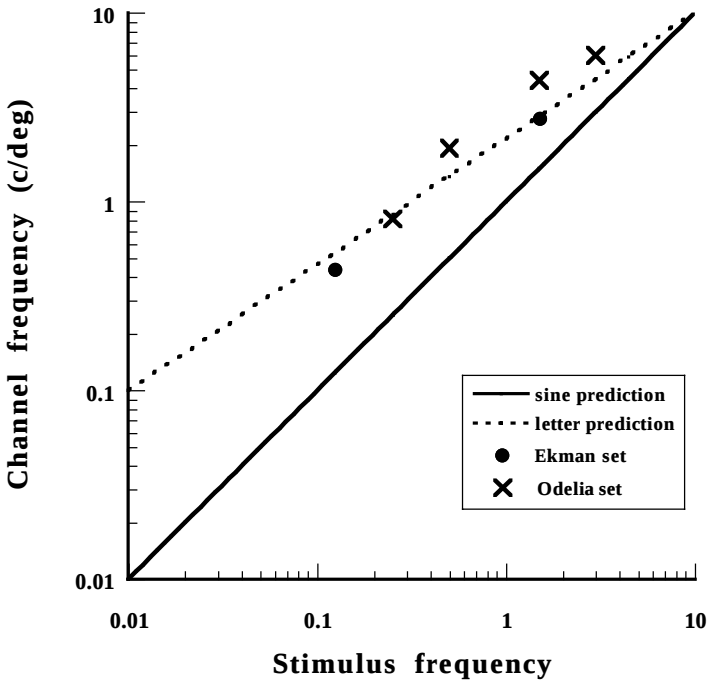


Fig. 6 – The graph on top is Lisa's results and the graph on the bottom is Max's.

General Discussion

Despite the popular idea that faces and letter are at opposite ends of the object spectrum, with faces processed holistically, and letters by parts, we find that the same spatial-frequency channel processes both letters and faces, unlike that for detecting sinusoidal gratings. Over a 36-fold range of face size (corresponding to nominal spatial frequency of 0.25 to 9 c/deg) the face results all fall on or near the letter result. This leads us to conclude that there is some universal aspect to object perception and faces are like objects.

Pelli (1999) found that shape perception was size-dependent. In a block portrait by Chuck Close, it was found that seen from near, the blocks are too large to be integrated into a three dimensional shape, and the face appears flat. When viewed from afar, the angular size of the blocks is greatly reduced, and the viewer can see the face as solid, with a protruding nose. This is relevant to our data in that face recognition is also size-dependent, as the face results lie on the letter result, which has a slope of 2/3. The channel frequency we use to process faces is different at each stimulus frequency (3/size).

Earlier on, the question of top-down or bottom-up was also raised. Looking back to Fig. 5b, the observer is using essentially the same channel (to identify the face) in the presence of low- or high-pass noise. In one case the noise is mostly at lower frequencies (to the left on the graph); in the other case the noise is mostly at higher frequencies (to the right on the graph). If the observer could choose which channel to use, we would expect the observer to avoid the noise by using a low frequency channel (to the left) when the noise is high frequency(right), and vice versa. The results show that the observers use essentially the same channel for the same signal, failing to avoid the noise.

Acknowledgements

I would like to thank Professor Denis Pelli most sincerely for his excellent guidance and knowledge and for the generous use of his lab. A thank you is in order to Melanie C. Palomares whose constant support has been invaluable to me throughout the past year. I would also like to thank MariaLuisa Martelli for her aid and support these past few months. Lastly, to Max, who has replicated my data, and all the students in Professor Pelli's lab who have encouraged me throughout the past year, thank you.

Bibliography

Ekman, P. (1992) Are There Basic Emotions? *Psychological Review*, *99*, 550-553

Farah, M. J. (1994) Specialization within visual object recognition: clues from prosopagnosia and alexia. In M. J. Farah and G. Ratcliff (Eds.), *The Neuropsychology of High-Level Vision* (pp. 133-146), Hillsdale, NJ: Lawrence Erlbaum Associates.

Fletcher, H. (1940) Auditory patterns. *Reviews of Modern Physics*, 12, 47-65.

Pelli, D. G. (1999) Close Encounters – An Artist Shows that Size Affects Shape. *Science*, 844-845

Pelli, D. G., and Farell, B. (1999) Why use noise? *Journal of the Optical Society of America A*, 16 (3).

Solomon, J. A., and Pelli, D. G. (1994) The visual filter mediating letter identification. *Nature, 369*, 395-397.

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