Counting Features:
quantifying discrete parts in visual object identification

Henny Admoni
J. L. Miller Great Neck North High School
35 Polo Road
Great Neck, NY 11023

Mentor: Professor Denis Pelli
Psychology and Neural Science
6 Washington Place, Room 959
New York University
New York, NY 10003

Counting Features

Abstract

People visually detect simple gratings, made of alternating bars of black and white, by detecting the bars independently. More generally, it seems that the visual identification of ordinary objects is based upon the independent detection of several *features:* discrete components of the image that are detected independently of each other (Robson and Graham, 1981; Pelli, Farell, and Moore, 2003). However, the number of features we use to identify a complex image remains unknown. Assuming that identification requires the detection of a certain number of features, and that features are detected independently over time, the probability of identification will grow as a binomial function at a rate determined solely by the number of features required. Thus, measuring the proportion of correct identifications as a function of duration should reveal the number of features used by the observer. The accuracy of this method was confirmed in specific cases where the number of features used was already known, and successfully applied to key examples of the general case.

Introduction

Vision is arguably the most important human sense, and has received more study than any other. Despite this, the complex process of visual perception remains poorly understood. The brain identifies objects by separately detecting and then integrating discrete visual components, called *features* (Pelli, Farell, and Moore, 2003). Features are independently detected components of the image. The number of features in most complex images is unknown.

Past work has successfully modeled object detection by assuming first (A1) that features are detected independently over space and time and second (A2) that detecting any one feature is enough to detect the object (Brindley 1965; Watson, 1979; Robson and Graham, 1981). Consider a brief presentation of an image that affords a single glimpse of duration t_{glimpse}. If the object contains six features, we may imagine this as analogous to flipping six appropriately weighted coins to model the probability of detection for each feature. We suppose that landing heads up corresponds to detecting a feature. Allowing another glimpse corresponds to flipping all of the coins again. We could then count the heads again and add up the total across glimpses. The probability of detecting the object is equal to the probability of getting at least one head, because only one feature is required to determine the presence of an object. Part of assumption A1 is that extended presentation is equivalent to a series of glimpses, each of duration t_{glimpse}, which is expected to be on the order of 100 ms (Cornsweet, 1970). This model works very well for the detection of gratings (Watson, 1979; Robson and Graham, 1981).

In this study the detection model is extended to the identification of a visually complex object by adding the third assumption (A3) that identification requires k feature

detections. Objects presumably vary in complexity by the number of features they contain, and may vary in the number that are required for their identification. We suppose that an object is correctly identified if k or more features are detected; otherwise the observer guesses blindly. For simplicity, assume that all features are detected with equal probability.

With those three assumptions, the probability of identification as a function of number of glimpses was plotted. The number of glimpses $\frac{t}{t_{\text{glimpse}}}$ is determined by stimulus duration t, divided by the duration of one glimpse, t_{glimpse}. Note that in Figure 1 the horizontal axis is a number that represents a ratio between these two other quantities. Thus, glimpse duration in itself does not affect the slope, which is the important element of the model function. The vertical axis represents the probability of correct identification, p. The slope of each graph, on a log scale, depends solely on k. To learn the value of k, it is sufficient to simply measure correct identification as a function of stimulus duration, plot it against number of glimpses, and find the corresponding model slope.

For each feature, for each glimpse, the probability of detection p corresponds to flipping a weighted coin that lands heads up with that probability p. Bernoulli's formula gives the probability p_i of exactly i heads among m coin flips

$$p_i = \frac{m!}{i!(m-i)!} p^i (1-p)^{m-i} \qquad (1)$$

Each coin flip corresponds to a potential feature detection, and each coin that lands "heads" is a detection; i is the actual number of feature detections. The potential number of feature detections is $m = n\, n_{\text{glimpse}}$ where n is the number of features in the object and

$n_{\text{glimpse}} = \dfrac{t}{t_{\text{glimpse}}}$ is the number of glimpses. By assumption A2 the probability of detection p_{det} is the probability of not having zero detections, plus the probability g of guessing blindly when there are zero detections,

$$p_{\text{det}} = (1 - p_0) + g\,p_0 = 1 - (1 - g)\,p_0 \qquad (2)$$

where p_0 is the probability of zero detections, which is equal to p_i when $i = 0$. By assumption A3 the probability of identification p_{id} is the probability of not having fewer than k detections (i.e. the probability of having k or more detections), plus the probability of guessing blindly when there are too few detections:

$$p_{\text{id}} = 1 - \sum_{i=0}^{k-1} p_i + g \sum_{i=0}^{k-1} p_i = 1 - (1 - g) \sum_{i=0}^{k-1} p_i \qquad (3)$$

The detection and identification probabilities are equal when k is 1.

The probability of identification p_{id} is graphed in Figure 1 as a function of the number of potential detections m. Recall that this value is the product of the number of features and the number of glimpses, $m = n\,n_{\text{glimpse}}$. Probability of identification increases with duration. The steepness of the curve depends only on k, the number of features required to identify the image. To plot human data on these scales it is helpful to estimate the duration of a glimpse; however, since the horizontal scale is logarithmic, changes in the estimated duration of a glimpse will merely shift the data left and right horizontally without affecting the slope.

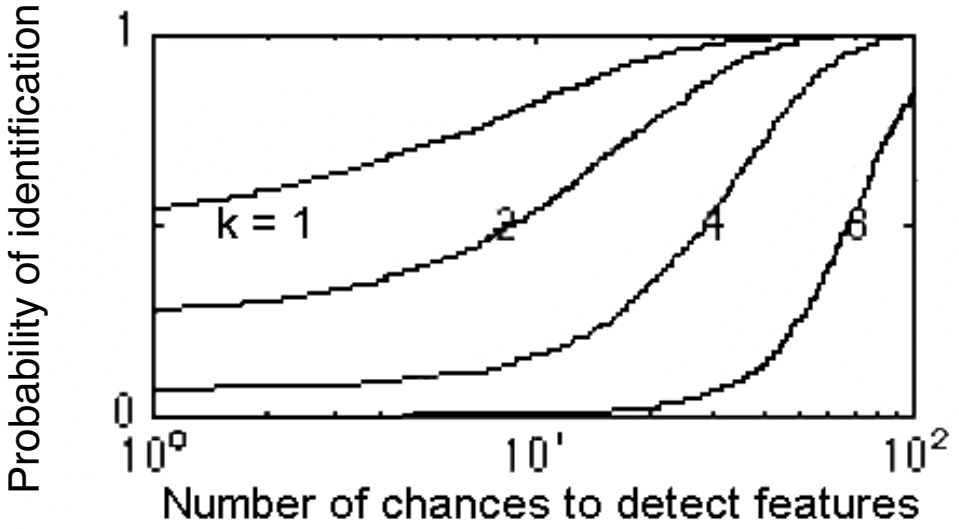


Figure 1: Probability of identification as a function of number of chances to detect features. Note the increasing steepness of the function as more features are required. Assumed value for guessing parameter from Equation 2, g, was zero.

This model is broadly consistent with the feature-integration theory of attention, which asserts that features are registered early and automatically by the brain, while objects are identified at a later stage, with focused attention (Treisman and Gelade, 1980).

Gabors, special grating patterns, were used for the initial experiment to test the assumptions, as it has already been established that small patches of a grating pattern are detected independently over time and space (Watson, 1979; Robson and Graham, 1981). In the second experiment, facial displays of emotion were used to discover how many features observers need for identification of common images.

Method

Stimuli

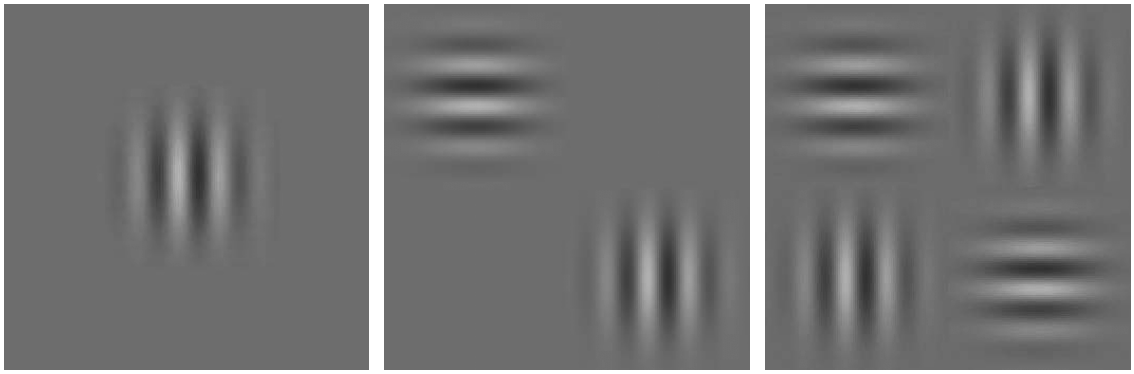
This test was run on a gamma-corrected ViewSonic E790B computer monitor that displays 72 pixels per inch and has a background luminance of 14.7 cd/m^2. A video attenuator that drives only the green gun of the monitor was used to display all stimuli. Color computer monitors utilize three electron guns displaying red, green, and blue in differing proportions to create a wide range of colors. Each gun is assigned a numerical value that represents the relative amount of that color desired. When using all three guns in a display, it is impossible to achieve the fine gradations of contrast that are necessary in the presentation of the image in this experiment. Thus, the red and blue guns were disabled, and only the green gun fired, making the screen appear green but allowing for very subtle changes in contrast. The images presented in this experiment were drawn on the screen using the programming language MATLAB (Mathworks), and the Psychophysics Toolbox (Brainard, 1997; Pelli 1997). The observers' viewing distance was 50 cm.

The stimuli were gabors and emotion faces. A gabor is a type of grating pattern, a sine wave seen through a Gaussian window. The sine wave has three cycles per degree and appears as a three-bar grating with soft edges (Figure 2a). Gabors allowed the number of features required for identification to be controlled (for instance, doubling the number of gabors doubled the number of features necessary to identify that image).

Three gabor sets were used in this experiment: Indy A had one gabor, oriented either horizontally or vertically; Indy B had two gabors, each oriented either horizontally or vertically; Indy C had four gabors, each oriented either horizontally or vertically. Thus,

Indy A had two images in the set, Indy B had four images in the set, and Indy C had sixteen images in the set (Figure 2).

To correctly identify an image in the set, every gabor pattern must be detected. For instance, in Indy C, though the orientation of three gratings may be determined, there is always the variability of the last grating to be taken into account. Though the chances at random guessing are improved in that situation, fewer than k features ($k = 4$, in the Indy C case) are detected and thus the observer is reduced to blind guessing.



(a) (b) (c)

Figure 2: These three images are samples of the (a) Indy A, (b) Indy B and (c) Indy C sets used in this experiment.

The face set consisted of four photographs from the Ekman and Friesen series cropped above the eyebrows and below the chin to a size of 6.35 centimeters by 6.35 centimeters. These photographs have been widely used in cross-cultural studies and more recently in neuropsychological research. Each photograph showed Professor Ekman displaying one of four primary emotions: happiness, sadness, fear and disgust (Figure 3).

Figure 3: The four Ekman emotional faces, from left to right, display happiness, sadness, fear, and disgust.

Observers and their learning curves

The observers were four women between the ages of 17 and 36: HA, JE, AM and LA. Each had normal or corrected vision. HA is the author.

A learning curve is defined as an observer's efficiency for identifying characters from a new alphabet. It increases quite rapidly upon initial exposure to that alphabet; after 2,000 trials the observer becomes an "expert" at the alphabet and the efficiency rate increases much more slowly (Pelli, Burns, Farell, & Moore, 2003). Only HA and LA participated in the gabor section. Their first 2,000 Indy trials were discarded as part of the learning period.

All four observers participated in the Ekman emotion faces section of the experiment. For this portion, there was no need for observers to undergo a learning period because the "alphabet" used was not new. The images display basic emotions that are encountered in daily life, and thus the observers have already reached a plateau in their learning curve from basic human interactions.

Contrast

Contrast is the difference in luminance between the lightest and darkest areas of an image. The contrast of an image affects the time necessary to detect that image. No

matter how brief a single presentation of an image may be, the afterimage will persist for several hundred milliseconds if the stimulus is of sufficient contrast to be easily identified (Sperling, 1960). Threshold is observer specific, and it is therefore necessary to lower the contrast to just above the observer's threshold.

We used the QUEST program (Watson and Pelli, 1983) to determine contrast threshold, defined as the point at which contrast is so low the observer can no longer detect an image. This program estimates threshold by interpreting the observer's data for that run (30 trials). If the observer correctly identifies the stimulus, QUEST lowers the contrast, making the task more difficult. If the observer incorrectly identifies the stimulus, the contrast is raised. At the completion of the run, the program returns the contrast at which the observer could identify the stimulus correctly 82% of the time at 400 ms., a duration which was chosen arbitrarily. The observer-specific contrast was used at all durations.

Procedure

Each observer was provided a consent form to sign prior to experimentation. The observer was then instructed to observe the center of the screen and to match the image they saw appear with an identical image in the response screen. The observer was informed that at times the image would flash too quickly to see; however, every trial required a response, even if it was only a blind guess.

Following these instructions, the observer was placed directly in front of the computer screen. The program presented either an Indy letter or an Ekman emotion face at low contrast for some duration that varied from 25 to 3200 ms. This range of durations

elicited the full spectrum of responses from blind guessing at the shortest duration to a correct identification every time at the longest duration.

Following the stimulus presentation, an answer screen appeared that displayed the stimulus among the set of images to which it belonged. The observer used the mouse to select one image, and the program would beep if that response was correct. Nothing would occur if the response was not. This trial ran 30 times at one duration, then the process would repeat for the next duration. Four trials were administered in order of lengthening duration, then all four were repeated. This comprised one run. In one sitting the observer would compete two runs, for a total of 480 trials.

Results

Indy alphabet

Figure 4 shows the results for two observers identifying Indy letters superimposed on the model curves for one, two and four features, represented by the dashed lines. Note how well the model matches the human data. Table 1 compares the actual and model derived number of features required to do this task, showing perfect agreement.

Table 1: Predicted v. actual number of detected features required for each image identification.

Image	**How many images in set**	**Predicted *k***	**Actual *k***
Indy A	1	1	1
Indy B	4	2	2
Indy C	16	4	4
Emotion Face Set	4	??	3

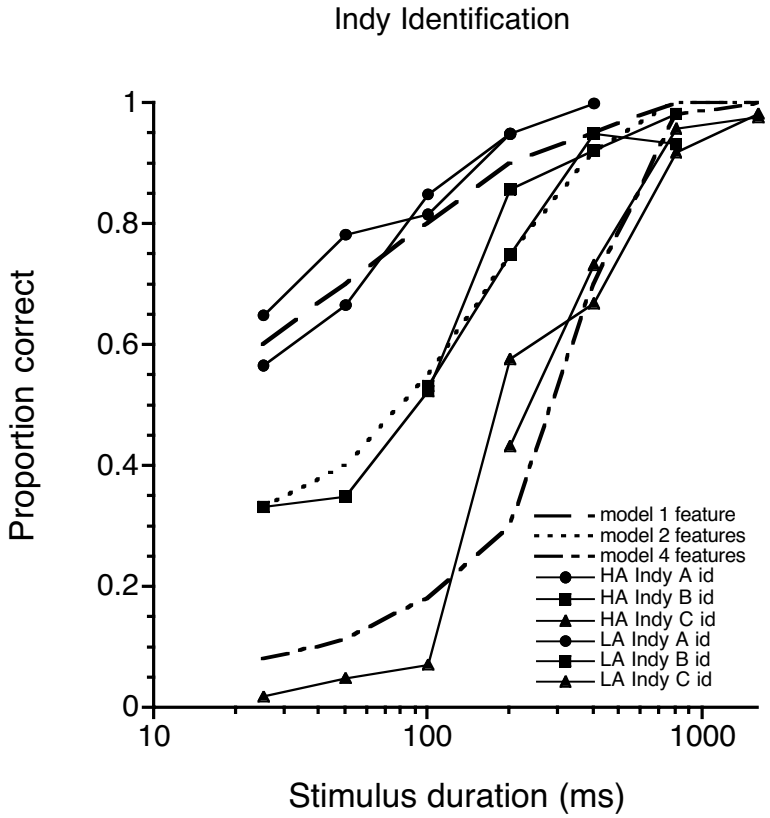


Figure 4: Identification of Indy letters by LA and HA (solid lines) and the theoretical model (dashed lines). Each shade represents a different number of features required, *k*. (g = 0; $t_{glimpse}$ = 0.1 s)

Ekman's face set

Figure 5 shows the results for the identification of facial expressions by four observers. The solid lines superimposed on the model indicate that observers use 2 to 4 feature detections to correctly identify a facial expression.

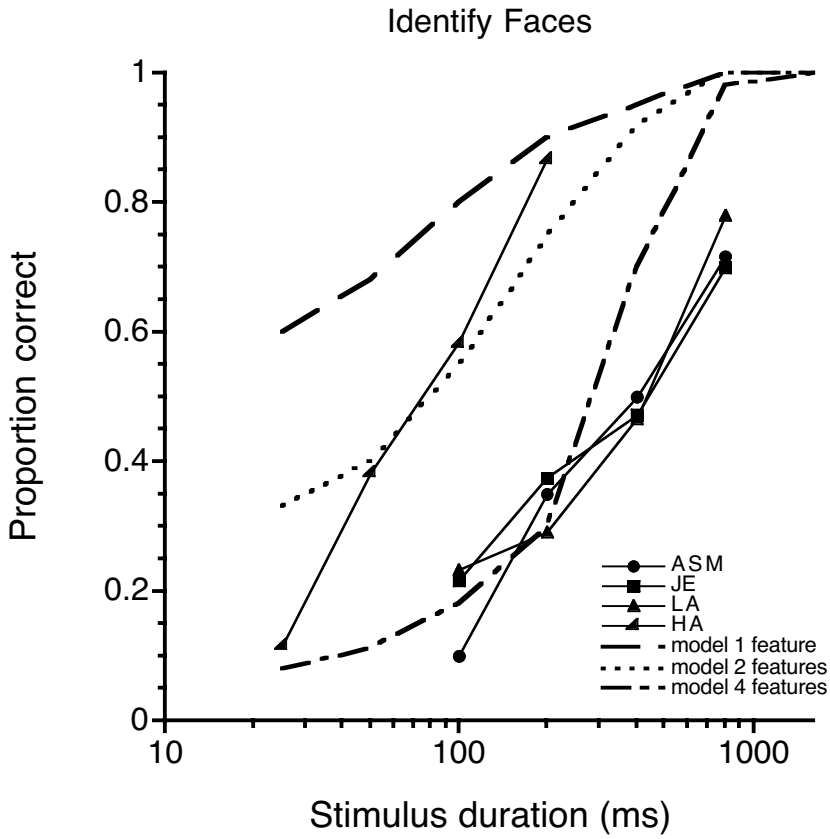


Figure 5: Identification of facial expression by four observers (solid lines) and the model (dashed line). (g = 0; $t_{glimpse}$ = 0.1 s)

Discussion

Independent feature detection has previously been affirmed for detecting gratings (Robson and Graham, 1981) and identifying words (Pelli, Farell, and Moore, 2003). The results show that this concept can be extended to identifying complex objects with multiple features, such as the Indy characters and facial expressions.

Using an artificial alphabet to control for the number of features available, a psychometric test was successfully created to measure how many feature detections, k, are needed to correctly identify a complex image. Comparing the model to the human observer data showed an excellent match, and thus validates the measurement of probability versus duration as a way of estimating the number of features used by the observer. Table 1 shows that there was perfect agreement between the predicted and actual number of features used, for experiments with one-, two-, and four-feature Indy letters. This agreement validates the model.

In this new psychometric function, the slope of the probability of identification as a function of log duration counts the number of features used by the observer.

This model is universal and can be applied to identification of both simple and complex. Results for identifying facial expressions are also fit by the model, and consistent among most observers. The results indicate that three features are required to identify the expressions on these faces.

Conclusion

The probability summation model that has been so successful in explaining feature detection is extended here to account for object identification by supposing that a certain number of features must be detected in order to identify an object. A new psychometric test is presented in which plotting the slope of the probability of identification as a function of log duration allows the determination of the number of features used by the observer. This model can be applied to the identification of any complex image. Results for identifying letters composed of gabors validate the model. Results with faces indicate that observers use three features to identify facial expression.

Acknowledgements

I'd like to acknowledge the many sources of inspiration and motivation for this research. Interminable thanks to Professor Denis Pelli, who guided me so well from start to finish. Many praises to the entire Pelli lab – to Dr. Marialuisa Martelli, who always took the time to answer my queries; to Tracey Berger, who brought me right into the swing of things; to Michael Su for his rational and excellent advice; to Najib Majaj for always explaining things in just the right way; to Betty Kolod and Lyuba Azbel for moral support and their Ekman face set, without which the experiment (and experience) would not have been the same. Many thanks to my Intel advisor Mr. David Keith, who sat through so many revisions. My deepest appreciation to Mr. Edward Baluyut for a fine read-through. A sincere thank you to Mr. Alan Schorn and Mr. Tom Elkins, who helped me find lab placement well before the summer began. Thank you, certainly, to my father, who read through every draft of the paper, and to my family for their support and interest in this long but exciting process.

References

Cornsweet, TN. (1970) *Visual Perception*. New York: Academic Press.

Ekman, P, Friesen, WV. (1976) *Pictures of Facial Affect*. Palo Alto, California: Consulting Psychologists Press.

Erdmann, B, Dodge, R. (1898) *Psychologische Untersuchungen uber das Lesen auf experimenteller Grundlage*. Halle: Niemeyer.

Pelli, DG, Burns, CW, Farell, B, Moore, DC. (2003) Identifying letters. *Vision Research*, in press.

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

Robson, JG, Graham, N. (1981) Probability summation and regional variation in contrast sensitivity across the visual field. *Vision Research 21*, 409-418.

Sperling, G. (1960) The information available in brief visual presentation. *Psychological Monographs, 74* (11, Whole No. 498), 29.

Treisman, AM, Gelade, G. (1980) A feature-integration theory of attention. *Cognitive Psychology 12*, 97-136.

Watson, AB. (1979) Probability summation over time. *Vision Research 19*, 515-522.

Watson, AB, Pelli, DG. (1983) QUEST: A Bayesian adaptive psychometric method. *Perception and Psychophysics 33*, 113-120.