Jeff Stott

Pelli Lab

8/14/2005

Spatial and Temporal Crowding in the Recognition of Words and Faces

Faces and words are very different, but are they identified the same way? Previous studies indicate that both faces and words are recognized by-their parts. Here, we study the effects of both spatial and temporal crowding on face and word recognition, supposing that in order to recognize these objects we must isolate the objects in time and their parts (letters and facial features) in space. Using RSVP in the peripheral visual field, we measured the proportion of words and faces correctly identified as a function of part-to-part spacing and presentation rate. We reject the hypothesis that the spatial and temporal isolation processes are statistically independent, and show that while faces require more time and larger spacings between parts for recognition, the space-time tradeoff is identical for both object types.

Introduction

One of the underlying questions in object recognition research is whether there is a general-purpose brain mechanism for recognizing objects or whether there are different mechanisms specialized for different classes of objects (Grill-Spector, Kourtzi, & Kanwisher, 2001). Some researchers have proposed that words and faces are very different classes of object that are recognized by different brain mechanisms. Farah (1991), in particular, has suggested that the brain has separate modules for different kinds of object, with faces processed as wholes and words processed by parts. fMRI studies have identified the fusiform gyrus as the likely anatomical area specialized for face recognition (Kanwisher, McDermott, & Chun, 1997). Other evidence for a category-specific face region comes from studies of brain-damaged patients who show selective deficits for face recognition (Farah, 1991). Also, human infants preferentially gaze at human faces, suggesting an innate component to face recognition (Goren, Sarty, & Wu, 1975; Johnson, Dziurawiec, Ellis, and Morton, 1991).

Likewise, several brain regions have been proposed as candidate areas in which visual specialization for words could occur, such as Petersen's medial extra-striate region (Petersen et al., 1990), Howard's angular gyrus (Howard et al., 1992) or Cohen's occipito-temporal region (Cohen et al., 2000).

The hypothesis that words and faces are processed by different mechanisms. makes intuitive sense. Words and faces are opposite in several important respects. Faces are very complex images in comparison to words, and may therefore require more complex machinery for identification. Faces are part of the natural world, shaped by evolution. Letters and words are human constructs and presumably, have not been around long enough to influence brain evolution. Words have definite parts—letters—whereas faces do not have clearly defined parts. Since words and faces are so qualitatively different, and emerged at vastly different time periods, it makes sense that the brain would have different mechanisms for processing them. In particular, it would not be surprising if faces turn out to be identified by an innate mechanism, and that words by a different learned mechanism.

Despite all the evidence for the special and unique standing of face recognition, there are some indications that faces—although having a brain area devoted to their identification—may be processed by the same general mechanism as words. Some researchers assert that faces show special effects only because we become expert at identifying them through years of practice (Diamond & Carey, 1986; Gauthier & Tarr, 1997; Gauthier, Skudlarski, Gore, & Anderson, 2000). Psychophysically, Martelli et. al. (2005) have demonstrated that both faces and words are recognized by parts. Moreover, they showed that the critical spacing of facial parts required for recognition is similar to the critical spacing between letters required for word recognition. Until the computations that underlie face and word identification are known, it is difficult to determine whether or not they are processed by different mechanisms.

Most research on crowding has dealt with what we label here as “spatial crowding”, i.e., impairment of the recognizability of a target object by other nearby objects. Here, we use the term “crowding” more broadly to indicate both spatial and temporal processes. Spatial crowding can be thought of as the integration of features over too large an area. This excessive integration can take place between a target object and neighboring “flanker” objects, or it can take place between the parts of a single object. We call the former category “between-objects” crowding and the latter category “within-object crowding.” In addition to the effects of between-objects crowding, both words and faces suffer from within-object crowding. In the peripheral visual field, the letters of a word and the parts of a face (nose, mouth, eyes) crowd one another and impair recognition. By increasing the spacing between letters (or facial parts), crowding is attenuated. Below is an example of spatial crowding. Fixate on the cross and try to read the words.

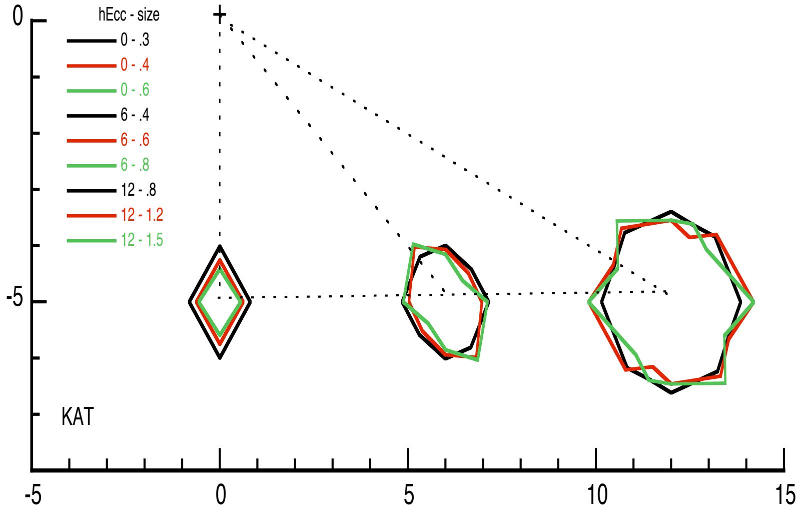
Taxi

+

T a x i

The words above and below the cross are identical and they are at the same eccentricity from the fixation point. However, the word on top is difficult to read, whereas the bottom word is easy to identify. One can see that the top word has a capital ‘T’ as the first letter, but the inner letters are jumbled and impossible to distinguish.

During “within-object” crowding, the parts of an object are too close to one another for the brain to be able to spatially isolate them. One can think of the brain as having different sized “isolation fields” that integrate the features within their bounds (Pelli et al., 2004). In the case of words, each isolation field integrates the different features of the letter within its scope. Figure 1 illustrates this concept.



It

Figure 1. Data collected in our lab using a paradigm developed by Gordon Legge (Legge, Mansfield, & Chung, 2001). Letter triplets were presented at different distances from the fixation point. The middle letter was identified at an 80% correct threshold while varying the spacing of the two adjacent flanker letters. The resulting data points represent the critical spacings at which the flanker letters no longer interfere with target identification (for an 80% correct threshold). Connecting the points maps out the shape of the isolation fields at different distances from the fovea. The size of the isolation field increases with distance from the fovea and it bears a circular, or slightly elliptical, shape. Also, notice that the print size does not effect the critical spacing (the different colored-lines all overlap), which is a typical effect of spatial crowding (Pelli et al., 2004).

In the case of temporal crowding, the flankers are other stimuli presented before and after the target within very short time intervals (on the order of milliseconds). Examples of temporal crowding include quickly scanning faces in a crowd or trying to read several words in one glimpse. Treisman & Schmidt (1982) describe temporal crowding as “a speed limit on object processing in a serial attentive scan.” Pelli et. al. (2004) agree with Treisman in supposing temporal crowding to be bottleneck due to attention. This conception fits well with the “spotlight theory of attention,” first advanced by William James in 1890. For James, there is a temporal constraint on what we can attend to. In James’ analogy, it takes time to direct our spotlight of attention to a new object, and anything presented before the spotlight can be redirected is not perceived.

Temporal and spatial crowding are both obstacles to object identification. Temporal crowding determines a maximum “speed limit” at which successive objects can be perceived. Spatial crowding determines a minimum “space limit” that must exist between the parts of an object (or between two or more whole objects). To recognize several peripheral objects presented serially in time, both temporal and spatial crowding need to be avoided. Each object must be isolated in time and the parts of each object must be isolated in space for recognition to occur.

This study investigated whether spatial and temporal crowding are perceptual constraints that are statistically independent of one another, or whether the two effects interact. A helpful example of independence is that of a coin toss. Suppose you are tossing two coins simultaneously, one in your right hand and one in your left hand. Your chances of getting both heads for any given toss is equal to the probability of getting a head in your right hand (.5) times the probability of getting a head in your left hand (.5). Therefore, your probability of getting both heads is .25. Your right and left hand do not affect one another during the coin toss. They are independent, and the probability of getting both heads is the product of the two individual probabilities.

We investigated the applicability of an independence model to face and word identification. In this case, an independence model predicts that the probability of someone correctly identifying an object is equal to the probability of identifying it in space (independent of time pressure) multiplied by the probability of identifying it in time (independent of space pressure): In symbolic terms, $\mathrm{P}(A \text{ « } B) = \mathrm{P}(A)\mathrm{P}(B)$. The results of this empirical test indicate whether face and word recognition are similar or different with respect to the spatial and temporal crowding mechanisms.

Methods

Subjects

One subject (also the author) with corrected-to-normal acuity performed these experiments. All experiments were performed on an Apple Power Macintosh computer using MATLAB software with the **Psychophysics Toolbox** extensions (Brainard, 1997; Pelli, 1997). The background luminance was set to the middle of the monitor range, about 18 cd/m. All stimuli were shown at full contrast. The paradigm for our computer program was Rapid Serial Visual Presentation, or "RSVP" (Potter, Kroll, and Harris, 1980). In RSVP, stimuli are presented sequentially to the observer.at the same point in space. In this experiment, all the stimuli were presented at 10 degrees eccentricity "due south" of the fixation point in the inferior visual field.

Stimuli

Two types of stimuli were used, words and faces. Face stimuli were gathered from the Parker Brothers 1991 card game "Real People." All the faces were black and white head shots. We excluded faces with mustaches or with hair falling over the eyes or brow. The purpose of this exclusion was to weed out faces that could be recognized by a single outstanding feature. Cards were scanned at a resolution of 1200dpi and cropped using Adobe Photoshop 7.0. Each face was individually cropped so as to include the section of the face between the eyebrows and the chin, excluding the facial outline (again, minimizing outstanding single features). All cropped images were exactly square in shape.

For both words and faces, the set sizes were equal to four. Originally, three different sets sizes were going to be used: 4, 26, and 250. However, since an analysis of our data for the set size of four showed a clear negation of our hypothesis, we chose not continue with the larger set sizes. Word stimuli were four-letter words picked randomly from the list of Kucera & Francis, 1967. The font used was "Helvetica" (Microsoft Word, 2001).

Task

Two experiments were performed to generate two different types of curve, the "tradeoff curve" (Figure2), and the psychometric function (figures 3-6). However, the task required of the observer was the same for both experiments. The exact same paradigm, with one light modification, was used to collect all the data presented here.

The fixation point was a 0.15 deg white square that was present for the entirety of each trial. Stimuli were presented at 10 deg eccentricity in the inferior visual field. For a given "run", the stimuli were either all words or all faces chosen at random from their respective set of four. Each run consisted of 15 trials. In each trial, 6 stimuli were presented sequentially. The experiment was self-paced and self-scored. The subject initiated each trial himself by pushing the space bar, and scoring was done orally. The observer read the words aloud as they appeared during the trial. For faces, the observer spoke the names of the faces aloud (having memorized a list of four names arbitrarily assigned to the faces). A tape recorder was used to record the oral response and the number of incorrectly identified words or faces was entered into the computer after each trial.

In order to manipulate spatial crowding with the individual faces, we varied the size of the faces, effectively changing the spacing between parts. The eyes, nose, and mouth were defined as the relevant face parts for our purposes, following Martelli et al., 2005. Spatial crowding effects are independent of the size of the stimuli (Pelli et al., 2004). Therefore, varying the size of the faces (or words) did not bring another variable into play. Spacing mediates crowding, not size. The distances between facial features was measured using a ruler and pencil for each of the four faces. The numbers were averaged for each face [(nose to eye1+nose to eye2+nose to mouth)/ 3], and then averaged across all four faces. This number, divided by the length of one side of the square (each cropped photo was presented to the observer in the shape of a square), yielded a ratio of the spacing between facial features to the width of the stimulus. We multiplied the threshold widths of the face stimuli by this ratio to determine the threshold spacing. For words, the center-to-center spacing between individual letters was used as the standard for spacing between parts. These definitions for "parts" (i.e. letters for words, eyes, nose, and mouth for faces) have been supported by another study (Farah, Wilson, Drain, & Tanaka, 1998).

Temporal crowding was manipulated by increasing or decreasing the rate of presentation of sequential stimuli. Our MATLAB program did this automatically by means of the QUEST staircase algorithm (Watson & Pelli, 1983; King-Smith, Grisby, Vingrys, Benes, & Supowit, 1994). Each run kept stimulus size constant and varied stimulus duration, or vice versa. Criterion performance was set at 80% correct. *The QUEST algorithm manipulated the dependent variable and honed in on the threshold value at 80% correct.* Data from runs with lower than %75 correct performance were not used. The experimental task was always the same, whether the run tested faces or words.

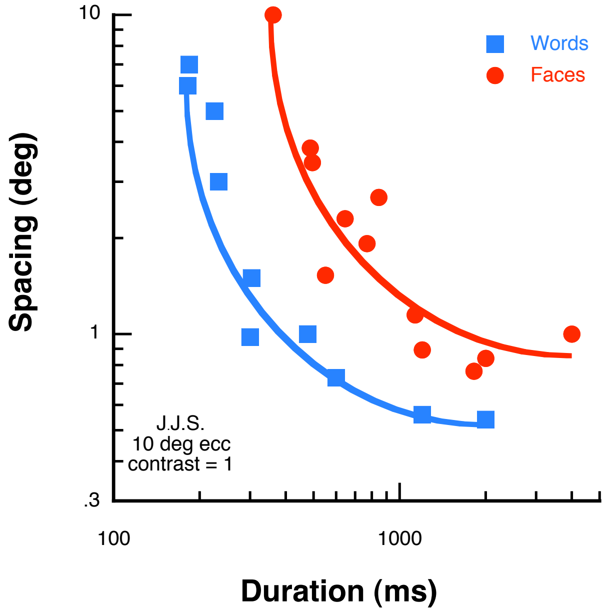
The psychometric functions were collected in the same manner as above with one minor adjustment. Instead of having the MATLAB program manipulate one of the variables (duration or spacing), we had it keep both variables constant. We treated the observer's percent-correct as the dependent variable. QUEST did not hone in on an 80% correct threshold, but presented the exact same stimuli (in terms of duration and spacing) regardless of observer performance.

Modeling

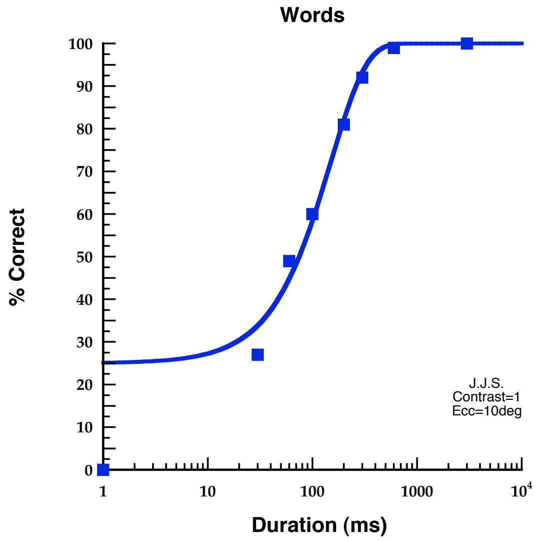
The curves for the psychometric functions (figures 3-6) were fitted to the data points using Weibull functios. The curves predicted by the independence model (figures 7-8, in black) were generated by multiplying the two corresponding psychometric functions. That is, percent correct as a function of spacing, **f(s)**, was multiplied by percent correct as a function of duration, **f(t)**, to yield a predicted space-time tradeoff curve, **s(t)**. Only the solutions equal to .8, or 80% correct, were used to generate points for the independence model curves (becasue that threshold corresponds to the experimental task).

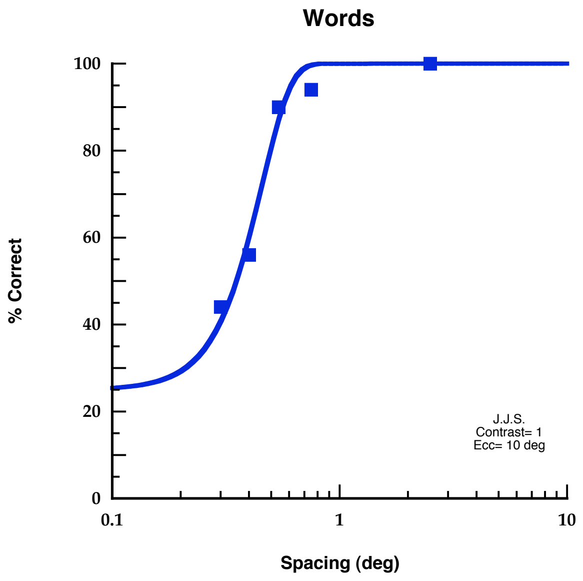
Results

Figure 2, below, shows the spacing-duration tradeoff curves for both words and faces.

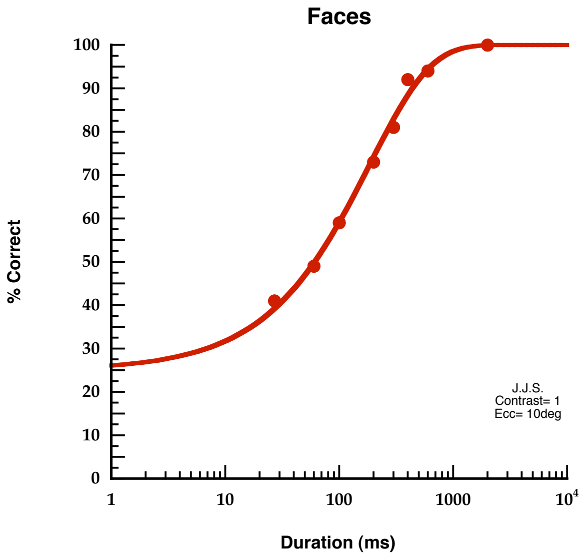


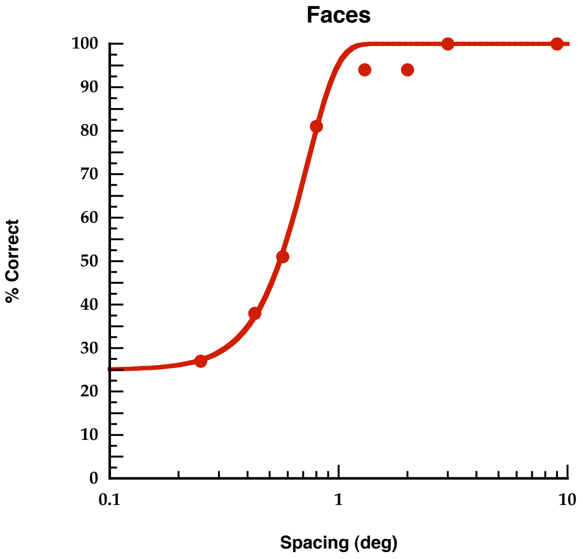
The blue squares represent runs in which words were used as the stimulus. The blue line is a curve fit by hand to the data. The red circles show the data for faces. The red line is the same curve, shifted up and to the right. The horizontal axis is the duration of the stimulus (one word or one face) measured in milliseconds. The vertical axis is the spacing between parts of the stimulus, measured in degrees of visual angle. Both axes are measured in log units. Figures 3 and 4 (below, left and right) show the psychometric functions for words..





Figures 5 and 6 (below, left and right) show the psychometric functions for faces.





In all four psychometric functios, the x-axes measure either spacing (in degrees of visual angle) or the duration of the stimulus (in milliseconds). The y-axes, in all four graphs, measure the percent correct obtained by the observer. Duration and spacing are measured in log units. Percent-correct is measured in linear units. Figure 7, below, shows the spacing-duration tradeoff for words (same as in figure 2) along with the curve (in black) that the independence model predicts.

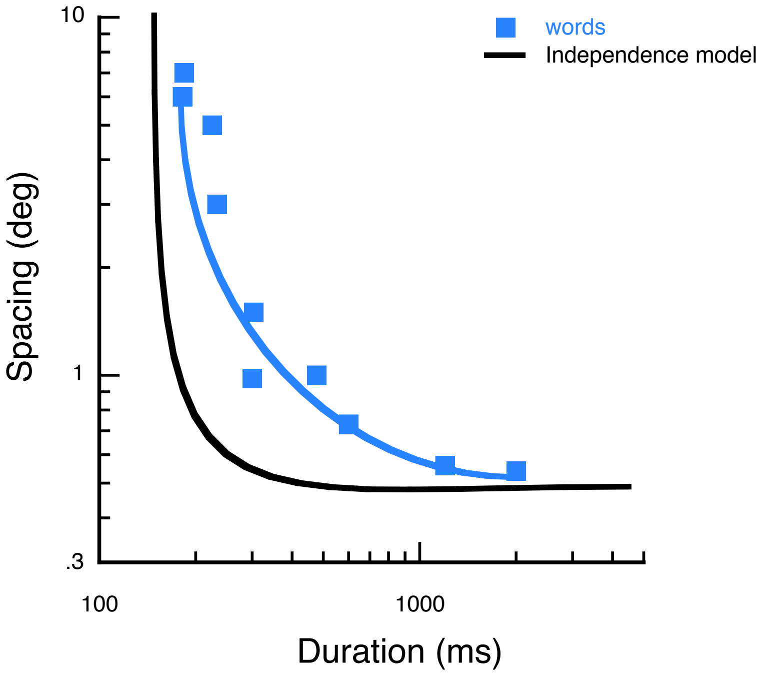
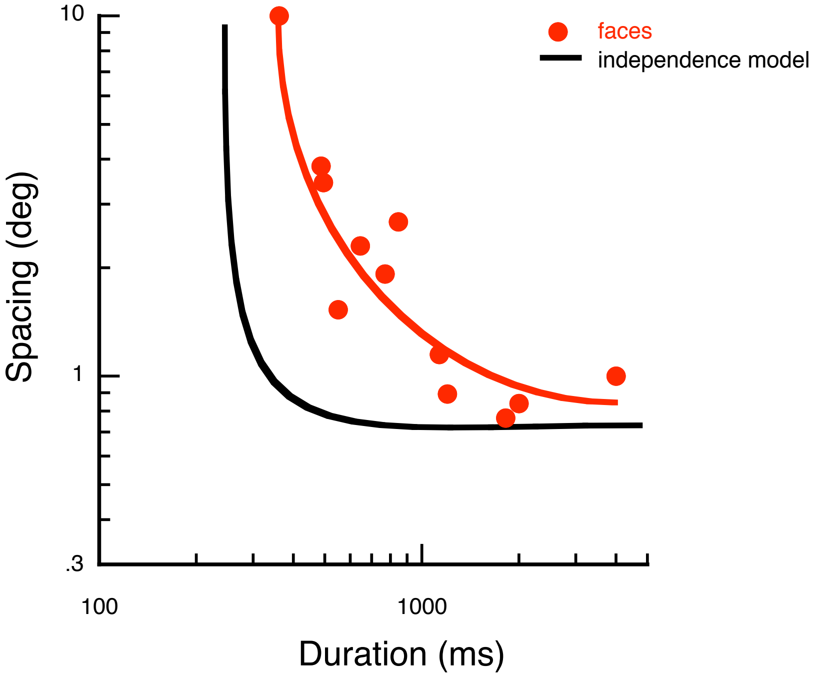


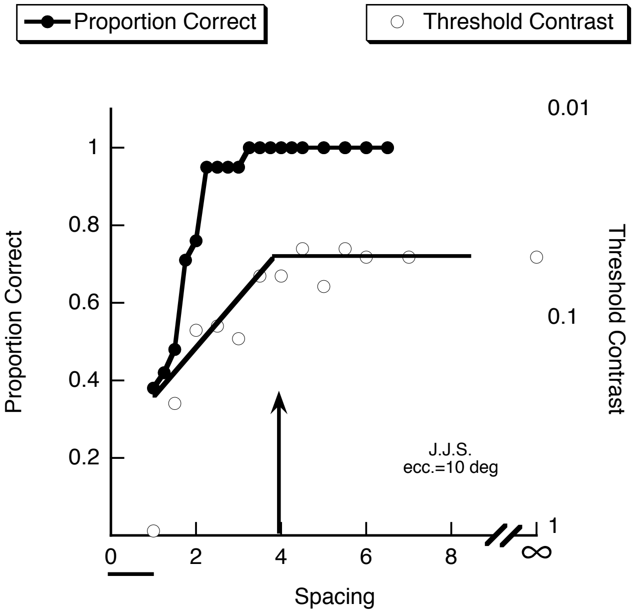
Figure 8, below, shows the spacing-duration tradeoff for faces and the curve (in black) that the independence model predicts.



Discussion

Figure 2 illustrates the tradeoff between isolating an object in space and isolating an object in time. You will notice that as duration decreases, spacing increases, and vice versa. When the observer has ample time to identify each word or face, he can tolerate more spatial crowding and still get 80% correct. As the time pressure increases (going from right to left on the x-axis), the observer needs progressively larger spacings to maintain 80% correct performance. Notice the two asymptotes for each curve. The horizontal asymptote corresponds to the smallest spacing at which the stimulus can be perceived. Given ample time (two seconds or more), this asymptote shows how small the spacings can be for the observer to still identify the object.

We name the horizontal asymptote the "threshold spacing", as opposed to the "critical spacing" mentioned earlier. Bouma (1970) showed that critical spacing—operationally defined as the distance at which recognition is no longer impaired by adjacent flankers—is roughly one half of the viewing eccentricity. The spacings for the horizontal asymptotes in figures 7 and 8 are much smaller—by a factor of 4 or 5—than Bouma's predicted critical spacing at 10 deg eccentricity (i.e. 5 deg). Traditionally, the word "critical" denotes a break point—the knee joint—in a graph. Observe the graph on the left, which is a purely didactic example. It shows data from a 26-way letter identification task in which a center target letter was flanked by two adjacent letters on each side. In one condition, contrast equaled 1 and proportion correct was the dependent variable. In the other condition, threshold contrast was



the dependent variable. At about four degrees spacing between letters, there is a clear break point in the graph for both conditions. Four degrees is the critical spacing for this observer.

The tradeoff curves in Figure 2 do not show a sharp break point, but rather a gradually changing curve with an asymptote. The horizontal asymptotes in figure 2 measure the distances at which spatial crowding spoils recognition. In the example graph above, on the other hand, the four-degree critical spacing marks the distance at which spatial crowding *no longer* exerts any effect. Above four degrees spacing, performance does no longer improves. There is a qualitative difference between these two measures, and we therefore use separate terms. "Critical spacing" marks the outer bound of the isolation field, where spatial crowding disappears. It is distinguished by a break point in the graph. The "threshold spacing"—corresponding to the horizontal asymptotes in our tradeoff curves—marks the point at which spatial crowding spoils recognition for an arbitrarily defined threshold (in this case, 80% correct). The two measures are similar, but not identical. This means that we can compare the critical spacing and threshold spacing across observers and tasks, but must keep in mind that they ought not be identical for the same observer performing the same task. With that distinction in mind, notice that the threshold spacing for words is 0.5 deg, for faces, roughly 1 deg. This is a factor of 2 difference.

The vertical asymptotes in the tradeoff curves show the smallest amount of time needed to identify the object, given ample spacing. For words, this "critical duration" is 200 ms and for faces it is roughly 400 ms. Again, this is a factor of 2 difference. In this task, words and faces have quite different critical durations. For this observer, then, faces require more spacing between parts and more time than words do, by roughly a factor of 2. Although the threshold spacings and the critical durations are not the same for words and faces, their tradeoffs are identical. The tradeoff curves do not superimpose, but they have the same slope.

The curve for faces is identical to the words curve, but shifted up and to the right. Words and faces show a similar “give and take” between the spatial and temporal crowding constraints.

The difference in threshold spacing between words and faces makes some intuitive sense. For a typical print size—say 12pt—words are quite small. At talking distance, faces take up proportionately more visual angle. It stands to reason that if an object can be recognized with such small spacings (as in the case of printed text), then spatial crowding would be a smaller effect (on an absolute scale) than objects that must take up more visual angle.

Figure 7 answers the question of time-space independence for words. Does the independence model fit the data obtained empirically? Clearly, the two curves do not match up. The independence model has a much sharper corner at the 200 to 300 ms durations. In contrast, the observer’s performance is poorer in that region, making for a shallower tradeoff. The independence model curve for words is much closer to a stepwise function, where performance suddenly changes at a single threshold value. Indeed, this result can be seen from the steep slope of the psychometric functions for words.

Figure 8 shows the independence model prediction alongside the observer data for faces. This graph looks remarkably similar to the results for words. Again, the independence model does not fit the empirical data. The independence model curve predicts better performance—smaller spacings and durations in the 300 to 500 ms range—than attained by the human observer. For faces, the difference between the independence model and the observer data is more pronounced than it is in Figure 7, for words. However, just as the curves for words and faces in figure 2 are identical, so the curves for words and faces predicted by the independence model are nearly identical as well. That is, the independence model is equally ill-suited to the data for words and faces alike. We do not see a result where the model fits the data for one type of stimulus and not the other. Neither words nor faces fit an independence model for time-space tradeoff. For both words and faces, the variables of spatial and temporal crowding are not independent, but interact.

Conclusions

Using RSVP in the peripheral visual field, we measured the proportion of words and faces correctly identified as a function of part-to-part spacing and presentation rate. We show that while faces require more time and larger spacings between parts for recognition, the space-time tradeoff is identical for both object types. For a set size of four, we found words and faces to have different critical spacings and different critical durations. Critical duration has not been previously compared for words and faces. We reject the hypothesis that the spatial and temporal isolation processes are independent. Words and faces, in respect of their temporal and spatial processes, are processed the same way in the periphery.

Acknowledgments

This research was made possible through a National Science Foundation Summer Undergraduate Research Program (SURP) at New York Univeristy's Center for Neural Science. The author worked in the lab of Denis Pelli, Professor of Psychology at NYU. He recieved invaluable help from Denis Pelli, Mariluisa Martelli, and Katharine Tillman.

References

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

Cohen et al., 2000 L. Cohen, S. Dehaene, L. Naccache, S. Lehericy, G. Dehaene-Lambertz, M.A. Henaff and F. Michel, The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients, *Brain* **123** (2000) (Pt. 2), pp. 291–307.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, *115*(2), 107-117.

Ekman, P. (1992). Are there basic emotions? *Psychological Review, 99*(3), 550-553.

Farah, M. J. (1991). Patterns of co-occurrence among the associative agnosias: Implications for visual object representation. *Cognitive Neuropsychology*, *8*(1), 1-19.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3*(2), 191-197.

Goren, C. C., Sarty, M., & Wu, P. Y. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, *56*(4), 544-549.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10-11), 1409-1422.

Howard et al., 1992 D. Howard, K. Patterson, R. Wise, W.D. Brown, K. Friston, C. Weiller and R. Frackowiak, The cortical localization of the lexicons. Positron emission tomography evidence, *Brain* **115** (1992) (Pt. 6), pp. 1769–1782.

Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, *40*(1-2), 1-19.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302-4311.

King-Smith, P. E., Grigsby, S. S., Vingrys, A. J., Benes, S. C., & Supowit, A. (1994). Efficient and unbiased modifications of the QUEST threshold method: Theory, simulations, experimental evaluation and practical implementation. *Vision Research*, *34*(7), 885-912.

Kucera, H., and Francis, W.N. (1967) Computational Analysis of Present-Day Modern English. Providence, RU; Brown University Press.
James, W. (1890). *Principles of psychology* (Vol. 1, Chap. II). New York: Holt.

Legge, G. E., Mansfield, J. S., & Chung, S. T. L. (2001). Psychophysics of reading. XX. Linking letter recognition to reading speed in central and peripheral vision. *Vision Research, 41*(6)*,* 725-743.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision*, *4*(12), 1136-1169

Pelli, D.G., Martelli, M., & Majaj, N.J. (2005) Are faces processed like words? A diagnostic test for recognition by parts. *Journal of Vision*, 5, 58-70.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics. Transforming numbers into movies. *Spatial Vision*, 10(4), 437-442.
Petersen et al., 1990 S.E. Petersen, P.T. Fox, A.Z. Snyder and M.E. Raichle, Activation of extrastriate and frontal cortical areas by visual words and word-like stimuli, *Science* **249** (1990), pp. 1041–1044.

Potter, M.C., Kroll, J.F., and Harris, C. (1980) Comprehension and memory in rapid, sequential reading. In R.S. Nickerson (Ed.) *Attention and Performance, Vol. 8*. Hillsdale, NJ: Erlbaum.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14*(1), 107-141.

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception and Psychophysics*, *33*(2), 113-120.