Towards an image-computable visual text quality metric using deep neural networks

Ling-Qi Zhang[1, 2], Minjung Kim[1], James Hillis[1], Trisha Lian[1]
[1]Meta Reality Labs, [2]University of Pennsylvania

Abstract

Image quality metrics have become invaluable tools for image processing and display system development. These metrics are typically developed for and tested on images and videos of natural content. Text, on the other hand, has unique features and supports a distinct visual function: Reading. It is therefore not clear if these image quality metrics are effective or optimal as measures of text quality. Here, we developed a domain-specific image quality metric for text and compared its performance against quality metrics developed for natural images. To develop our metric, we first trained a deep neural network to perform text classification on a data set of distorted letter images. We then compute the responses of internal layers of the network to uncorrupted and corrupted images of text, respectively. We used the cosine dissimilarity between the responses as a measure of text quality. Preliminary results indicate that both our model and more established quality metrics (e.g., SSIM) are able to predict general trends in participants' text quality ratings. In some cases, our model is able to outperform SSIM. We further developed our model to predict response data in a two-alternative forced choice (2AFC) experiment, on which only our model achieved very high accuracy. Lastly, we demonstrated our model has the potential to generalize to novel perceptual dimensions that it has not been explicitly trained on.

Introduction

The ability to show high quality text is important for any display system, especially for productivity purposes. There have been many studies conducted on the legibility of text, usually through the perspective of reading speed [1, 2, 3]. However, even with equally legible text, numerous factors can still impact the visual *quality* of text. Here we are defining quality as the overall aesthetic appearance of the text in addition to if the text is legible , which can be crucial for the user experience. While text quality can be assessed through user studies, they can be costly, time consuming, and limited to a few dimensions of interest. The goal of the current paper is to develop an image-computable metric that is able to predict text quality directly from rendered images of text. Such model will provide a valuable tool for the design of any display system for which text quality is an important consideration.

Related work

To the best of our knowledge, there is no general model of text quality. However, many metrics has been developed for quality assessment of images of natural content. Earlier attempts at developing image quality metrics used heuristic and empirical measurements of human perception to design algorithms that transform the image to the appropriate perceptual space [4, 5, 17]. An alternative approach was to use the fact that natural images have distinct statistical regularities [7]. Thus, distortions on natural images can be assessed through characterizing the deviations from these regularities. In addition, since these statistics can often be computed on a single image, a high-quality "reference image" is no longer required. For example, in [8], the wavelet domain coefficients are used for assessing and removing noise from images while [9] developed a set of statistical metrics for measuring the "naturalness" of images in the spatial domain. These two different approaches also roughly correspond the full-reference and reference-free methods of quality assessment. Empirical measurements of the human visual system have also played important role in advancing image quality metric, for example, by incorporating the contrast sensitivity function [10, 11].

More recently, deep convolutional neural networks and learning-based methods have achieved superior performance on various image processing tasks [12]. Thus, they have also been used for building image quality metrics. In [13], the authors showed that the internal representation of neural networks trained on natural images is highly effective in predicting quality judgements of human participants. Related studies also demonstrated it it possible to use user data (e.g., 2-alternative-forced choice, or 2AFC) to either fine-tune [13], or directly train a neural network model for image quality assessment [14, 15]. In particular relevance to us, previous work also demonstrated that deep neural network is able to capture the statistical regularities of text for task such as deblurring [16].

In our current work, we developed a visual quality metric based on a convolutional neural network that is specialized for text. We demonstrate that our model outperforms generic image quality metric for predicting text quality, and also fine-tune our model using 2AFC data of text quality judgement.

Convolutional neural network model

The success of neural network models depends on their ability to extract relevant perceptual features from images. Since text has unique features that are distinct from natural images, our first step is to develop a neural network that is specialized for text. To this end, we trained a simple convolutional neural network to classify images of letters.

The neural network consists of six consecutive 2D convolutions with kernel size 3. Each convolution is followed by a rectified linear unit (ReLU), a 2D batch normalization, and a maximum pooling operation. The output of the last convolution block then goes to three fully connected layers, before being converted into class probability through softmax. A dropout layer with a drop probability of 0.2 is applied right before the first layer, and a 1D batch normalization is applied right after it. The network's

task is simply identify the letter. We also assumed that the upper and lower case letters as the same category.

For the training data, we generated an image dataset of isolated, 26 English letters in either upper or lowercase, that varies in location, size, and contrast. The letter images then go through a random affine transformation to simulate geometric artifacts, and are blurred by a Gaussian point spread function with varying sizes, before corrupted by randomly sampled 1/f pink noise of different magnitudes (**Figure 1**). The network is optimized through the Adam optimizer using the cross-entropy loss function. We used a mini-batch size 64, an initial learning rate (LR) of 0.001 with an exponential LR decay with rate 0.95 per epoch. Although the performance is not our focus per se, the network can achieve a high accuracy of over 95% on the classification task.

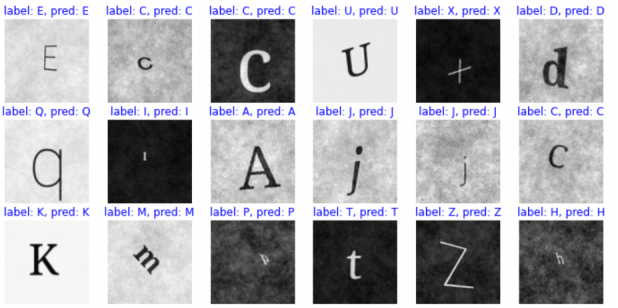

Figure 1. *Some example images from our test dataset. These are examples where the images have a low background noise. The network is able to correctly classify all the letters after training.*

Full-reference text quality assessment

After the network is optimized for the text classification task, we use the internal features extracted by the convolutional part of the model to build our text quality metric. Concretely, for model $f(\cdot)$, we have the features r computed from image I, $r = f(I)$. Given a reference image I_r and distorted image I_d, we can compute $r_r = f(I_r)$ and $r_d = f(I_d)$, respectively. The quality score s is then computed based on the difference between r_r and r_d. We used cosine similarity, $\cos(r_1, r_2) = (r_1 \cdot r_2)/(||r_1|| * ||r_2||)$. The final form of our text quality metric is $D(I_r, I_d) = \cos(f(I_r), f(I_d))$.

We compute the quality scores predicted by the model and compare them to an internal study where users were asked to provide a rating between 1 and 6 given an image of text. The images were rendered with either a different Gaussian blur size, font size, or text sharpening method. By running our model using the highest quality image as the reference, our method is able to qualitatively predict the user rating data. We also compared the results with standard image quality metric (i.e., SSIM). Although SSIM is also able to predict the user ratings for the blur and text sharpening condition, it failed to predict the effect of font size, presumably due to the lack of (size) invariance.

Building reference-free model with 2AFC data

In the previous section, we have shown that the internal features of the network capture key aspects of text quality. However, with the cosine similarity metric, a reference image is always required, and the numerical value of the score is not directly interpretable. Thus, we built an explicit score function using user data from a two-alternative forced choice (2AFC) task.

In particular, we assume a linear relationship between an internal judgement of quality score s produced by each user, and the features of the neural network r, namely $s = w^T r$. Here w is a weight vector. To model 2AFC data, we compute s for both images on each trial, $s_1 = w^T r_1$ and $s_2 = w^T r_2$, and a decision is made through a sigmoid nonlinearity $h(\cdot)$ based on the difference $s_1 - s_2$.

To estimate the weight vector w, we performed a logistic regression with L_1 regularization on a dataset of 2AFC judgements where users made choices regarding quality of text images with different font size, display resolution, and optical blur.

Consider the binary choice made by the user in each trial i, $d_i \in \{0,1\}$. The model predicts the user has a probability of $p_i = h(w^T r_{i1} - w^T r_{i2})$ in choosing image number 1 as the one with better text quality. Logistic regression minimizes the cross-entropy loss function: $L = -\sum_{i=1}^{N} \{d_i \log p_i + (1-d_i)\log(1-p_i)\}$. The L_1 regularization encourages the weight vector to be "sparse", result in the final form of the loss function: $L = -\sum_{i=1}^{N} \{d_i \log p_i + (1-d_i)\log(1-p_i)\} + \beta \sum_i |w_i|$. The β parameter was chosen using a cross-validation procedure.

There are a total of $4,000$ binary choices across 10 subjects. We randomly selected $3,600$ trials for model fitting and cross-validation, and report the model testing performance on the remaining 400 trials (**Figure 2**). Our model is able to achieve over 90% accuracy on the test set. As a comparison, we also built a control model that uses the difference in pixel values of the two images as regressors, and perform logistic regression with the choice data, which only achieved an accuracy of 60%. Further, we found that SSIM is unable predict the binary choices above chance level.

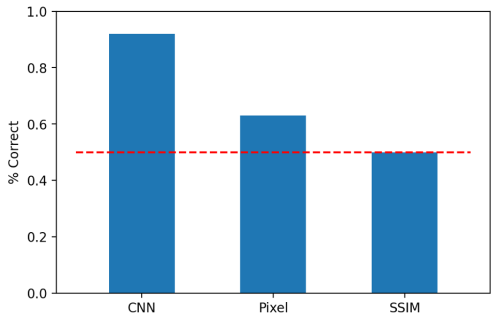

Figure 2. *Performance, in percent correct, in predicting the binary choices on the test dataset, after fitting the linear score function on the training dataset. The dotted line indicate chance level performance (50%).*

With the weight vector w estimated from 2AFC data, we now have a completely reference-free image quality metric of the form $s = w^T \cdot f(I)$. In addition, since the quality score is adjusted through choice data, it can be interpreted in the unit of just noticeable difference (JND). That is, a difference of 1 in quality score predicts that an average user can barely see the quality differences between the text images.

Factors impacting text quality

To visualize the prediction of the model on how different factors impact text quality, we plot the mode-predicted quality score

s as a function of these factors. Below, we show two examples of a contour plot of the predicted score, as two different factors that impact text quality are varied (**Figure 3**).

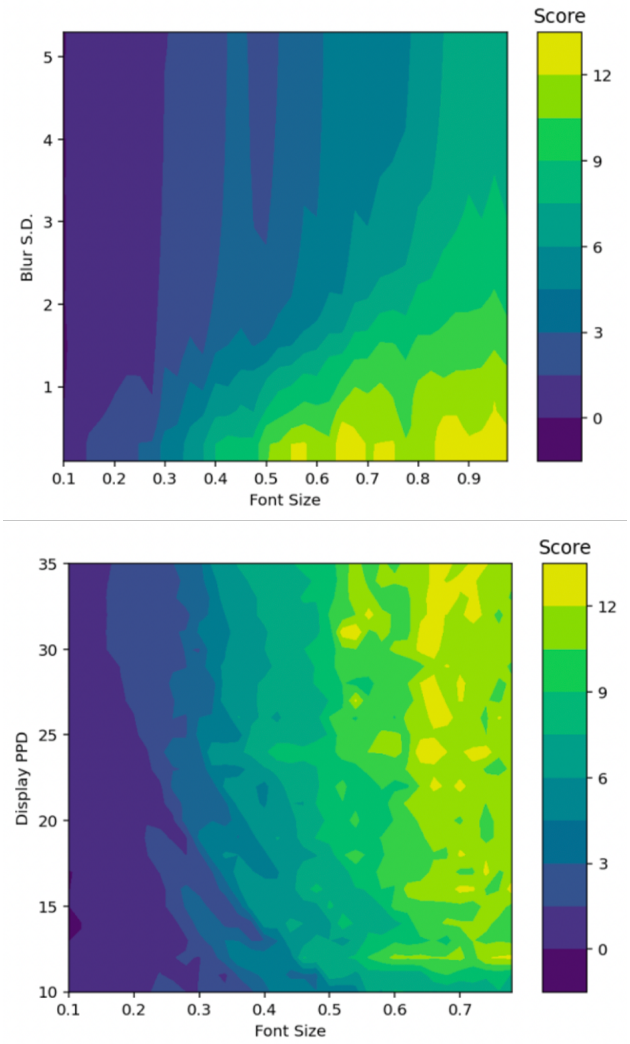

Figure 3. *The quality score, predicted by the model, as a function of the font size (x-axis) and optical blur (y-axis) of the rendered text image (top); and the font size (x-axis) and image resolution (y-axis) of the rendered text image (bottom). The score is averaged over a randomly selected set of sentences.*

As expected, as the font size increases and the optical blur decreases, the quality of the text is higher (**Figure 3**, top). Similarly, larger font size and image resolution results in better text quality (**Figure 3**, bottom). Most importantly, the lines in the contour plots represent iso-quality lines: The combinations of the parameters along them are predicted to produce text with the same perceptual quality. This is important for understanding trade-offs that are relevant for text quality in display design. In addition, although our model was optimized using only binary choices, it produces quality score predictions that are mostly continuous and smoothly varying, indicting the model is properly extrapolating from the training data. However, we can still see some discontinuities, such as in bright yellow regions in the bottom plot of **Figure 3**. We expect that these should be resolved with a larger dataset.

Predicting text quality

Our model would be extremely valuable if it is able to predict text quality beyond the dimensions that has been explicitly trained on. To test this possibility, we next examine the effect of additive contrast on text quality. Below (**Figure 4**) we show two example images where a relatively low (left) and high (right) additive contrast, respectively.

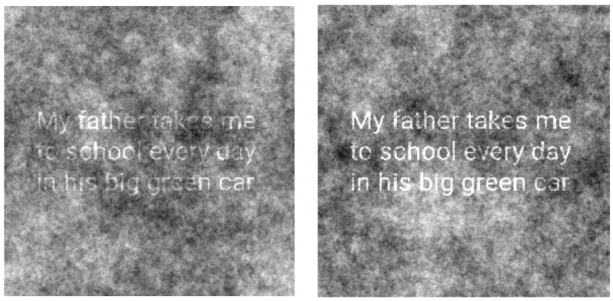

Figure 4. *Two example images of text with a low additive contrast (left) and high additive contrast (right), respectively.*

Previous research has examined the relationship between the additive contrast of the display, and the legibility and quality of the rendered text [17]. Although text legibility saturates when the contrast is high enough, that is, after a certain contrast threshold, subjects were able to recognize text equally well; Text quality increases monotonically (See **Figure 3** in [17]). Thus, subjects prefer higher contrast text, even if it does not make the text more legible. Our model was never trained explicitly with contrast variation, both for the initial letter recognition task, and the 2AFC prediction task. We computed the predicted quality score for a series of text images with increasing additive contrast (**Figure 5**). Our model is able to recapitulate the exact monotonic relationship observed in [17], which demonstrates its ability to generalize to novel dimensions.

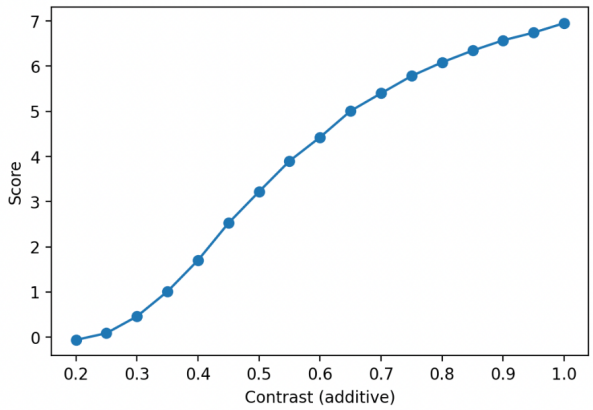

Figure 5. *The quality score predicted by our model, as we increase the additive contrast in the rendered text images. The score is averaged over a randomly selected set of sentences.*

To further elucidate how well our model is able to general-

ize, we further predicted the 2D contour plots, similar to **Figure 6**, but using additive contrast as one of the axes. Although the model was never explicitly trained on how contrast, font size, and optical blur jointly determines text quality, it produces quality score predictions that are sensible. For example, in the top plot of **Figure 6**, the model predicts that a higher contrast and larger font size produces higher quality text, and the two factors also trade-off in a precise way, as demonstrated by the iso-quality contour. Similar results are also observed in the bottom panel, where we simulated the interaction between contrast and optical blur. These predictions can be verified using 2AFC experiment in the future.

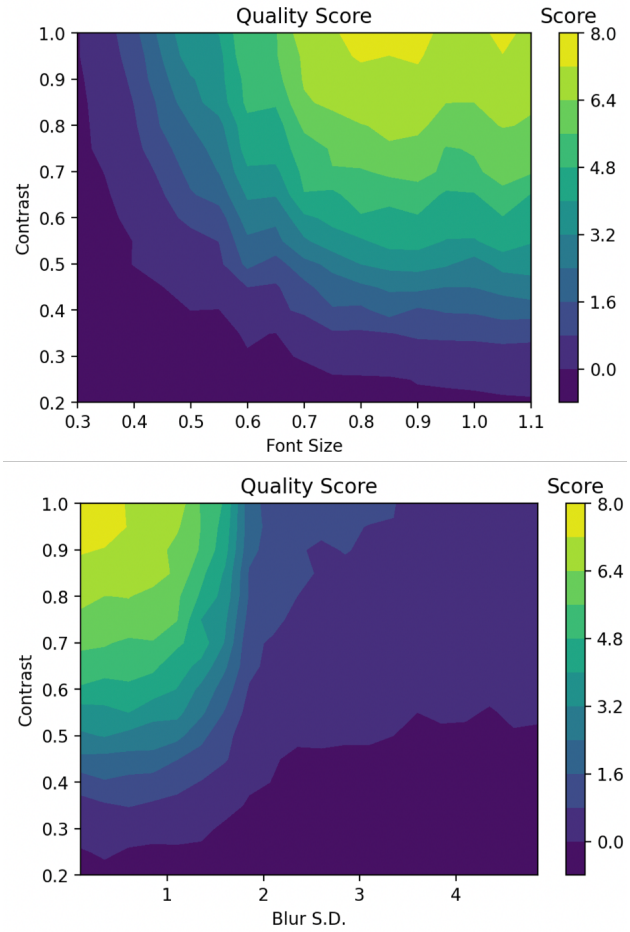

Figure 6. *The quality score, predicted by the model, as a function of the font size (x-axis) and additive contrast (y-axis) of the rendered text image (top); and the optical blur (x-axis) and additive contrast (y-axis) of the rendered text image (bottom). The score is averaged over a randomly selected set of sentences.*

Discussion

In this article, we built an image-computable text quality metric, based on the internal features of a convolutional neural network trained on simple letter recognition task. Our initial comparison of the model prediction to rating data on text quality indicated the model is able to recapitulate key aspects of user's judgements. Next, we extended the model quantitatively by fitting the model to a dataset of 2AFC judgements by fitting a linear score function using logistic regression with a sparse regularization. Not only can our model achieve a high performance on test data, and can predict text quality score in a meaningful way, including the interaction and trade-off of different factors that impact text quality.

We further showed that our model can generalize to a novel dimension that it was not explicitly trained on. In particular, we found that although the model was not explicitly trained with variations in the additive contrast of the text, it predicts the relationship between contrast and text quality as previous reported [17], and also how contrast interacts with other factors. Future experiments will be able to validate the prediction of the model.

Our work adds to a growing literature of showing the similarities between convolutional neural network and human perception [13, 18, 19], but in the specialized domain of text. In addition to the fact that our simple model is able to achieve high performance in predicting the quality judgements of human participates, we also found that only a small number of features (i.e., $10-20$) is enough for making these predictions **Figure 7**. In contrast, neural network models that are not trained specially on text require a much larger set of features to be able to predict behavior **Figure 7**. This indicates that our model trained specifically on text does extract the most relevant perceptual features, compare to models that are trained on natural images and others. This also offers us a potential opportunity to interpret these features that are predictive of text quality judgements, which will be one of our most important goals in future work.

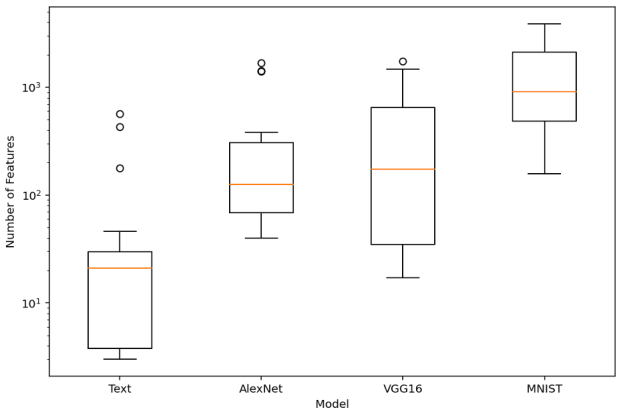

Figure 7. *The number of features required for obtaining the same predictive performance of behavior (y-axis), for four different neural network models (x-axis) that are trained on text (our model), object recognition (AlexNet and VGG), and MNIST dataset.*

It is worth noting, however, that our model is still limited to the user data it was trained on. Although we have showed the model is able to generalize to novel dimension, we expect that it still will need to be trained on a more comprehensive dataset that includes a wide range of conditions [20]. In addition, our definition of text quality is confined mostly to the task of reading. It could be possible that with alternative task demand, for example, visual search [21], factors that impact text quality will be different. However, our modeling framework should also be applicable, given the appropriate dataset for model training.

In summary, we have shown that we are able to build spe-

cialized model of text quality metric that achieves good predictive performance of user judgement of text quality. Our model is able to capture how different factors that impact text quality interact and trade-off, and also generalize to novel dimensions. The prediction of our model should also be testable by simply running psychophysics experiment, for example, to see the effect of geometric distortions on text quality, just as those we used in our training dataset. Our model will provide a valuable tool for the design of display and text rendering pipeline.

References

[1] Legge GE, Pelli DG, Rubin GS, Schleske MM. Psychophysics of reading—I. Normal vision. Vision research. 1985 Jan 1;25(2):239-252.

[2] Legge GE, Bigelow CA. Does print size matter for reading? A review of findings from vision science and typography. Journal of vision. 2011 May 1;11(5):8-8.

[3] Chung ST, Legge GE, Pelli DG, Yu C. Visual factors in reading. Vision research. 2019 Aug;161:60-62.

[4] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing. 2004 Apr 13;13(4):600-12.

[5] Zhang L, Zhang L, Mou X, Zhang D. FSIM: A feature similarity index for image quality assessment. IEEE transactions on Image Processing. 2011 Jan 31;20(8):2378-2386.

[6] Zhang X, Silverstein DA, Farrell JE, Wandell BA. Color image quality metric S-CIELAB and its application on halftone texture visibility. InProceedings IEEE COMPCON 97. Digest of Papers 1997 Feb 23 (pp. 44-48). IEEE.

[7] Srivastava A, Lee AB, Simoncelli EP, Zhu SC. On advances in statistical modeling of natural images. Journal of mathematical imaging and vision. 2003 Jan;18(1):17-33.

[8] Portilla J, Strela V, Wainwright MJ, Simoncelli EP. Image denoising using scale mixtures of Gaussians in the wavelet domain. IEEE Transactions on Image processing. 2003 Oct 27;12(11):1338-1351.

[9] Mittal A, Moorthy AK, Bovik AC. No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing. 2012 Aug 17;21(12):4695-4708.

[10] Mantiuk R, Kim KJ, Rempel AG, Heidrich W. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Transactions on graphics (TOG). 2011 Jul 25;30(4):1-4.

[11] Mantiuk RK, Denes G, Chapiro A, Kaplanyan A, Rufo G, Bachy R, Lian T, Patney A. Fovvideovdp: A visible difference predictor for wide field-of-view video. ACM Transactions on Graphics (TOG). 2021 Jul 19;40(4):1-9.

[12] LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015 May;521(7553):436-444.

[13] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 586-595).

[14] Prashnani E, Cai H, Mostofi Y, Sen P. Pieapp: Perceptual image-error assessment through pairwise preference. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018 (pp. 1808-1817).

[15] Kim J, Lee S. Deep learning of human visual sensitivity in image quality assessment framework. InProceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 1676-1684).

[16] Mei J, Wu Z, Chen X, Qiao Y, Ding H, Jiang X. Deepdeblur: text image recovery from blur to sharp. Multimedia tools and applications. 2019 Jul;78(13):18869-85.

[17] Blanc-Goldhammer DR, MacKenzie KJ. The effects of natural scene statistics on text readability in additive displays. InProceedings of the Human Factors and Ergonomics Society Annual Meeting 2018 Sep (Vol. 62, No. 1, pp. 1281-1285).

[18] Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the national academy of sciences. 2014 Jun 10;111(23):8619-24.

[19] Yamins DL, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience. 2016 Mar;19(3):356-65.

[20] Hebart MN, Zheng CY, Pereira F, Baker CI. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nature human behaviour. 2020 Nov;4(11):1173-85.

[21] Wolfe JM. What can 1 million trials tell us about visual search?. Psychological Science. 1998 Jan;9(1):33-9.

Author Biography

Ling-Qi Zhang is a PhD student in Neuroscience at the University of Pennsylvania. His research focuses on how regularities in the visual world shape sensory perception using computational models and psychophysics experiment.

Minjung Kim is a research scientist at Reality Labs. She works on bridging basic vision science with applied research, with an emphasis on developing models of contrast and color perception to predict text legibility in head-mounted displays. `https://www.minjung.ca/`

James Hillis is a research scientist at Reality Labs. His work applies knowledge from perception science to optimize design and development of display hardware, control systems and graphics pipelines for a broad range of products.

Trisha Lian is a research scientist at Reality Labs. Her research focuses on developing imaging system simulations alongside models of the human visual system, in order to better understand visual artifacts in head mounted displays.