The bottlenecks in human letter recognition: A computational model

NEW YORK UNIVERSITY

Avi J. Ziskind[1], Olivier Hénaff[2], Yann LeCun[2,3], & Denis G. Pelli[1,2]

[1]Psychology Dept., NYU [2]Center for Neural Science, NYU [3]Computer Science Dept.

Abstract

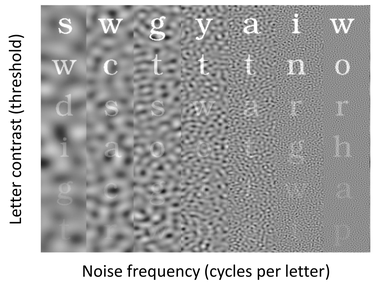
We have implemented two machine-learning models of object recognition by human observers. Both models capture three hallmarks of human performance that cannot be accounted for by template matching:

(1) spatial frequency channels,
(2) crowding,
(3) effects of letter complexity.

One model is a Convolutional Neural Network (ConvNet), and the other is a texture statistics model followed by a linear classifier. With appropriate hyper-parameters and training, both models account for spatial-frequency channels, crowding, and effects of letter complexity.

PHENOMENON

Channel tuning

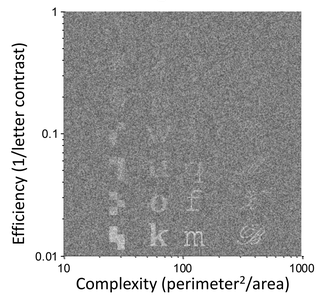


Demo: The outline of barely visible letters traces out your threshold for seeing letters as a function of spatial frequency.

Crowding

Demo: Each target letter is presented between two fixed-contrast flankers. The flankers are tightly spaced on the left and loosely spaced on the right. While fixating the top plus sign, identify the target letter on the left, and on the right. Now try the next line, and so on. The faintest target letter that you identify indicates your threshold contrast for identification in the presence of flankers. Nearby flankers raise threshold much more than distant flankers do.

Complexity

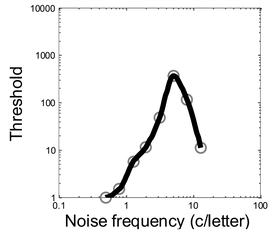


Demo: In each column, the height of the faintest identifiable letter indicates your efficiency. Across columns, your efficiency drops as complexity increases.

OBSERVER

Human observer

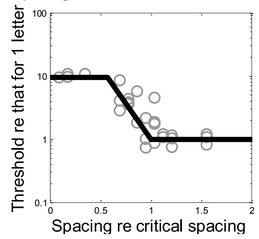
Estimated human threshold for identifying a letter in narrow band noise as a function of center frequency of the noise. Note the bandpass character of the curve. This is called a "channel".



Here, we specify font size by its k-height. Solomon and Pelli used x-height, which is smaller by a factor of 1.5.

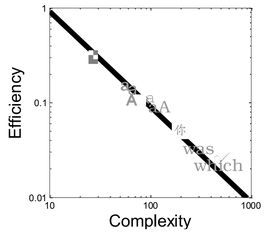
Solomon & Pelli, 1994

Threshold energy required by an observer to identify a target letter as a function of the center-to-center spacing to a flanker. The eccentricity was 20°, and the critical spacing was 11°.



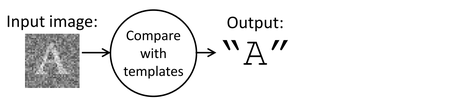
Pelli et al., 2004

Efficiency is the threshold energy required by an observer for identification relative to that of the optimal classifier. The log-log plot shows the reciprocal relationship between efficiency and complexity of font over a 30-fold range.

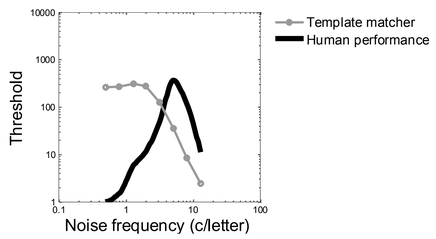


Pelli et al., 2006

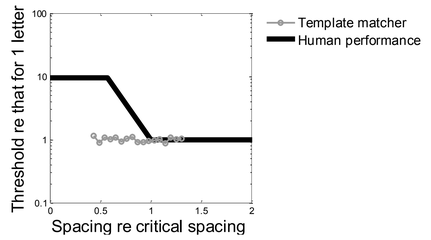
Template matcher ✗



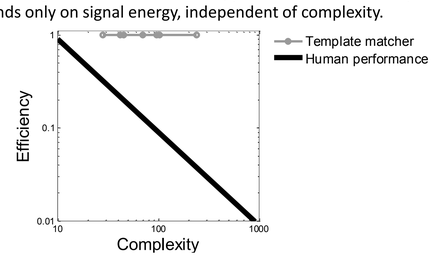
Unlike human, the channel of the template matcher is low pass.



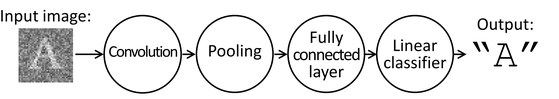
Unlike human, the template matcher is unaffected by the presence of additional letters.



Unlike human, for a given level of white noise, performance of the template matcher depends only on signal energy, independent of complexity.

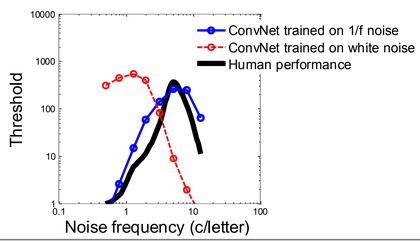


Convolutional network (ConvNet) ✓

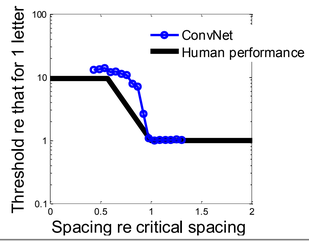


LeCun et al., 1998

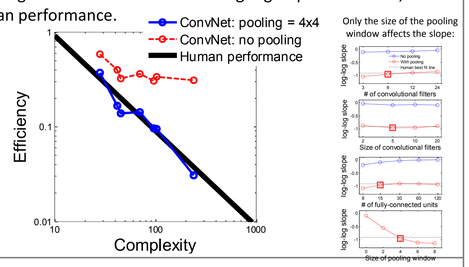
ConvNet is low-pass when trained on letters in white noise (red), and bandpass when trained in 1/*f* noise (blue), which is prominent in human vision (*Raghavan and Pelli*, in prep.).



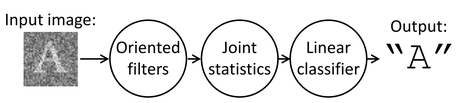
We trained a ConvNet on a single letter and then tested its ability to identify either of two letters when presented simultaneously. It performs well only when the second letter falls outside its field of view.



ConvNet's efficiency is inversely proportional to complexity. The critical parameter that controls this dependency is the extent of the pooling window. A pooling window of 4x4 results in a log-log slope of about -1, matching human performance.

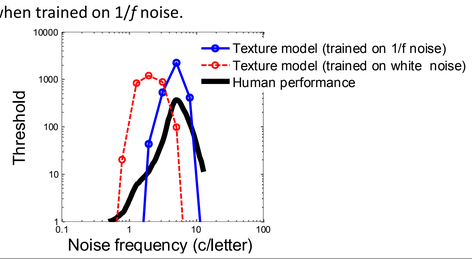


Texture statistics ✓

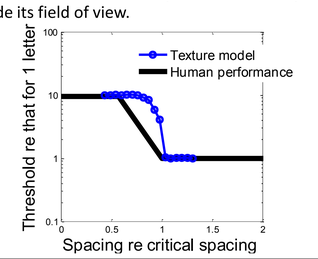


The linear classifier classifies images by their Portilla & Simoncelli (2000) texture statistics.

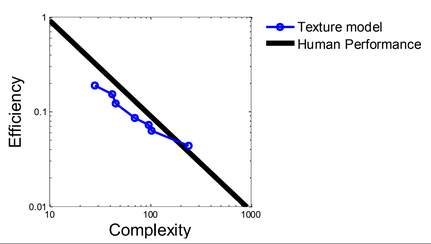
The threshold curve of the texture statistics model is band-pass like that of human observers. The center of the channel matches that of human observers when trained on 1/*f* noise.



The texture statistics model can identity a single letter, but performs very poorly when two letters are present. It performs well only when the second letter falls outside its field of view.



The texture statistics model's efficiency is inversely proportional to complexity, with a log-log slope (-0.7) close to that of human observers (-1).



Bibliography

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.
Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision research*, *46*(28), 4646-4674.
Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *Int'l J of Computer Vision*, *40*(1), 49-70.
Raghavan, M, & Pelli, D.G. Photon and cortical noises limit what we see. (*in preparation*)
Solomon, J. A., & Pelli, D. G. (1994). The visual filter mediating letter identification. *Nature*, *369*(6479), 395-397.

Citation

Ziskind, A.J., Hénaff, O., LeCun, Y., & Pelli, D.G. (2014) The bottleneck in human letter recognition: A computational model. *Vision Sciences Society*, St Pete Beach, FL, May 16-21, 2014, 56.583. http://psych.nyu.edu/pelli/posters.html

Acknowledgments

We thank Eero Simoncelli, Jonathan Winawer, Najib Majaj, Johannes Ballé, Elad Ganmor and Neil Rabinowitz for very helpful discussions.