Supporting Information

Suchow and Pelli 10.1073/pnas.1218438110

SI Materials and Methods

Composite II′: Ruling Out Stimulus Artifacts. As a check on our work, we implemented composite II in two ways. First we created the composite observer by cascading the two bionic crutches that had already been programmed to work with the human, but now without the human. The ideal detector passes an 18-bit feature map, indicating which features were detected, to the ideal combiner. However, one reviewer noted that we create special stimuli, high-contrast letters, to test the human in composite IH, and perhaps some artifact in those stimuli is affecting our results. So we also implemented an ideal to identify the stimuli presented to the human in composite IH. This alternate implementation of composite II is formally equivalent, and gave identical results, assuring us that no stimulus artifact had intruded.

Composite HI′: Ruling Out Unwanted Human Combination Learning in Composite HI. The whole point of the bionic observer with ideal combination is to isolate the human detection step, so it is essential to rule out combination learning by the human participant; we did that in three ways: by shuffling the order of detection trials to discourage pattern learning, by testing for it after training, and by running a modified control experiment that makes such learning impossible (i.e., a modified implementation of the bionic observer).

Distributing a letter's 18 possible features over 18 detection trials does not eliminate the letter's pattern; it merely converts a brief spatial pattern into a prolonged spatiotemporal pattern, extended over 18 detection trials. In principle, the observer could improve his detection performance by learning each letter's pattern, combining information across sequential presentations.

We discourage pattern learning by shuffling the order of the 18 detection trials that constitute each identification trial; in fact, the order is irrelevant, so this does not rule out pattern learning, but would likely hinder it. Furthermore, because shuffling makes it hard to guess the location of the next feature presentation, it has the incidental benefit of extending the relevant area to be attended to include the whole letter, not just one feature. Thus, the human participant is expected to attend to the whole letter area both when he detects as part of the bionic observer and when he identifies unaided.

At the end of training, we use a recognition test to discover any unwanted combination learning. On each trial, the human participant is shown two Gabor letters, one after the other, feature by feature, and is asked to indicate which of the two was used in training. As in training, on each trial, each letter's features are presented in random order. One of the letters is old, the other new. The old letter is a random sample from the training alphabet, an eight-letter subset of IndyEighteen. The new letter (the foil) is a random sample from a specially constructed alphabet, also an eight-letter subset of IndyEighteen, which has the same features as the old alphabet, rearranged into different combinations. This new alphabet is formed by shuffling the features of the eight old letters, across letters, to create the eight letters of the new alphabet. The two alphabets, old and new, have the same number of features of each type (first-order statistics), but differ in how these features are combined (higher-order statistics). Only if the observer has learned these combinations will he be able to distinguish old from new. No such learning was found: observers correctly indicated which letter was old on 46% of trials, not significantly different from chance (50%), or from performance measured before training, 52%.

To be absolutely sure of this key point, we also created a modified implementation of the bionic observer that eliminates the possibility of combination learning by the human detector. Before, the 18 detection trials were all based on a single target letter. In the new implementation, there is still a target to be identified (by the ideal combiner), but each detection trial is based on an independently selected letter from the eight-letter alphabet, not necessarily the target. Each detection trial is conducted as a yes/no test, but scored as right or wrong. If the human detector is right, then the bionic combiner receives this feature correctly (i.e., present or absent), as in the target. If the human detector is wrong, then the combiner receives this feature wrongly (i.e., present if absent in the target; absent if present in the target). For the human, this abolishes the letter patterns (which are co-occurrences of features within a letter) while preserving the frequency of each feature in the alphabet as a whole; for the bionic combiner, this presents a feature vector reflecting human accuracy in detecting each feature.

Despite the several differences in our two implementations of detection, the outcome—separability—was the same for both the original and the variant, which suggests that the detection step is performed similarly, just as efficiently, in both cases. In the future, it will usually be enough to test for recognition after training to confirm that the pattern learning is negligible. If such a test reveals that combination learning has occurred, this modified implementation of the bionic observer with human detection and ideal combination can be used to eradicate it.

Participants. Six human participants (JS, AK, DF, CH, SS, and MC) performed unconstrained H. Four of them (DF, CH, SS, and MC) also performed in composites HI′ and IH; the other two performed in composite HI. JS is an author. The others were naive to the purpose of the experiment. All participants gave informed consent in writing. Testing of human observers was approved by the NYU University Committee on Activities Involving Human Subjects (UCAIHS).

Stimuli. Signals are IndyEighteen letters. Each letter is a gray square with a combination of several Gabors oriented 45° from vertical, placed at various locations on a 3 × 3 grid. The superimposed Gabors are orthogonal, and, despite forming a plaid, they are detected and perceived independently (1, 2). The center-to-center spacing of adjacent Gabors is 1.4× the wavelength, 1.4/*f*. A vertical Gabor pattern is

$$L(x,y) = \left[1 + c\,\sin(2\pi f x)\exp\left(-\frac{x^2+y^2}{\lambda^2}\right)\right]L_0,$$

where background luminance $L_0 = 21$ cd/m^2, spatial frequency $f = 2$ cycle/degree, spatial extent $\lambda = 0.61$ degree, and contrast c is chosen using the estimate provided by QUEST.

Noise is added independently to each pixel of the stimulus, such that the luminance of any particular pixel is the sum of the luminance assigned to that pixel by the signal and a random increment or decrement in luminance sampled from a zero-mean Gaussian distribution, truncated at ±2 SDs. The rms contrast of the noise is 0.20. There are 25.4 pixels/degree, horizontally and vertically. The power spectral density N is $10^{-4.21}$ deg^2.

Presentation. Stimuli are rendered by an Apple Macintosh computer running MATLAB in conjunction with the Psychophysics Toolbox extension (3, 4). Stimuli are displayed on a cathode-ray tube monitor, driving only the green gun to achieve 12-bit accuracy, at a background luminance of 21 candela/m^2 (5). The display resolution is set to 1,024 × 768 at 60 Hz, 29 pixels/cm. The viewing distance is 50 cm.

Procedure. Each threshold measurement is based on a run of 40 letter-identification trials. The identification trial is performed by the human, either unconstrained or as dictated by the kind of composite: two steps, one step, or none. Each correct identification is rewarded with a short beep. The observer is asked to fixate a central white dot subtending 0.10° on the monitor. The observer initiates the run by clicking a mouse. When the human acts alone or as a combiner, 1,000 ms later the stimulus appears for 200 ms, followed by a blank screen for 250 ms, followed by a noise-free response screen containing all of the letters. The observer uses a mouse-controlled cursor to select a letter from the response screen. Any response automatically initiates the next trial, 1 s later. When the human acts as a detector, he performs 18 feature-detection trials for each letter-identification trial. On each detection trial, he reports the presence or absence of the Gabor by key press. There is no detection-specific feedback; the only feedback is the identification reward at the end of the identification trial, i.e., after the 18th detection trial. The feedback indicates whether the human and ideal together chose the correct letter.

QUEST. The QUEST sequential estimation procedure provides threshold estimates over the course of learning (6). The QUEST procedure estimates from already-known information regarding both the task and observer (assumed stationary), as well as from the observer's performance throughout the run, to provide a maximum posterior probability estimate of threshold contrast, the signal contrast (ratio of luminance increment to background luminance) at which the observer correctly identifies the signal at criterion performance (75% correct). After each trial, the QUEST procedure calculates a threshold estimate. We place each new trial at the current threshold estimate. In practice, if the observer correctly identifies the signal, the next trial presents a lower contrast. If he incorrectly identifies the signal, the next trial presents a higher contrast. QUEST is initialized at the beginning of each run with log threshold estimate -1 ± 2 (±SD), β 3.5, lapse rate 0.01, and guess rate 0.125, and is updated after each identification trial.

Calculating the Slope of Learning. For each unconstrained or composite observer, for each participant, we fit a line to the data (log threshold contrast as a function of log trial) by linear least-squares regression. Extrapolating any of these rising lines makes the impossible prediction that the human will eventually beat the ideal. In fact, improvement must saturate eventually, after huge amounts of practice. Even so, Pelli et al. (7) found good straight-line fits to letter-learning data out to 50,000 trials. The ideal does not learn; it is unaffected by practice, so we display best-fit horizontal lines for I and II in Fig. 2.

SI Methods for Table 1

Here we provide the methods used to estimate the log-log slope of learning from the 13 studies presented in Table 1, top to bottom.

This Study, Composite Observer HI. The slope, −0.03, is the average across all participants and is reported in the main text.

Pelli et al. (8), Familiar Letters. Experiment 3.4 of Pelli et al. (ref. 8, p. 4,658) measured improvement in threshold contrast for the identification of a letter. Participant RA performed 2,000 trials of the identification task using familiar letters. His efficiency increased from 6% (at 40 trials) to 7.3% (at 2,000 trials). We fit a straight line, in log coordinates, to these two points using linear least-squares regression; its slope was 0.050. Because efficiency is inversely proportional to threshold contrast squared, the log-log slope of efficiency is −2× that of threshold contrast. Therefore, the log-log slope of contrast learning is $0.050/-2 = -0.0250$. Two other participants, AW and DM, performed ∼2,500 trials (in blocks of 40) of an identification task using 2×3 checkerboard patterns. Figure 10 of ref. 1 (p. 4659) shows the data. The vertical axis plots the efficiency estimated from each block. The data are fit with a straight line on log-log axes. The slope of efficiency learning is 0.076 for participant AW and 0.100 for participant DM, and so the slope of contrast learning is −0.038 and −0.050, respectively. Thus, the average log-log slope of contrast learning across the three participants is −0.04.

Furmanski et al. (9). Furmanski et al. (ref. 9, figure 2a, p. 574) show improvement in threshold contrast for the detection of a Gabor. The learning curve is the average across six participants and shows learning over the course of a month. The reported "normalized threshold" is proportional to threshold and does not affect our estimate of the slope. We fit a line, in log coordinates, to the 34 normalized thresholds reported in the figure; its slope is −0.06.

This Study, Composite Observer IH. The slope, −0.11, is the average across all participants and is reported in the main text.

Suchow and Pelli (10). The figure in Result III shows improvement in efficiency for the identification of an unfamiliar letter from the Armenian alphabet. Two participants in Suchow and Pelli (10), SAS and JWS, performed 3,000 trials of the identification task in blocks of 40 trials. The vertical axis plots the efficiency estimated from each block. The data are fit with a straight line on log-log axes. The log-log slope of efficiency learning is 0.21 for participant SAS and 0.21 for participant JWS. Thus, the average log-log slope of contrast learning is −0.11.

Pelli et al. (8), Unfamiliar Letters. We used the same method described above. Pelli et al. (ref. 8, figure 10, p. 4659) also reports seven learning curves for participants identifying unfamiliar letters. Each curve includes between 1,500 and 5,000 trials of an identification task. Participants SE, JB, and AW identified 4×4 checkerboard patterns; participants DM and AW identified Devanagari letters; participant AW identified Hebrew letters; participant JF identified English letters. The average log-log slope of contrast learning was −0.11.

Lu and Dosher (11). Lu and Dosher (ref. 11, figure 4a, p. 50) show improvement in threshold contrast for the identification of the orientation of a Gabor tilted ±8° from diagonal. This task required a fine discrimination, which was initially unfamiliar to the participants. Participants were tested at each of two criteria (70.7% and 79.3% correct) at each of eight levels of added noise (rms contrast ranging from 0 to 0.33). In the text, the authors report that at the highest level of external noise, threshold contrast improved from 0.72 (for sessions 1 and 2 of 10, coded as session 1.5) to 0.48 (for sessions 9 and 10, coded as session 9.5). We fit a line, in log coordinates, to these two points; its slope was −0.22. The slope is the same, −0.22, if the line is instead fit to the data from all sessions at the highest noise level, not just the first and last two. Lower noise levels produced more shallow slopes of learning. The average slope across all noise levels and criteria is −0.12 (ranging from −0.014 to −0.24).

This Study, Observer H. The slope, −0.16, is the average across all participants and is reported in the main text.

Gold et al. (12), Noise Texture. Gold et al. (ref. 12, figure 3, p. 177) show improvement in efficiency for the identification of a noise texture. For each of the two participants, AMC and JMG, we fit a straight line, in log coordinates, to the points using linear least-squares regression. The average log-log slope of contrast learning was −0.32.

Gold et al. (12), Face. Gold et al. (ref. 12, figure 3, p. 177) show improvement in efficiency for the identification of a face.

For each of the two participants, AMC and CGB, we fit a straight line, in log coordinates, to the points using linear least-squares regression. The average log-log slope of contrast learning was −0.39.

Michel and Jacobs (13). Michel and Jacobs (ref. 13, figure 6, p. 9) show improvement in efficiency for discrimination of shapes in filtered noise. The authors defined efficiency as the ratio of the sensitivity index d′ of human and ideal, which is similar, but not identical, to our definition as the ratio of threshold energies. For each of the three participants who showed evidence of learning (BVR, WHS, and RAW), we fit a straight line, in log coordinates, to the points using linear least-squares regression. The average log-log slope of contrast learning was −0.48.

Fine and Jacobs (14). Fine and Jacobs (ref. 14, figure 6, p. 3217) show improvement in threshold contrast for the discrimination of a complex plaid pattern. The high spatial frequency component of the plaid was placed at a different contrast than the low spatial frequency component, and so for analysis we separately measured the slope using the contrast of each component, and then averaged the slopes together to produce the final estimate. The across-participant average threshold contrast for sessions 1 and 2 (coded as session 1.5) was 0.081 and 0.27 for the low and high spatial frequency components, respectively. After the final sessions, 7 and 8 (coded at session 7.5), thresholds dropped to 0.024 and 0.078, respectively. We fit a straight line, in log coordinates, to the points using linear least-squares regression. The average log-log slope was −0.78.

1. Watson AB, Robson JG (1981) Discrimination at threshold: Labelled detectors in human vision. *Vision Res* 21(7):1115–1122.
2. Kim J, Wilson HR (1993) Dependence of plaid motion coherence on component grating directions. *Vision Res* 33(17):2479–2489.
3. Kim J, Wilson HR (1993) Dependence of plaid motion coherence on component grating directions. *Vision Res* 33(17):2479–2489.
4. Brainard DH (1997) The Psychophysics Toolbox. *Spat Vis* 10(4):433–436.
5. Pelli DG, Zhang L (1991) Accurate control of contrast on microcomputer displays. *Vision Res* 31(7-8):1337–1350.
6. Watson AB, Pelli DG (1983) QUEST: A Bayesian adaptive psychometric method. *Percept Psychophys* 33(2):113–120.
7. Pelli DG, Burns CW, Farell B, Moore-Page DC (2006) Feature detection and letter identification. *Vision Res* 46(28):4646–4674.
8. Pelli DG, Burns CW, Farell B, Moore-Page DC (2006) Feature detection and letter identification. *Vision Res* 46(28):4646–4674.
9. Furmanski CS, Schluppeck D, Engel SA (2004) Learning strengthens the response of primary visual cortex to simple patterns. *Curr Biol* 14(7):573–578.
10. Suchow JW, Pelli DG (2005) Learning to identify letters: Generalization in high-level perceptual learning. *J Vis* 5(8):712, (abstr).
11. Lu ZL, Dosher BA (2004) Perceptual learning retunes the perceptual template in foveal orientation identification. *J Vis* 4(1):44–56.
12. Gold J, Bennett PJ, Sekuler AB (1999) Signal but not noise changes with perceptual learning. *Nature* 402(6758):176–178.
13. Michel MM, Jacobs RA (2008) Learning optimal integration of arbitrary features in a perceptual discrimination task. *J Vis* 8(2):, 3.1–16.
14. Fine I, Jacobs RA (2000) Perceptual learning for a pattern discrimination task. *Vision Res* 40(23):3209–3230.

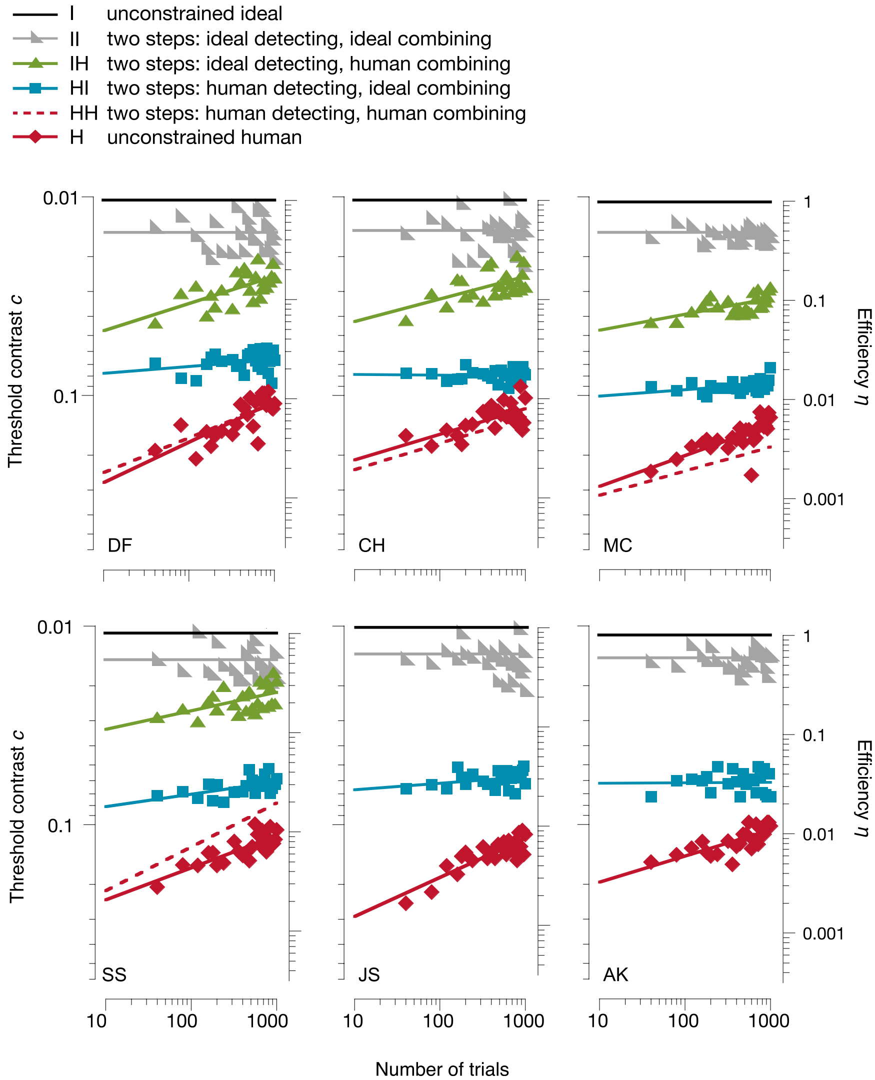

Fig. S1. Learning curves for all participants (as in Fig. 2). The log-log slope of contrast learning, averaged across participants, is 0 (I), 0.019 (II), −0.11 (IH), −0.027 (HI), −0.16 (HH), and −0.16 (H). The threshold contrast at 1,000 trials, averaged across participants, is 0.010 (I), 0.015 (II), 0.026 (IH), 0.069 (HI), 0.13 (HH), and 0.12 (H).