To appear in: Proc. SPIE, Conf. on Human Vision and Electronic Imaging, XX
Vol. 9394, Eds. B Rogowitz, TN Pappas, H de Ridder,
San Francisco, CA, 9-12 Feb 2015.

Geometrical and Statistical Properties of Vision Models obtained via Maximum Differentiation

Jesús Malo* and Eero P. Simoncelli†
jesus.malo@uv.es, http://isp.uv.es

ABSTRACT

We examine properties of perceptual image distortion models, computed as the mean squared error in the response of a 2-stage cascaded image transformation. Each stage in the cascade is composed of a linear transformation, followed by a local nonlinear normalization operation. We consider two such models. For the first, the structure of the linear transformations is chosen according to perceptual criteria: a center-surround filter that extracts local contrast, and a filter designed to select visually relevant contrast according to the Standard Spatial Observer. For the second, the linear transformations are chosen based on statistical criterion, so as to eliminate correlations estimated from responses to a set of natural images. For both models, the parameters that govern the scale of the linear filters and the properties of the nonlinear normalization operation, are chosen to achieve minimal/maximal subjective discriminability of pairs of images that have been optimized to minimize/maximize the model, respectively (we refer to this as MAximum Differentiation, or "MAD", Optimization). We find that both representations substantially reduce redundancy (mutual information), with a larger reduction occurring in the second (statistically optimized) model. We also find that both models are highly correlated with subjective scores from the TID2008 database, with slightly better performance seen in the first (perceptually chosen) model. Finally, we use a foveated version of the perceptual model to synthesize visual metamers. Specifically, we generate an example of a distorted image that is optimized so as to minimize the perceptual error over receptive fields that scale with eccentricity, demonstrating that the errors are barely visible despite a substantial MSE relative to the original image.

Keywords: Vision Models, Multi-layer Networks, Image Quality Metrics, Maximum Differentiation, Redundancy Reduction, Visual Metamers.‡

Introduction

It is widely believed that visual perception emerges through a cascaded sequence of neural transformations having a similar "canonical" form.[1–4] It is thus natural to construct models for image quality assessment using cascades, in which each stage analyzes factors of increasing complexity (e.g. luminance, contrast, and structure).[5–7] Critical to this endeavor is the problem of parameter optimization: How can we select parameters of a multi-stage cascade model so as to mimic human discrimination performance? Recent progress in object recognition has exploited such cascaded constructions, along with new methods of supervised machine learning applied to large databases of natural images, to achieve state-of-the-art results [e.g., 8–10]. But these

*Image Processing Lab, Universitat de València, Spain

†Howard Hughes Medical Institute, Center for Neural Science, and Courant Institute of Mathematical Sciences, New York University, USA

‡This work was partially supported by the projects TIN2012-38102-C03-01 and TIN2013-50520-exp.

machine learning methods rely on data sets (images, and category labels) that are many orders of magnitude larger than what can be feasibly obtained in experiments with human subjects.

Here, we assume a cascade model in which each stage is a linear-nonlinear transformation. The linear part is a convolution with one or more filters, and the nonlinear part implements a local "divisive normalization" operation (see[11] for review). Such representations have been proposed previously for perceptual image distortion measures,[5,7,12,13] as well as other image processing applications.[14] We use perceptually or statistically motivated choices of the filters for each stage.[7,13] Given these filters, we determine the remaining scalar parameters of the normalization operations based on human subject responses in a Maximum Differentiation (MAD) task.[15] Specifically, subjects examine pairs of images synthesized with maximal/minimal distortion according to models with different parameter settings, and choose the image pairs that are most easily differentiated. We refer to this as "MAD optimization".

Preliminary results lead to models in which each transformation step provides a successively more powerful measure of perceptual quality, as measured by its ability to synthesize images with minimally visible distortion but with substantial mean squared error. Moreover, each step reduces the statistical redundancy (mutual information) of natural images. Both models are highly correlated with human quality scores from the TID2008 database.[16] Finally, we show that a "foveated" version of the perceptually-optimized model, in which filters grow in diameter with eccentricity (as in the early visual system), can be used to synthesize images with minimally visible distortions of even larger mean squared error.

Structure of cascaded models

We assume a model constructed as a cascade of stages:

$$\mathbf{x}^{(0)} \xrightarrow{T_1} \mathbf{x}^{(1)} \xrightarrow{T_2} \mathbf{x}^{(2)} \tag{1}$$

where $x^{(0)}$ is a pixellated input (luminance) image containing d pixels, each T_i performs a *linear+nonlinear* transformation on the vector $\mathbf{x}^{(i)}$ to give the set of responses in the vector $\mathbf{x}^{(i+1)}$. The quality measure is simply the MSE measured on the output layer, $\mathbf{x}^{(2)}$.

The high dimensionality of the signals can imply a large number of parameters in each transformation T_i. For instance, a general linear operation preserving the dimensionality in the i-th stage is parameterized by a matrix with $d \times d$ free parameters. The number of parameters in the nonlinear part may be even bigger. In order to keep the number of parameters small, the structure of the stages has to be constrained.

We assume the linear stages are convolutional,[17] and choose the filters according to either perceptual or statistical criterion. The resulting transforms are somewhat similar, as suggested by the Efficient Coding Hypothesis.[13,18,19]

- **Perceptual design** The filters are chosen using well-known conventions in the perception and image quality communities. The first stage computes the local contrast by subtracting and normalizing by the local luminance (computed with a Gaussian kernel of width σ_1):

$$\mathbf{x}^{(1)}_k = \frac{\left(I - H^{(1)}_{kl}\right)\mathbf{x}^{(0)}_l}{b_1 + H^{(1)}_{kl}\,\mathbf{x}^{(0)}_l} \tag{2}$$

where I is the identity matrix. The second stage applies a linear Contrast Sensitivity filter and a masking nonlinearity:

$$\mathbf{x}^{(2)}_k = sign(H^{CSF}_{kl}\,\mathbf{x}^{(1)}_l)\,\frac{|H^{CSF}_{kl}\,\mathbf{x}^{(1)}_l|^{\gamma_2}}{b_2 + K^{(2)}_{kl}\;|H^{CSF}_{lm}\,\mathbf{x}^{(1)}_m|^{\gamma_2}} \tag{3}$$

where H^{CSF} is the convolution by the impulse response corresponding to the Contrast Sensitivity Function of the Standard Spatial Observer CSF.[20] Overall, we just have two parameters in the first stage (σ_1, b_1) and three parameters in the second stage (σ_2, b_2 and γ_2).

- **Statistical design**. We set the linear part, $H^{(i)}$, to be a decorrelating transform (PCA), over a database of natural images.[21] On the other hand, we assume the nonlinear part is a divisive normalization, which has been shown to reduce redundancy:[13, 19, 22, 23]

$$\mathbf{x}_k^{(i)} = D_{kk}^{(i)} \, sign(H_{kl}^{(i)} \, \mathbf{x}_l^{(i-1)}) \, \frac{|H_{kl}^{(i)} \, \mathbf{x}_l^{(i-1)}|^{\gamma_i}}{b_k + K_{kl}^{(i)} \; |H_{lm}^{(i)} \, \mathbf{x}_m^{(i-1)}|^{\gamma_i}} \tag{4}$$

  The normalization pooling, corresponding to $K^{(i)}$, is computed with a Gaussian blur of width σ_i, and $D^{(i)}$ is a diagonal matrix containing a weighting function with an exponential decay in the diagonal, with characteristic length κ_i (i.e., it applies larger weights to the first coefficients of the decomposition defined by $H^{(i)}$). Since $H^{(i)}$ is chosen based on image statistics (it is the eigenvector matrix of the covariance of samples at the $(i-1)$-th stage), there are only four unknowns per stage: the width, σ_i, the scalar parameters b_i and γ_i, and the width κ_i of the post-weighting.

Perceptual optimization of normalization parameters

To obtain perceptual estimates of the scalar parameters of either model, we used the Maximum Differentiation (MAD) competition methodology.[15] Specifically, for a given set of parameters, we synthesize images that are maximal/minimal in the MSE of their model responses relative to those of an original image, but have the same MSE in the image domain (again, relative to the original image). Synthesis is achieved as follows. First we add white noise to the original image to achieve a desired pixel-domain Euclidean distance (MSE). Then, we modify this initial image through gradient ascent/descent on the MSE of the model response, while constraining the image to lie on the sphere of desired image-domain MSE. The gradient is computed by concatenating the Jacobian matrices corresponding to each stage of the transformation (i.e., the "chain rule").

We also implemented a more efficient calculation based on a a second-order approximation of the model response distance.[7, 13] Under this approximation, MAD competition reduces to injecting noise into the subspace corresponding to low/high eigenvalues of the transformed Riemannian metric. This subspace describes the least/most visible directions according to the model. Interestingly, we observed that even for image-domain PSNR $\sim$ 25, this second order approximation of the model response distance,[7, 13] produces very similar results to the full gradient-descent optimization.

According to the model with the selected parameter settings, the two synthesized images are deemed the most perceptually different pair of images at that level of (image-domain) MSE. By comparing image pairs synthesized for different parameter settings, a human observer can select those that are most distinguishable, thereby indicating the choice of parameters that agrees best with their perceptual judgements.

We developed an iterative psychophysical "coordinate descent" procedure to estimate the scalar normalization parameters for both models. In an outer loop, we alternated between optimizing the parameters of the first stage (initially, assuming an unnormalized transformation for the second stage), and then optimizing those of the second stage. Within each stage, we optimized one parameter at a time, holding the others fixed at their current values. MAD optimization of one parameter was achieved by showing observers pairs of extremal images synthesized with six different values of that parameter (but with the same image-domain MSE) and asking them ro specify which pair exhibited the largest difference in quality. After each trial, new pairs were drawn, keeping the selected parameter value from the previous trial, but using six different values of a different parameter. A staircase procedure in which the ranges of the parameters were progressively reduced was used to search for the optimal value of each parameter. This was repeated for multiple initial conditions and observers and the optima averaged to obtain an overall estimate for the parameters. We looped over the set of parameters four times when optimizing each stage, and we iterated over the two stages until stable results (within the standard deviation of the parameters) were obtained. In the explored cases, this only required two or three iterations. Experimental data were collected for two subjects, and original images were drawn from randomly chosen patches of standard images (Barbara, Einstein, Baboon, Cameraman, Boats, Goldhill).

Perceptual properties: Differentiation of cascade stages

Figure 1 demonstrates the perceptual capabilities achieved by each model stage, for both models. Specifically, each row shows sequences of images with identical MSE in the image domain, but with minimal MSE at at each successive stage of the corresponding model cascade. As expected, the visibility of errors in the perceptual model are progressively less noticeable, since the model stages are optimized to represent perceptual distortion. More surprising is that this approximately holds for the statistical model, despite the fact that its linear stages are optimized solely to decorrelate their inputs, as derived from natural images.

Statistical properties: Redundancy reduction

Table 1 demonstrates the reduction of statistical redundancy achieved by successive model stages, for both models. Specifically, we measured mutual information (in bits) between one coefficient and 8 neighbors at different stages of the network. Specifically, we used the eight adjacent coefficients of 3×3 spatial neighborhoods for the perceptual stages, and the eight adjacent frequencies in DCT-like domains arising in the statistical stages. We averaged these values over different coefficients and across a set of 10000 patches of size 64×64 from natural images from a calibrated database.[21] For both models, we see a significant reduction in redundancy. Such a reduction is expected for the statistically optimized model, but consistent with ref.,[19] it is also found in the model that is perceptually derived, suggesting that it is a property of the early visual system.

Model \ Stage	Input	Linear 1	Non-linear 1	Linear 2	Non-linear 2
Perceptual	0.579	0.155	0.152	0.270	0.228
Statistical	0.579	0.032	0.022	0.003	0.002

Table 1. Mutual information (in bits) between groups of neighboring coefficients at intermediate stages of each hierarchical model.

Properties of the distance: Comparison to subjective quality ratings

We compared image quality predictions of our two models to human subject data from the TID2008 database.[16] Figure 2 shows the alignment between subjects' Differential Mean Opinion Scores (DMOS) and distances computed according to our MAD-optimized models (in blue) as

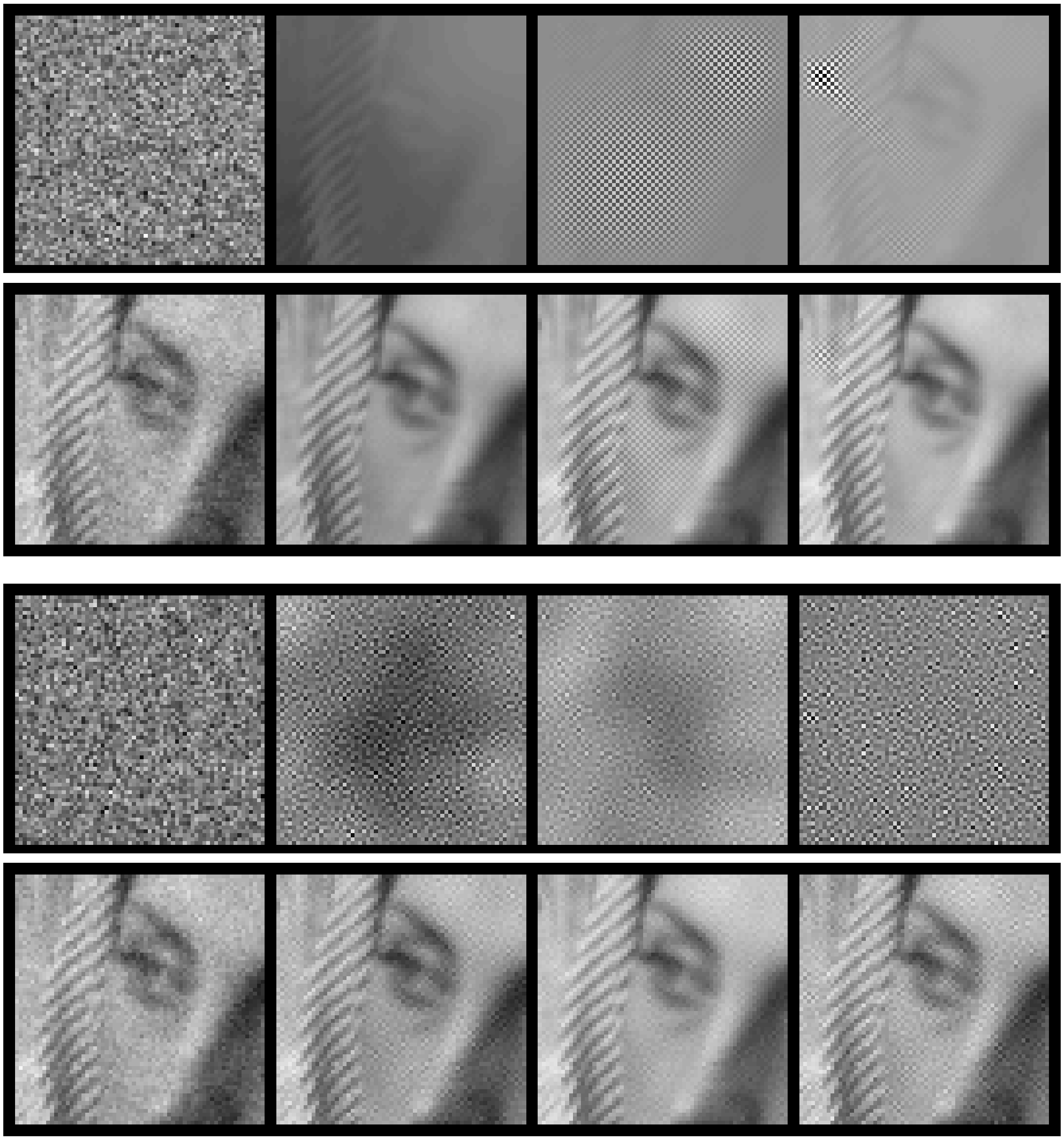

Figure 1. Example images synthesized to have minimal distortion according to successive model stages, but identical MSE in the image domain (PSNR = 28.1). The sampling frequency used in the example implies that images should subtend 1 deg. Left to right: output of first linear transform, output of first normalization transform, output of second linear transform, output of final normalization transform. Top: perceptual model. Bottom: statistical model. The distortion associated with one-layer linear models (left) is similar to white noise because the corresponding metric is strictly the identity in the statistical case and not far from the identity in the perceptual case.

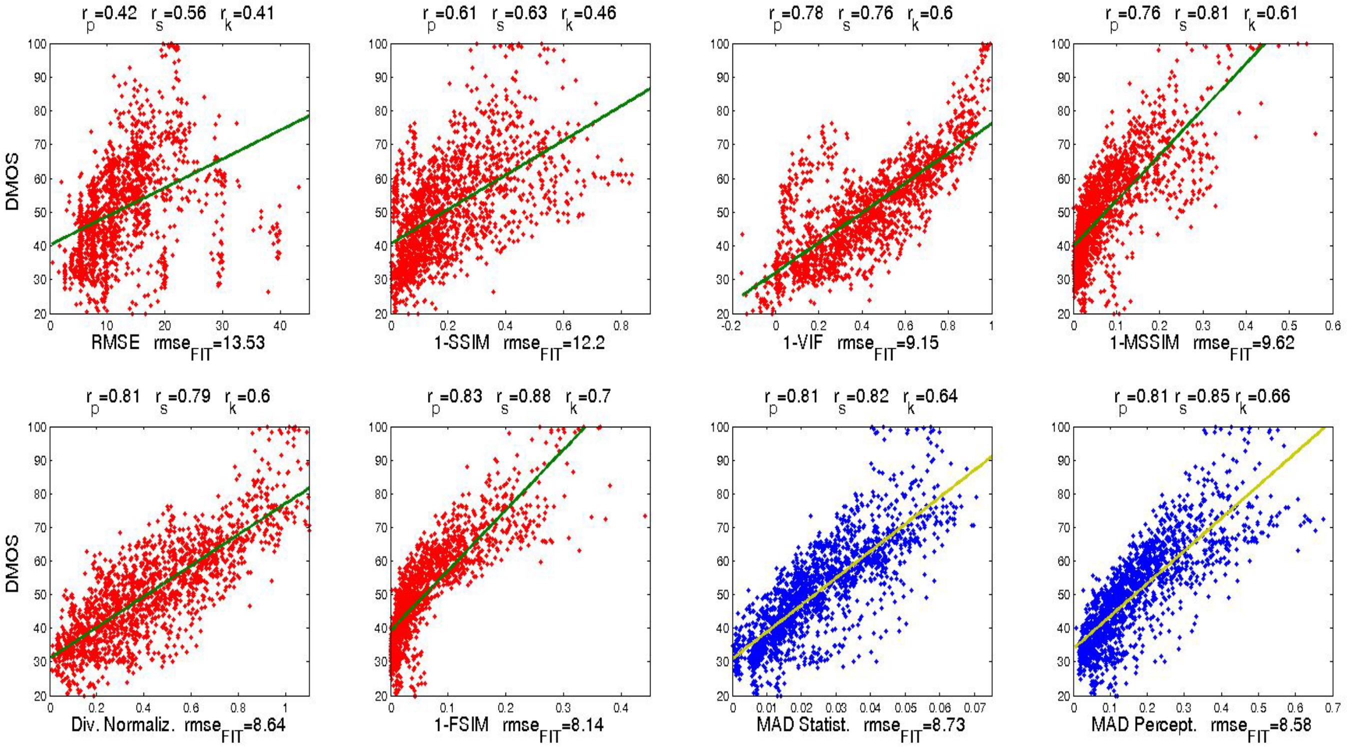

Figure 2. Direct comparison of subjective image distortion data (DMOS) and distances computed according to different distortion measures: our MAD-optimized models (scatter plots in blue) as well as several models from the literature (red - from upper left: RMSE, SSIM,[5] VIF,[25] MSSIM,[24] divNorm,[7] FSIM[26]). The values of the Pearson, Spearman and Kendall correlations (r_p, r_s and r_k, respectively) are provided above each plot. The value of the RMSE of the best linear fit (in DMOS units), from which r_p is computed, is provided underneath each plot.

well as those of several other distortion measures (in red). We find that both models show higher correlation with the human data than SSIM,[5] MSSIM,[24] VIF[25] as well as those based on divisive normalization on wavelets.[7, 12] Moreover, it provides a more linear scatter plot and more balanced residuals than the current state-of-the-art FSIM method.[26]

Perceptual properties: Synthesis of foveated visual metamers

The models proposed here are both based on spatially local computation. In the early visual system, the receptive fields of neurons are localized, but grow approximately linearly with eccentricity away from the fovea. This suggests that a better model might be achieved by scaling the spatial filters with eccentricity. A related approach has shown that a local representation of visual texture, when suitably scaled with eccentricity, can generate highly distorted versions of images that are perceptually indistinguishable from an original, when viewed under proper fixation.[27, 28]

To demonstrate this concept, we modified the perceptual model by introducing an eccentricity dependency in the widths of all the convolutional kernels in Eqs. 2-3 while keeping the other parameters constant. In particular, for eccentricity smaller than 1 degree, the impulse response widths were those found in the MAD experiment. For bigger eccentricities, widths were linearly

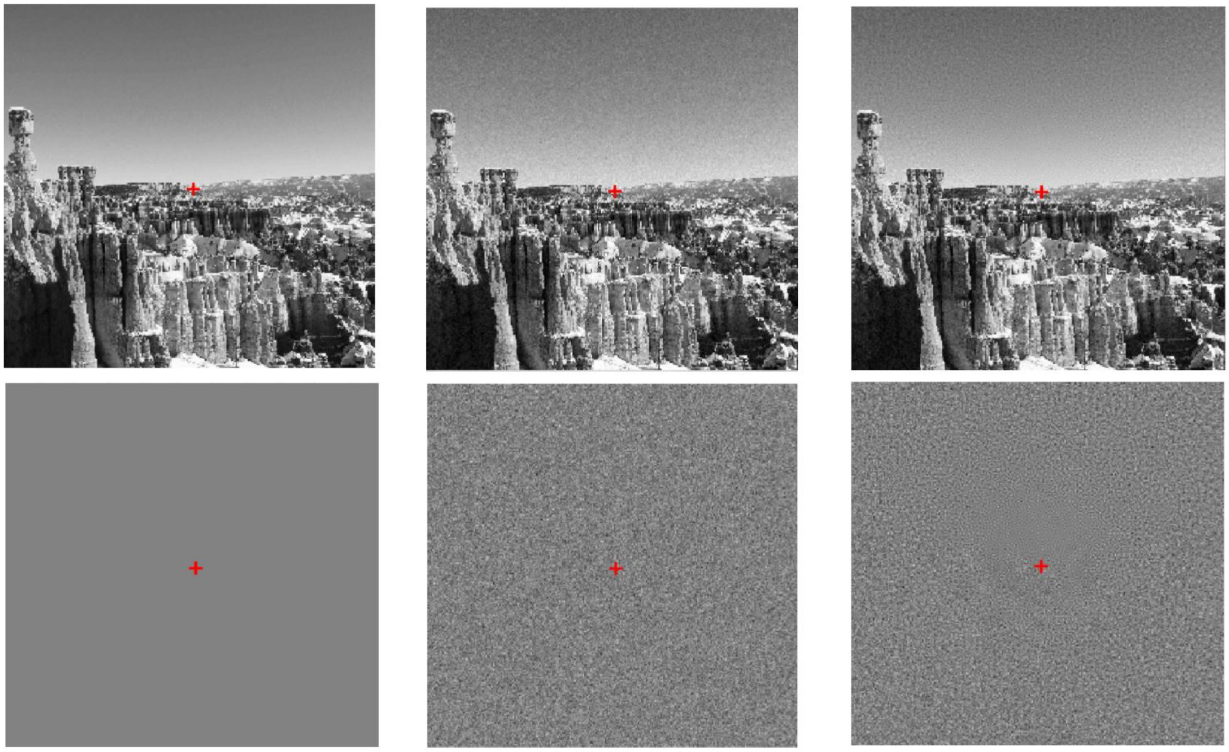

Figure 3. Minimally visible distortion, generated using a foveated version of the perceptual model. The sampling frequency used in the example implies that images should subtend 8 deg. Top row show images, and bottom row shows the difference between that image and the original. Left: original image. Middle: original image corrupted by homogeneous white noise. Right: original image corrupted by noise that is minimally visible, according to the foveated perceptual model. Mean squared error of the two corrupted images is identical (image-domain PSNR = 28.1).

increased from their original values at a rate of 0.06 subtended degrees per eccentricity degree. Then using the gradient-descent MAD procedure described above, we synthesized an image with noise that is barely visible according to this foveated model.

Figure 3 shows an example, in comparison to an image corrupted by white noise of the same average amplitude. Note that the perceptual model injects noise that is modulated over both space and spatial frequency, depending on a combination of eccentricity and local image content.

Conclusion

We've developed an iterative psychophysical methodology, MAD-optimization, for selecting the parameters of perceptual distortion models, and have used it to optimize two example models. Both optimized models show successive reductions in statistical redundancy, as well as reductions in the visibility of distortions that they deem minimal, over each model transformation. And both models show high correlation with human quality ratings. Armed with this optimization methodology, we hope to extend these models with additional stages, thus mimicking the hierarchical structure of the human visual system.

REFERENCES

[1] Fukushima, K., "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics* **36**(4), 193–202 (1980).

[2] Douglas, R. J., Martin, K. A., and Whitteridge, D., "A canonical microcircuit for neocortex," *Neural Computation* **1**, 480–488 (1989).

[3] Heeger, D. J., Simoncelli, E. P., and Movshon, J. A., "Computational models of cortical visual processing," *Proc. Nat'l Academy of Science* **93**, 623–627 (January 1996).

[4] Riesenhuber, M. and Poggio, T., "Hierarchical models of object recognition in cortex," *Nature Neuroscience* **2**, 1019–1025 (1999).

[5] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on* **13**, 600–612 (Apr. 2004).

[6] Seshadrinathan, K. and Bovik, A. C., "Unifying analysis of full reference image quality assessment.," in [*ICIP*], 1200–1203, IEEE (2008).

[7] Laparra, V., Muñoz Marí, J., and Malo, J., "Divisive normalization image quality metric revisited," *JOSA A* **27**(4), 852–864 (2010).

[8] Ranzato, M., Huang, F., Boureau, Y., and LeCun, Y., "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in [*Proc. Computer Vision and Pattern Recognition Conference (CVPR'07)*], IEEE Press (2007).

[9] Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y., "What is the best multistage architecture for object recognition?," in [*Proc. International Conference on Computer Vision (ICCV'09)*], IEEE (2009).

[10] Salakhutdinov, R. and Hinton, G., "An Efficient Learning Procedure for Deep Boltzmann Machines," *Neural Computation* **24**, 1967–2006 (Apr. 2012).

[11] Carandini, M. and Heeger, D. J., "Normalization as a canonical neural computation.," *Nature reviews. Neuroscience* **13**, 51–62 (Jan. 2012).

[12] Teo, P. and Heeger, D., "Perceptual image distortion," *Proceedings of the SPIE* **2179**, 127–141 (1994).

[13] Malo, J., Epifanio, I., Navarro, R., and Simoncelli, E. P., "Non-linear image representation for efficient perceptual coding," *IEEE Trans Image Processing* **15**, 68–80 (Jan 2006).

[14] Lyu, S. and Simoncelli, E. P., "Statistically and perceptually motivated nonlinear image representation," in [*Proc. SPIE, Conf. on Human Vision and Electronic Imaging XII*], Rogowitz, B., Pappas, T. N., and Daly, S. J., eds., **6492**, 67–91, Society of Photo-Optical Instrumentation, San Jose, CA (January 28-30 2007).

[15] Wang, Z. and Simoncelli, E. P., "Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual discriminability," *Journal of Vision* **8**, 1–13 (Sep 2008).

[16] Ponomarenko, N., Carli, M., Lukin, V., Egiazarian, K., Astola, J., and Battisti, F., "Color image database for evaluation of image quality metrics," *Proc. Int. Workshop on Multimedia Signal Processing* , 403–408 (Oct. 2008).

[17] LeCun, Y. and Bengio, Y., "Convolutional networks for images, speech, and time-series," in [*The Handbook of Brain Theory and Neural Networks*], Arbib, M. A., ed., MIT Press (1995).

[18] Barlow, H., "Possible principles underlying the transformation of sensory messages," in [*Sensory Communication*], Rosenblith, W., ed., 217–234, MIT Press, Cambridge, MA (1961).

[19] Malo, J. and Laparra, V., "Psychophysically tuned divisive normalization approximately factorizes the PDF of natural images," *Neural Computation* **22**(12), 3179–3206 (2010).

[20] Watson, A. and Malo, J., "Video quality measures based on the standard spatial observer," *Proc. IEEE ICIP* **3**, 41–44 (2002).

[21] van Hateren, J. and van der Schaaf, A., "Independent component filters of natural images compared with simple cells in primary visual cortex," *Proc.R.Soc.Lond. B* **265**, 359–366 (1998).

[22] Schwartz, O. and Simoncelli, E., "Natural signal statistics and sensory gain control," *Nat. Neurosci.* **4**(8), 819–825 (2001).

[23] Lyu, S. and Simoncelli, E. P., "Nonlinear extraction of 'independent components' of natural images using radial Gaussianization," *Neural Computation* **21**, 1485–1519 (Jun 2009).

[24] Wang, Z., Simoncelli, E. P., and Bovik, A. C., "Multi-scale structural similarity for image quality assessment," in [*in Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers, (Asilomar*], 1398–1402 (2003).

[25] Sheikh, H., Bovik, A., and de Veciana, G., "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing* **14**, 2117–2128 (Dec 2005).

[26] Zhang, L., Zhang, L., Mou, X., and Zhang, D., "Fsim: A feature similarity index for image quality assessment.," *IEEE Transactions on Image Processing* **20**(8), 2378–2386 (2011).

[27] Rosenholtz, R., "What your visual system sees where you are not looking," in [*Proc. SPIE: Human Vision and Electronic Imaging*], **XVI** (2011).

[28] Freeman, J. and Simoncelli, E. P., "Metamers of the ventral stream," *Nature Neuroscience* **14**, 1195–1201 (Sep 2011).