

# Learning to detect and combine the features of an object

Jordan W. Suchow<sup>a</sup> and Denis G. Pelli<sup>b,1</sup>

<sup>a</sup>Department of Psychology, Harvard University, Cambridge, MA 02138; and <sup>b</sup>Department of Psychology and Center for Neural Science, New York University, New York, NY 10003

Edited by Wilson S. Geisler, The University of Texas at Austin, Austin, TX, and approved November 19, 2012 (received for review October 23, 2012)

To recognize an object, it is widely supposed that we first detect and then combine its features. Familiar objects are recognized effortlessly, but unfamiliar objects—like new faces or foreign-language letters—are hard to distinguish and must be learned through practice. Here, we describe a method that separates detection and combination and reveals how each improves as the observer learns. We dissociate the steps by two independent manipulations: For each step, we do or do not provide a bionic crutch that performs it optimally. Thus, the two steps may be performed solely by the human, solely by the crutches, or cooperatively, when the human takes one step and a crutch takes the other. The crutches reveal a double dissociation between detecting and combining. Relative to the two-step ideal, the human observer's overall efficiency for unconstrained identification equals the product of the efficiencies with which the human performs the steps separately. The two-step strategy is inefficient: Constraining the ideal to take two steps roughly halves its identification efficiency. In contrast, we find that humans constrained to take two steps perform just as well as when unconstrained, which suggests that they normally take two steps. Measuring threshold contrast (the faintness of a barely identifiable letter) as it improves with practice, we find that detection is inefficient and learned slowly. Combining is learned at a rate that is 4× higher and, after 1,000 trials, 7× more efficient. This difference explains much of the diversity of rates reported in perceptual learning studies, including effects of complexity and familiarity.


object recognition | sensitivity | letter identification

The world is full of objects, and we spend our lives identifying them. Reading an hour a day for a year means identifying millions of letters and words. Each letter is a good basic-level object: simple, common, useful, and with its own name and shape (1–4). Identifying a letter requires two steps of visual processing: the observer first detects the letter's features and then combines them to recognize the letter (5).

However, what is a feature? Interpretation of learning studies that use traditional letters and other everyday objects is hindered by the infinite number of possible features, which include physical properties, like size and shape, as well as abstract properties, like function and beauty (6–8). To avoid this morass, we narrowly define features as discrete components of an image that are detected independently of each other (5).

When letters share features (perhaps, the vertical bar in a D and an L), detecting one feature is not always enough to tell which letter it is, so multiple features must be detected and combined for reliable identification. Both steps—detection and combination—are liable to errors that impede identification. For example, if a letter is faint or seen in dim light, a reader may incorrectly identify it because she fails to detect a feature that is present or because she spuriously “detects” a feature that is absent. Identifying an unfamiliar letter can be difficult even when all of its features have been correctly detected. For example, a novice reader may mistakenly identify a plainly visible letter, confusing the shape of one for that of another.

Whether struggling to detect or to combine, with more practice, observers fail less. They learn. Feature detection and combination can both be learned through practice (9–11).

To study features, it is helpful to use Gabors. A Gabor is a grating patch that is made by vignetting a sinusoidal grating with a Gaussian window, which restricts its spatial extent to a few bars . Gabors are fairly well matched to the receptive fields of simple cells in the primary visual cortex measured physiologically, and to the tuning of spatial frequency channels measured psychophysically. Gabors can differ in position, orientation, and spatial frequency. If Gabors are sufficiently different along these dimensions, they are detected independently and can be distinguished by a single feature detection (12, 13). Practice improves detection of a Gabor (14). This learning is specific to the trained stimulus and location (15, 16).

Tasks requiring feature combination also improve with practice. Merely detecting the presence of an object does not require combining its features, but identifying it usually does; this is because detecting any feature reveals the object's presence, but, depending on the other possible objects, usually several features are needed to specify which object is present. Fine and Jacobs (17) measured improvement with practice in identifying compound gratings, which are multifeature objects composed of several superimposed Gabors, and found that learning transferred across orientations, unlike learning in detection tasks. Likewise, Kovács et al. (18) measured improvement of search for orientation-defined contours and found that learning transferred between eyes and to other orientation-defined contours, again unlike learning in detection tasks.

There are hints that the two steps, detection and combination, may be learned at different rates. Learning of familiar letters is slow and has been attributed to improved feature detection (19). Unlike the slow learning of familiar letters, the learning of new letters is initially fast, but slows as the letters become familiar (5, 20–22). This learning might involve improvement at either step. Identification involves both detecting and combining of features, so, when identification performance improves, one would like to know how much of this learning is due to improved detection rather than improved combination of features.

Here, through the use of six variously enhanced observers performing the same letter-identification task, we dissociate detecting and combining, revealing each step's contribution to learning. Of the six kinds of observer, two are “unconstrained” and four are “composite.” Unconstrained is the traditional situation of presenting a faint target and asking the observer to identify it, with no constraints. We test both the human (H) and the ideal (I) observer in this way. The ideal is an algorithm that chooses the most probable hypothesis, maximizing expected accuracy. Composite observers are new. Of the four composite observers, two are bionic. They are human in only one of the steps. The other step is delegated to a bionic crutch, either the ideal detector or the ideal combiner. In these two cases, the two perform as a team:

Author contributions: J.W.S. and D.G.P. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

<sup>1</sup>To whom correspondence should be addressed. E-mail: denis.pelli@nyu.edu.

This article contains supporting information online at [www.pnas.org/lookup/suppl/doi:10.1073/pnas.1218438110/-DCSupplemental](http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1218438110/-DCSupplemental).

Either the human combines what the ideal detects (composite IH), or the ideal combines what the human detects (composite HI). Having broken up the task into two parts, we can also assign both parts, in distinct sessions, to the same observer, so that the human (composite HH) or the ideal (composite II) takes both steps. The bionic crutches test for double dissociation: Is the human identification process actually separable into two distinct steps of detection and identification?

Are the bionic crutches overkill? Is it not enough, for our purpose, just to note the different learning rates for tasks that do and do not require combining? No. That comparison is suggestive, but has not led to any published conclusions about distinct learning rates in separate stages. Each task yields a rate, but no one has managed to link the task and the model strongly enough to draw conclusions that distinguish the two kinds of learning in one model. Adding to the confusion, much of the perceptual learning results are for fine discrimination of one feature, like orientation, which may require combining the activity of several feature detectors, but has usually been taken to reflect learning at an early stage. Our bionic crutches provide a double-dissociation paradigm that rises above these vagaries, showing how the observed pattern of results is diagnostic evidence for independent processes. Our paradigm requires manipulations (the bionic crutches) that selectively affect the two presumed processes. The strong conclusion is well worth the bother.

To separate the steps, we need to know the letters' features; they are uncertain for traditional letters, so we use Gabor letters instead (Fig. 1). Based on the probability summation literature, we suppose that our Gabors are features, detected independently (12, 13). The juxtaposition of  $n$  Gabors creates an  $n$ -feature "letter" (23, 24). Incidentally, though Gabors are very well-suited to be the elements of our stimuli, they are not essential; simple bars might do as well. By using Gabor (or bar) letters, we can precisely specify the features that constitute each letter, while maintaining the essence of an alphabet: a set of many distinguishable objects sharing a common visual style. Our Gabor letters are similar to Braille letters in that they each consist of a binary array of features. Braille behaves well when presented visually (5, 25). Even so, the conclusions of this paper do not depend on our claim that Gabor "letters" are letters; it is enough that they are objects.

We created the IndyEighteen alphabet. In general, an Indy $N$  alphabet is the set of all possible combinations of  $N$  features. Suppose we are asked to identify a randomly selected letter from



**Fig. 1.** Eight Gabor letters. The letters of the IndyEighteen alphabet are composed of Gabors. Each of the 18 possible Gabors is oriented  $\pm 45^\circ$  from vertical and is at one of nine locations in a  $3 \times 3$  grid. When a right-tilted and a left-tilted Gabor coincide, they form a plaid, but vision still responds to them independently. We suppose that the Gabors are detected independently, so that each Gabor is a feature. With two orientations and nine locations, there are 18 possible Gabors, i.e., features. The eight letters displayed here are a randomly selected subset of the  $2^{18}$  letters in the whole alphabet. Note that within this subset, some features are common to many letters (e.g., six of the eight letters contain a right-tilted Gabor at the top right corner), whereas some features are common to just a few (e.g., two of the eight letters contain a right-tilted Gabor at the bottom left position).

this alphabet of  $2^N$  letters. Because the presence of each feature is statistically independent of the rest, all  $N$  features must be detected to identify the letter reliably. In most traditional alphabets, however, a letter can be identified without detecting all of its features.\* To better match this property of traditional alphabets, we created several eight-letter subsets drawn randomly from IndyEighteen. One such subset appears in Fig. 1. Reducing the number of possible letters makes identification easier. In general, in a subset of Indy, the features are no longer independent or equally frequent, so fewer feature detections are needed for identification, and some features are less informative than others. At the extremes, a feature may be unique to a letter and thus diagnostic of its identity, or common to all of the letters and thus irrelevant to the task of distinguishing among them (26).<sup>†</sup>

For each unconstrained or composite observer, we create a new alphabet consisting of eight IndyEighteen letters. On each trial, we ask the observer to identify a letter drawn from that eight-letter alphabet. We measure threshold contrast, the lowest contrast (faintness) sufficient to identify the letters correctly 75% of the time. We then convert threshold contrast to efficiency. Efficiency is a useful way to characterize performance of a computational task (27, 28); this pits the actual observer against the ideal observer, an algorithm that performs the whole task optimally, not constrained to taking two steps. Efficiency is defined as the fraction of the signal energy used by an observer that is required by the ideal to perform just as well. Contrast energy is proportional to the contrast squared, so the efficiency of the actual observer is

$$\eta = \frac{c_1^2}{c^2}, \quad [1]$$

where  $c$  and  $c_1$  are threshold contrasts of the actual and ideal observers.

## Results

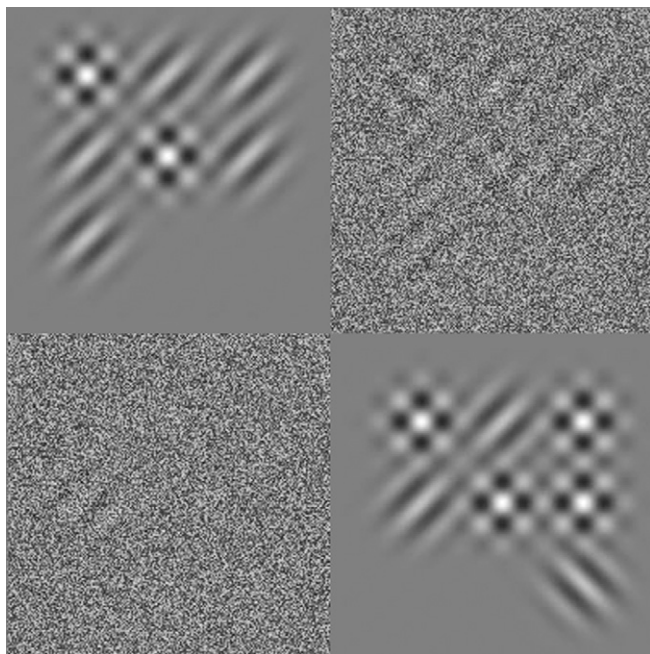
**Dissociating Detecting from Combining.** Fig. 2 shows learning for two participants, plotting threshold contrast as a function of the number of completed identification trials. (Results for all six participants appear in Fig. S1.) The right-hand vertical scale shows the efficiency corresponding to each threshold contrast. There are two graphs (Fig. 2, *Left* and *Right*), one per participant. Within each graph appear all results for that participant, unconstrained and composite. The top line (Fig. 2, solid black line) is the unconstrained ideal (I), the baseline for calculating efficiency. The bottom solid line is the unconstrained human. The other four lines, sandwiched in between, are for composites. Solid lines are fits to data, and the dashed line is a prediction derived from the other lines (Eq. 3). The vertical positions of the lines show that threshold contrast (and efficiency) are best for the unconstrained ideal, slightly worse for the two bionic crutches working together, and get worse, from line to line, as we ask the human to do part or all of the work (Fig. 2, bottom solid line). At trial 1,000, the composite-observer efficiency with the human doing just the combining (IH, 15%) is 7 $\times$  that with the human doing just the detecting (HI, 2.1%). The lines in which the human does

\*Pelli et al. (5) found that human observers need  $7 \pm 2$  feature detections for threshold letter identification for all traditional alphabets tested, over a 10-fold range of complexity. Assuming that feature count is proportional to complexity, as proposed in ref. 5, then, even if the least-complex alphabet tested had only seven features per letter, the most complex had 70 features per letter. Thus the seven features detected at the threshold for identification of a complex letter are only a small fraction of the letter's features.

<sup>†</sup>Using a Monte Carlo simulation, we determined that 4–14 feature detections are required to achieve criterion performance of 75% correct for identifying a letter from a set of eight randomly selected IndyEighteen letters, depending on the false alarm rate. A false alarm occurs when an absent feature is "detected." Sometimes, by chance, enough features are falsely detected such that the letter appears more similar to one of the foils than to the target. Additional feature detections, hits, are needed to compensate. We considered false alarm rates between 0% and 51%. At false alarm rates greater than 51%, it is impossible for the observer to achieve criterion performance, even with a hit rate of 100%.







**Fig. 4.** Stimuli. (*Upper Left*) A Gabor letter. When unconstrained, the human participant is presented with a Gabor letter faintly in noise (*Upper Right*). As the detector, the participant is presented with a single feature faintly in noise (*Lower Left*) and, as the combiner, with an imperfect set of detected features (*Lower Right*). In this last case, the high-contrast Gabors are easily seen, but are a less-than-faithful copy of the original letter's features, which makes it hard to guess what the original letter was.

## Materials and Methods

On each trial, we ask the unconstrained or composite observer to identify an IndyEighteen letter in added white noise. The letter and noise are both static, presented together for 200 ms. Testing of each unconstrained or composite observer begins with a new eight-letter alphabet and is performed in a single block of 25 runs, with 40 trials per run. Short (2-min) breaks are taken between runs, as needed, and longer (30-min) breaks are taken between blocks. The entire session was completed within 8 h in 1 d, without sleep or naps. The order of the blocks (one per task) is randomized for each observer to minimize any order effect in the group average.

**Unconstrained H: Human Identifies.** The human participant identifies, unconstrained. On each trial, we present a letter at threshold contrast (Fig. 4) to the human participant, who identifies the letter by selecting it from the response screen (Fig. 1). This trial challenges the human to identify, presumably by detecting and combining.

**Unconstrained I: Ideal Identifies.** The ideal observer identifies, unconstrained. The human participant plays no role. On each trial, we present a letter at threshold contrast. The ideal identifies the letter by choosing the most likely possibility; it compares the noisy stimulus to each letter on the response screen at the contrast of the signal, and selects the most similar (minimum rmsd; see appendix A of ref. 5). The ideal achieves the best possible expected performance, and this is the baseline for calculating efficiency.

**Composite HI: Two Steps (Human Detects, Ideal Combines).** The human participant detects and the bionic crutch (ideal combiner) combines. On each identification trial, instead of being shown the whole letter in a single presentation, the human performs 18 detection trials, one for each possible feature. (The 18 detection trials count as one identification trial in the horizontal axis of Fig. 2.) On each detection trial, the human participant reports whether the feature is present by responding "present" or "absent." The 18 present-vs.-absent decisions are recorded as an 18-bit string (1 if present; 0 if otherwise) that is passed to the bionic crutch (ideal combiner). The ideal combiner makes its selection by comparing the string received to the string for each letter on the response screen, selecting the most similar

(minimum Hamming distance) (33); this challenges the human participant to detect, without challenging combination.

**Composite IH: Two Steps (Ideal Detects, Human Combines).** The bionic crutch (ideal detector) detects and the human participant combines. On each identification trial, the crutch performs 18 detection trials. On each detection trial, the crutch selects the most probable hypothesis (present or absent), given the noisy stimulus and the frequency of that feature in the alphabet. Features judged by the crutch to be present are displayed at high contrast to the human participant, who identifies the letter by selecting it from the response screen; this challenges the human to combine, without challenging detection.

**Composite II: Two Steps (Ideal Detects and Combines).** The two bionic crutches together perform the whole task, in cascade. The human participant plays no role.

**Composite HH: Two Steps (Human Detects and Combines).** The human participant takes the two steps in separate sessions, one for each step. This trial challenges the human to detect in one session, and to combine in another session. The level of performance achieved by this two-step composite observer, HH, is computed from the measured performance of the other three two-step composites: HI, IH, and II. For this calculation, we suppose that the efficiency  $\eta$  of each two-step composite observer is the product of two factors,  $a$  and  $b$ , one for each step, and that each factor depends on whether that step is performed by the human (H) or an ideal bionic crutch (I), but it is independent of how the other step is performed:

$$\begin{aligned}\eta_{HH} &= a_H b_H \\ \eta_{HI} &= a_H b_I \\ \eta_{IH} &= a_I b_H \\ \eta_{II} &= a_I b_I.\end{aligned}\quad [2]$$

Because multiplication is transitive, we easily solve for the two-step human efficiency in terms of the others:

$$\eta_{HH} = \eta_{HI} \eta_{IH} / \eta_{II}.\quad [3]$$

This equation can be recast as a statement about thresholds, using Eq. 1 to substitute thresholds for efficiencies,

$$c_{HH} = c_{HI} c_{IH} / c_{II}.\quad [4]$$

Both equations correspond to the same dashed line in Fig. 2, using the threshold scale on the left (Eq. 4) or the efficiency scale on the right (Eq. 3). In future work, it will be interesting to study the human combination efficiency  $\eta_{HI}$ , for which we can solve Eq. 3,

$$\eta_{HI} = \eta_{HH} \eta_{II} / \eta_{IH}.\quad [5]$$

All of the terms on the right of Eq. 5 are easily accessible.  $\eta_H$  is easy to measure, and our work here suggests that, in future studies, one might assume that  $\eta_{HH} = \eta_H$ . The human efficiency of detecting  $\eta_{HI}$  seems to be conserved across many conditions, so that it could be estimated once. And the two-step efficiency  $\eta_{II}$  is easily computed by implementing the one- and two-step ideals. In this way, Eq. 5 could make it easy to routinely estimate the observer's combining efficiency  $\eta_{HI}$ .

Eq. 3 may seem odd if you did not expect the  $\eta_{II}$  term there for two-step efficiency; we can make it more intuitive by defining composite efficiency  $\tilde{\eta}$  relative to the composite ideal, II. Recall that standard efficiency is  $\eta = E_i/E$ . We now define composite efficiency  $\tilde{\eta} = E_{ii}/E$ . In this new notation, Eq. 3 becomes

$$\tilde{\eta}_{HH} = \tilde{\eta}_{HI} \tilde{\eta}_{IH}.\quad [6]$$

In words, for any observer whose efficiency is separable (Eq. 2), the overall composite efficiency is the product of the composite efficiencies of the steps. Eqs. 2–6 are all equivalent. Though the equation for  $\tilde{\eta}$  (Eq. 6) is simpler and more intuitive than the equation for  $\eta$  (Eq. 3), we chose to plot the traditional familiar efficiency  $\eta$  rather than our new-fangled composite efficiency  $\tilde{\eta}$  because they differ solely by the factor  $\eta_{II}$ , which is nearly 1.

**ACKNOWLEDGMENTS.** We thank Chris Berendes, Y-Lan Boureau, Charles Bigelow, Rama Chakravarthi, Hannes Famira, Judy Fan, Jeremy Freeman, Ariella Katz, Yann LeCun (isolating detection), Christine Looser, Najib Majaj,

Charvy Narain, Robert Rehder, Wendy Schnebelen, Elizabeth Segal, Eva Suchow, Steven Suchow, Katharine Tillman, Bosco Tjan (adding the unconstrained ideal), and Ed Vessel for helpful comments and discussion. We thank several

anonymous reviewers for many helpful suggestions. This is draft 146. This research was supported by National Institutes of Health Grant R01-EY04432 (to D.G.P.).

1. Rosch E, Mervis CB, Gray W, Johnson DM, Boyes-Braem P (1976) Basic objects in natural categories. *Cognit Psychol* 8:382–439.
2. Wong ACN, Gauthier I (2007) An analysis of letter expertise in a levels-of-categorization framework. *Vis Cogn* 15:854–879.
3. Pelli DG, Farell B, Moore DC (2003) The remarkable inefficiency of word recognition. *Nature* 423(6941):752–756.
4. Pelli DG, et al. (2009) Grouping in object recognition: The role of a Gestalt law in letter identification. *Cogn Neuropsychol* 26(1):36–49.
5. Pelli DG, Burns CW, Farell B, Moore-Page DC (2006) Feature detection and letter identification. *Vision Res* 46(28):4646–4674.
6. Treisman A (1988) Features and objects: The fourteenth Bartlett memorial lecture. *Q J Exp Psychol A* 40(2):201–237.
7. Pinker S (1984) Visual cognition: An introduction. *Cognition* 18(1-3):1–63.
8. Murphy GL (2002) *The Big Book of Concepts* (MIT Press, Cambridge, MA).
9. Gibson E (1969) *Principles of Perceptual Learning and Development* (Appleton-Century-Crofts, New York).
10. Fine I, Jacobs RA (2002) Comparing perceptual learning tasks: A review. *J Vis* 2(2):190–203.
11. Ahissar M, Hochstein S (1997) Task difficulty and the specificity of perceptual learning. *Nature* 387(6631):401–406.
12. Watson AB (1979) Probability summation over time. *Vision Res* 19(5):515–522.
13. Robson JG, Graham N (1981) Probability summation and regional variation in contrast sensitivity across the visual field. *Vision Res* 21(3):409–418.
14. Furmanski CS, Schluppeck D, Engel SA (2004) Learning strengthens the response of primary visual cortex to simple patterns. *Curr Biol* 14(7):573–578.
15. Mayer MJ (1983) Practice improves adults' sensitivity to diagonals. *Vision Res* 23(5):547–550.
16. Fahle M (2005) Perceptual learning: Specificity versus generalization. *Curr Opin Neurobiol* 15(2):154–160.
17. Fine I, Jacobs RA (2000) Perceptual learning for a pattern discrimination task. *Vision Res* 40(23):3209–3230.
18. Kovács I, Kozma P, Fehér A, Benedek G (1999) Late maturation of visual spatial integration in humans. *Proc Natl Acad Sci USA* 96(21):12204–12209.
19. Doshier BA, Lu ZL (1999) Mechanisms of perceptual learning. *Vision Res* 39(19):3197–3221.
20. Polk TA, Farah MJ (1995) Late experience alters vision. *Nature* 376(6542):648–649.
21. Chung ST, Levi DM, Tjan BS (2005) Learning letter identification in peripheral vision. *Vision Res* 45(11):1399–1412.
22. Suchow JW, Pelli DG (2005) Learning to identify letters: Generalization in high-level perceptual learning. *J Vis* 5(8):712, (abstr).
23. Levi DM, Hariharan S, Klein SA (2002) Suppressive and facilitatory spatial interactions in peripheral vision: Peripheral crowding is neither size invariant nor simple contrast masking. *J Vis* 2(2):167–177.
24. Levi DM, Sharma V, Klein SA (1997) Feature integration in pattern perception. *Proc Natl Acad Sci USA* 94(21):11742–11746.
25. Loomis JM (1981) On the tangibility of letters and braille. *Percept Psychophys* 29(1):37–46.
26. Seitz AR, Watanabe T (2009) The phenomenon of task-irrelevant perceptual learning. *Vision Res* 49(21):2604–2610.
27. Geisler WS (1989) Sequential ideal-observer analysis of visual discriminations. *Psychol Rev* 96(2):267–314.
28. Pelli DG, Farell B (1999) Why use noise? *J Opt Soc Am A Opt Image Sci Vis* 16(3):647–653.
29. Sternberg S (2003) Process decomposition from double dissociation of subprocesses. *Cortex* 39(1):180–182.
30. Lu ZL, Doshier BA (2004) Perceptual learning retunes the perceptual template in foveal orientation identification. *J Vis* 4(1):44–56.
31. Gold J, Bennett PJ, Sekuler AB (1999) Signal but not noise changes with perceptual learning. *Nature* 402(6758):176–178.
32. Michel MM, Jacobs RA (2008) Learning optimal integration of arbitrary features in a perceptual discrimination task. *J Vis* 8(2):3.1–16.
33. Hamming RW (1950) Error detecting and error correcting codes. *Bell Syst Tech J* 29(2):147–160.
34. Watson AB, Robson JG (1981) Discrimination at threshold: Labelled detectors in human vision. *Vision Res* 21(7):1115–1122.
35. Kim J, Wilson HR (1993) Dependence of plaid motion coherence on component grating directions. *Vision Res* 33(17):2479–2489.
36. Brainard DH (1997) The Psychophysics Toolbox. *Spat Vis* 10(4):433–436.
37. Pelli DG (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spat Vis* 10(4):437–442.
38. Pelli DG, Zhang L (1991) Accurate control of contrast on microcomputer displays. *Vision Res* 31(7-8):1337–1350.
39. Watson AB, Pelli DG (1983) QUEST: A Bayesian adaptive psychometric method. *Percept Psychophys* 33(2):113–120.