Revisiting perceptual distortion for natural images: mean discrete structural similarity index

Christopher Hillar*, Sarah Marzen*,†

*Redwood Center for Theoretical Neuroscience
University of California, Berkeley
Berkeley, CA, 94708, USA
chillar@msri.org

†Physics of Living Systems
Massachusetts Institute of Technology
Cambridge, MA, 02148, USA
semarzen@mit.edu

Abstract

A challenge in image processing is quantifying the perceptual quality of distorted images. Solutions to this problem allow lossy compression algorithms to be more easily and accurately evaluated. Motivated by failings of mean-squared error (MSE/PSNR), Wang, Bovik, and others proposed a perceptual image measure called mean structural similarity (MSSIM), which decomposes the distortion of image patches into three components: a difference in mean luminance, a difference in luminance variance, and a difference in structure. We present a new measure, mean discrete structural similarity (MDSSIM), that replaces the structural comparison of MSSIM with the Hamming distance between suitably discretized original and distorted image patches. To assess its performance, we apply this new image measure to a standard human psychophysics dataset, the LIVE Image Quality Assessment Database (Release 2). The high correlation of MDSSIM with human scores suggests, consistent with experiment and well-known results about lossy compression, that the human visual system may be fundamentally concerned with discrete structure in natural images.

Introduction

An active area of image processing research is finding accurate and easily calculated measures of an image's perceptual quality. Mean-squared error (MSE) is ubiquitous, but unlike the human visual system (HVS), MSE is very sensitive to luminance shifts and contrast increases, and is invariant to reordering of pixels, among other issues. Therefore, researchers have been searching for a perceptual distortion measure [1, 2] that can replace MSE and its relative, peak signal-to-noise ratio (PSNR).

In 2004, Wang and colleagues developed a full-reference perceptual distortion measure by assuming that the HVS utilizes "structure" in natural images [3] to discern differences. The structural similarity index (SSIM) decomposes the distortion of image patches into a difference in mean luminance, a difference in luminance variance, and a structural component, the cosine of the angle between the original and distorted image patch. These components are evaluated and multiplied together for each patch

2375-0359/17 $31.00 © 2017 IEEE
DOI 10.1109/DCC.2017.84

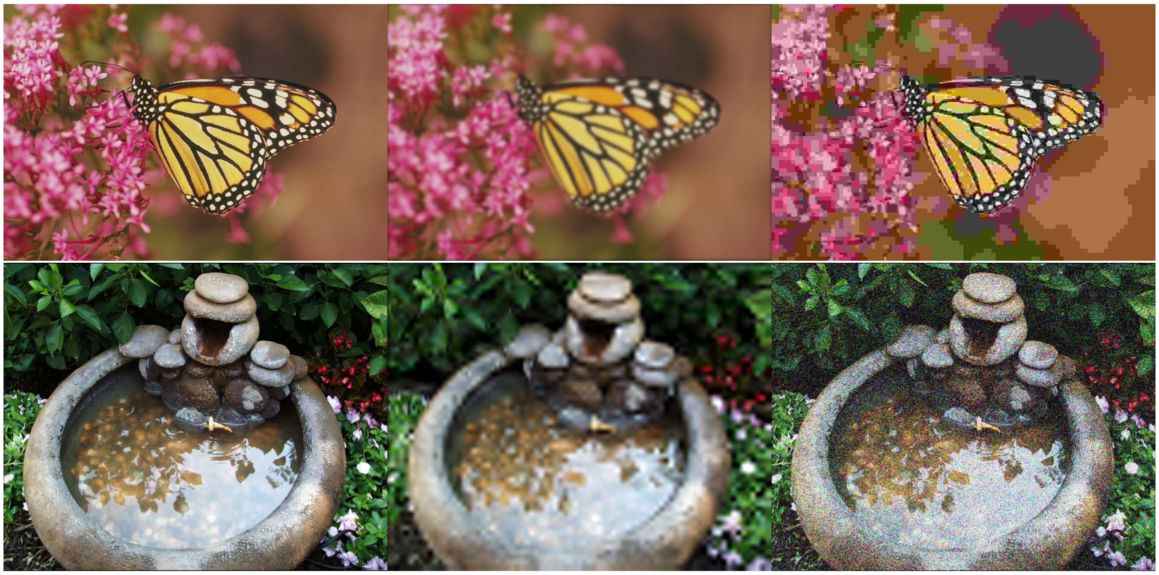

Figure 1: **Sample LIVE dataset images**. Top row (left-to-right): Original, Gaussian blur, Low-quality JPEG; Bottom row (left-to-right): Original, "Fast fading", White noise. (LIVE Dataset URL: http://live.ece.utexas.edu/research/quality/subjective.htm)

in a larger image, and local patch-wise SSIM scores are pooled to calculate the mean SSIM (MSSIM) of the image. MSSIM and variants thereof [4, 5] have been used widely, in some cases replacing MSE/PSNR, although the structural component in MSSIM's computation is directly related to a local (patch-wise) MSE [6, 7]. Workers looking to replace MSE have also used features of the DCT basis [8] or models of the HVS [9] to develop perceptual distortion measures.

Here, we propose a new measure of structural distortion for an image patch: the Hamming distance between suitable binarizations [10] of the distorted and reference patch. Our efforts stem from theorems in rate-distortion theory that identify discrete codings as optimal compressors [11, 12, 13], suggesting that perceptual structure – even in continuous-valued natural images – is perhaps fundamentally discrete.

In the next section, we give background for SSIM and our discrete approach. Then, we describe a new perceptual image distortion measure, called mean *discrete* structural similarity (MDSSIM), and study its performance using the LIVE image quality assessment dataset (Release 2); see Fig. 1 for examples from the dataset. We close with a short discussion examining the potential for our findings.

Background

In this section, we briefly review the MSSIM image measure, discrete recurrent neural networks, and the image compression approach of [10] that utilizes these networks.

Mean structural similarity index

MSSIM is a full-reference image quality assessment (IQA), in that distortion calculation requires access to the original (reference) image. Let $\mathbf{x} \in \{0, ..., 255\}^{L^2}$ be a reference image patch of dimensions $L \times L$, and let $\mathbf{y} \in \{0, ..., 255\}^{L^2}$ be the corresponding distorted image patch. SSIM indices, measuring the similarity between reference and distorted image patches, are the product of three factors:

$$SSIM(\mathbf{x}, \mathbf{y}) = l(\mathbf{x}, \mathbf{y})^{\alpha} c(\mathbf{x}, \mathbf{y})^{\beta} s(\mathbf{x}, \mathbf{y})^{\gamma}, \tag{1}$$

in which $l(\mathbf{x}, \mathbf{y})$ measures the dissimilarity in mean luminance, $c(\mathbf{x}, \mathbf{y})$ measures the dissimilarity in variance of luminance, and $s(\mathbf{x}, \mathbf{y})$ represents the measure of structural dissimilarity. Following the original work [3], we define:

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{2}$$

where $\mu_x = \frac{1}{L^2} \sum_i x_i,\ \mu_y = \frac{1}{L^2} \sum_i y_i$. We also define, as in the original work,

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{3}$$

where $\sigma_x^2 = \frac{1}{L^2-1} \sum_i (x_i - \mu_x)^2,\ \sigma_y^2 = \frac{1}{L^2-1} \sum_i (y_i - \mu_y)^2$. (The constants $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$ are there to prevent numerical instability.)

Finally, the structural dissimilarity $s(\mathbf{x}, \mathbf{y})$ is taken to be the cosine of the angle between $\mathbf{x}$ and $\mathbf{y}$, which is exactly related to the mean-squared error between them [6, 7]. The exponents are typically set to $\alpha = \beta = \gamma = 1$, and a common choice for patch size is $L = 11$. The mean SSIM index, MSSIM, averages the SSIM indices over all patches.

For the purposes of this work, the MSSIM numbers computed here used the MATLAB routines supplied by the authors of [3].

Discrete recurrent neural networks (DRNNs)

Our definition of structure in an image patch will be an attractor of a Hopfield network [14] (i.e., a symmetrically-weighted McCulloch-Pitts net [15]) whose weights are estimated to match the statistics of natural image patches. This section and the next provide background for this connection to theoretical neuroscience and natural patch modeling.

We start by reviewing the *Lenz-Ising model* [16]. Let $\mathbf{x} = (x_1, \ldots, x_n) \in \{0, 1\}^n$ be a length-n binary vector. The probability $p(\mathbf{x})$ of a particular state is given by:

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{i<j} W_{ij} x_i x_j - \sum_{i=1}^{n} \theta_i x_i \right) = \frac{1}{Z} \exp\left(-E_{\mathbf{x}}\right), \tag{4}$$

in which $W = W^{\top} \in \mathbb{R}^{n \times n}$ is a symmetric matrix with zero diagonal (the *weight matrix*), the column vector $\theta \in \mathbb{R}^n$ is a bias or *threshold* term, and $Z = \sum_{\mathbf{x}} \exp(-E_{\mathbf{x}})$ is the *partition function* (which normalizes p to sum to 1). States with high probability $p(\mathbf{x})$ have low *energy* $E_{\mathbf{x}}$ (defined in Eq. 4), and vice versa.

A Hopfield network equips this set of n "neurons" with a dynamics that, when given any initial configuration, finds a nearby state that is a local maximum of the

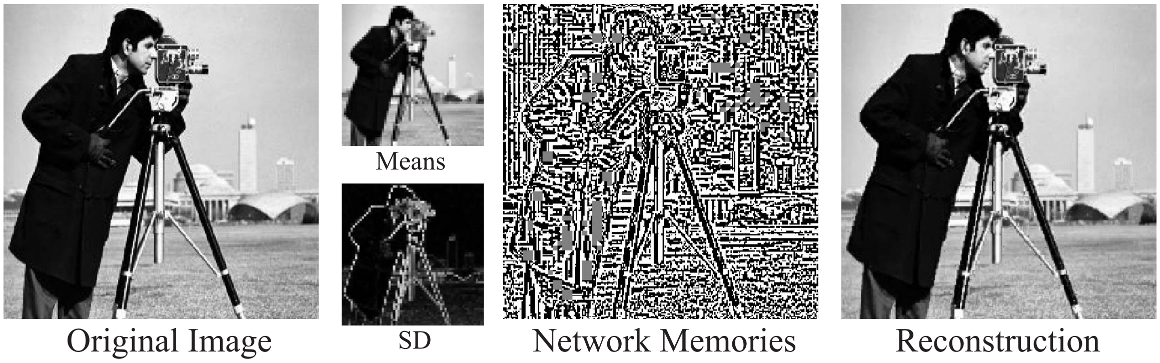

Figure 2: **Example** 4×4 **ON/OFF encoding** [10]: Grayscale 256×256 "cameraman", 64×64 patch means and standard deviations, discrete structural primitives (network memories/attractors), and reconstruction.

probability distribution. A *dynamics update* of $\mathbf{x}$ consists of replacing (in some fixed order through all nodes) each x_i in $\mathbf{x}$ with:

$$x_i = \begin{cases} 1 & \text{if } \sum_{j \neq i} W_{ij} x_j > \theta_i, \\ \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

A fundamental property of Hopfield networks is that dynamics does not increase energy $E_{\mathbf{x}}$. Using this fact, it can be shown that after a finite number of updates, each initial state $\mathbf{x}$ converges to its *attractor* $\mathbf{x}^*$ (or *memory*), which is a fixed-point of the dynamics. Sometimes this property is expressed by saying that the energy $E_{\mathbf{x}}$ is a "Lyapunov function" for the network dynamics. Another useful intuition is that the dynamics is an inference technique, "freezing" a noisy version of a memory into a probable nearby state.

In some applications, one knows *a priori* which memories need to be stored by the neural network. However, in the application considered here, no such knowledge is available. Instead, we find weights of the Hopfield network by fitting the distribution in Eq. 4 to the empirical distribution of image patches. One of the fastest ways to do this is using minimum probability flow (MPF) parameter estimation [17, 18]. Hopfield networks whose weights are estimated using MPF store more memories, more robustly, and more quickly than other better-known weight-training methods [18].

Discrete image patch coding

Consider a reference grayscale image patch $\mathbf{x}$ and a distorted image patch $\mathbf{y}$, both of size $L \times L$ and each normalized to have zero mean pixel intensity; i.e., $\mathbf{x}$ is replaced by $\mathbf{x} - \mu_x$, and similarly for $\mathbf{y}$. We shall explain in this section how to determine for each such patch a binary vector of size $2L^2$ representing its structural component. The structural similarity between the original patches is then computed from the Hamming distance between the binary components.

Consider first the following discretization scheme from [10] to turn a zero-mean patch into a binary vector of length $2L^2$. Every pixel x of the $L \times L$ patch is assigned

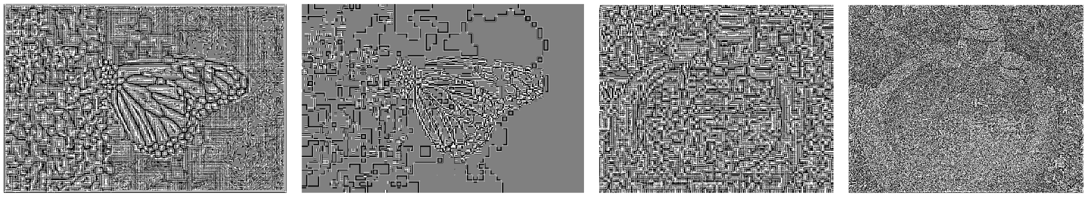

Figure 3: **Structural components**. Discrete structure in distorted images from Fig. 1. Left-to-right: Gaussian blur, low-quality JPEG, Fast fading, and White noise distortions.

two Hopfield neurons – one "ON" neuron, and one "OFF" neuron. We discretize each of these x to a pair of binary values $(ON, OFF) \in \{0, 1\}^2$ according to a parameter $\alpha \geq 0$. When $x > \alpha$, the discretized pixel is assigned $(ON, OFF) = (1, 0)$; similarly, when $x < -\alpha$, we have $(ON, OFF) = (0, 1)$; and finally, when $x \in [-\alpha, \alpha]$, we have $(ON, OFF) = (0, 0)$. We choose α so that 2α is the smallest pixel intensity difference that can possibly occur. Since we have integer-valued image intensities, this assigns $\alpha = 1/2$. We can thus convert any grayscale $L \times L$ image patch into a binary vector of length $2L^2$, a procedure that we call ON/OFF ternarization.

By collecting these binary vectors over natural images, we inherit a probability distribution over ternarized patches. And by matching a Lenz-Ising model / Hopfield network to the empirical distribution of ternarized natural image patches (obtained, e.g., from the van Hateran image database [19]), we determine a network with dynamics that can act on ternarized patches to output discrete attractors. In the case $L = 4$, it was found that the attractors of a trained network consisted of all *binary* attractors (i.e., the all-zero pattern or those with each ON/OFF pair having exactly one neuron firing). Thus, the ON/OFF network acts to coarse-grain the space of possible 3^{L^2} ON/OFF patterns into the subset of $(2^{L^2} + 1)$ binary ones.

Here, we use these attractors as structural representatives for determining image similarity; see Fig. 3 for some examples. However, in the image compression approach of [10], the attractor was utilized to "recover" a continuous primitive for reconstruction of a continuous patch given the attractor. In this case, representing each non-overlapping patch in an image with a mean, standard deviation, and continuous representative of an attractor corresponds to a high quality 4× lossy compression of a natural image. A sample encoding/decoding is depicted in Fig. 2, and more details can be found in [10, 20].

Results

Calculating MSSIM consists of two steps: quantifying the structural dissimilarity of corresponding image patches; and pooling the dissimilarity scores of all the image patches. Improvements to MSSIM (and other distortion measures) have been made by carefully considering the way in which patch-wise scores are pooled [5]. We instead focus on more accurately quantifying dissimilarity by developing a new notion of the structure of image patches. In other words, we propose altering only the structural dissimilarity measure $s(\mathbf{x}, \mathbf{y})$.

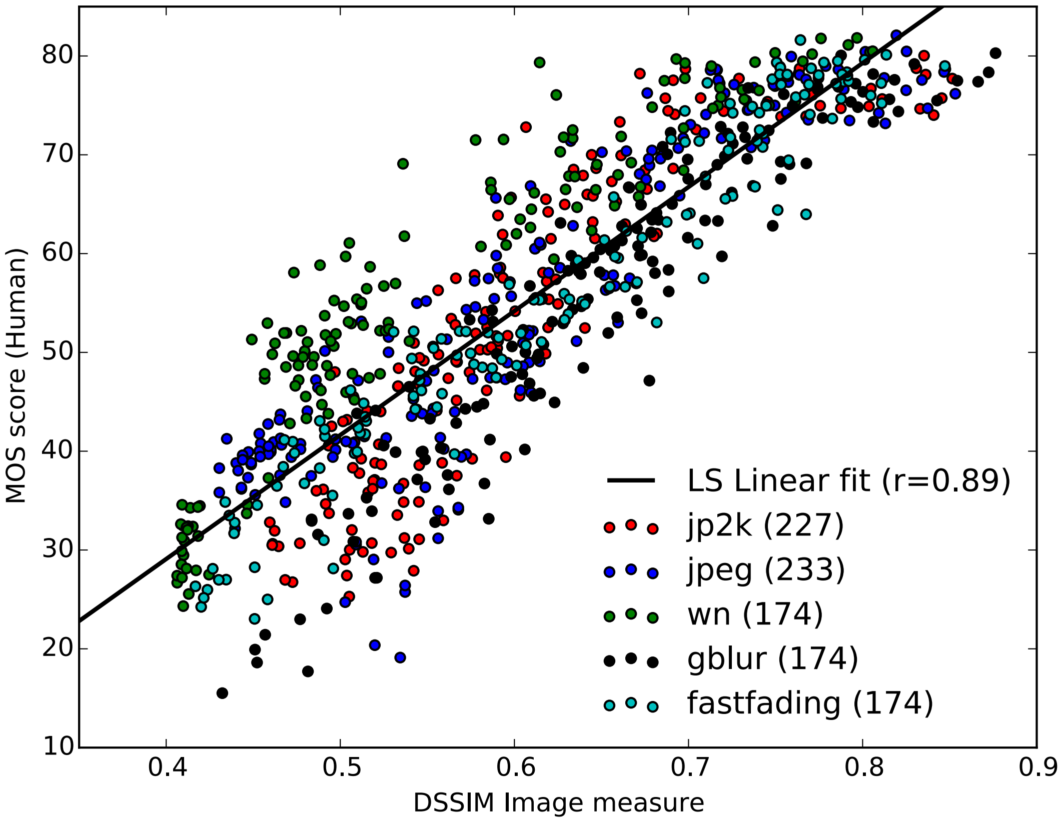

Figure 4: **Perceptual evaluation of DSSIM**. Mean opinion scores (MOS) against mean DSSIM for all reference and distorted images in the UT Austin LIVE dataset (Release 2), consisting of: 227 JPEG2000 compressed images (red dots), 233 JPEG compressed images (blue), 174 White noise (green), 174 Gaussian blur (black), 174 Fast fading (cyan).

Let D_x and D_y be the Hopfield attractors, as described above, corresponding to patches $\mathbf{x}$ and $\mathbf{y}$, respectively. The *discrete structural similarity index* (DSSIM) replaces the function $s(\mathbf{x}, \mathbf{y})$ with one minus the mean of the absolute difference between D_x and D_y (this is a constant multiple of Hamming distance).

The major finding here is that mean DSSIM (MDSSIM) is significantly correlated with mean opinion scores (MOS) from the UT Austin LIVE image quality assessment dataset, Release 2. This dataset consists of 779 distorted images (Fig. 1), each of which received a human quality score from 10 to 90. For simplicity in our analysis and to also be consistent with the scheme of [10], we computed a mean score over non-overlapping 4×4 patch pairs in the images. Thus, we averaged over all non-overlapping 4×4 regions a similarity score involving the corresponding means, variances, and discrete structures. In Fig. 3, we show how four different distortions from the LIVE dataset impact the ON/OFF discretization above.

In Fig. 4, we provide a scatterplot of the MDSSIM scores relative to MOS over all the reference/distorted image pairs in the LIVE dataset. The shading of the points in the figure separates the images into their distortion type. To quantitatively assess correlation between scores, we calculated Pearson correlation, Spearman correlation,

Table 1: Correlation coefficients with MOS scores

	Pearson (linear)	Spearman (logistic)	Kendall
PSNR	.80	.82	.62
MSSIM	.74	.85	.66
MDSSIM	.89	.90	.71

and Kendall's rank correlation using the Python package Scipy. Table 1 shows the results of these calculations.

It should be noted that MSSIM and PSNR predict MOS scores better when passed through certain nonlinearities. For simplicity here, we use straight linear correlation, but we stress this limitation in our analysis. In future work, we hope to more thoroughly compare our approach with others in the literature.

Discussion

Previous work suggests that the local structure in natural images is well-represented by cleverly discretized patches [10]. We have modified the structural similarity index described in [3], using this definition of structure, into an easily-calculable *discrete* structural similarity. This, in turn, results in a new perceptual distortion measure for images called mean discrete structural similarity, MDSSIM.

We have not yet investigated the effects of alternative pooling strategies or multi-scale extensions, which can greatly improve IQA performance, nor have we studied its relationship to more modern approaches [8, 9]. However, the results presented here are promising, suggesting that such variants on DSSIM will continue to yield improvements. Even without further modification, DSSIM is efficiently calculable, and so it could be used, for instance, to optimize various lossy compression algorithms (e.g., using classical rate-distortion theory [20]).

The literature on perceptual distortion contains several proposals for image measures constructed partly based on our understanding of the human visual system (e.g., [9]). Conversely, the surprisingly good performance of DSSIM suggests a potential new understanding of part of the transformation performed by the HVS. The structure in natural image patches postulated by DSSIM involves discretization of an image patch and subsequent convergence under the network dynamic. Experiments not shown here suggest that the binary discretization rather than the Hopfield dynamics were key to the high performance of DSSIM as a perceptual distortion measure. Although the attractors are not necessary for achieving high correlation scores with MOS, we suspect that they are still helpful for robustness of the image measure, much as the network was not necessary to see much of the productivity in the image compression scheme of [10].

Nonetheless, the findings here do suggest that the HVS may be far more responsive to local discrete structure in natural images than previously expected, pointing the way towards novel experiments in vision science, e.g. by extending the psychophysics experiments of [21] to the lossy image model of [10].

References

[1] Z. Wang and A. Bovik, "Mean squared error: love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[2] ——, "Modern image quality assessment," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1, pp. 1–156, 2006.

[3] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[4] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. IEEE, 2003, pp. 1398–1402.

[5] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.

[6] K. Seshadrinathan and A. Bovik, "Unifying analysis of full reference image quality assessment," in *2008 15th IEEE International Conference on Image Processing.* IEEE, 2008, pp. 1200–1203.

[7] R. Dosselmann and X. Yang, "A comprehensive assessment of the structural similarity index," *Signal, Image and Video Processing*, vol. 5, no. 1, pp. 81–91, 2011.

[8] T. Richter, "On the mDCT-PSNR image quality index," in *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on.* IEEE, 2009, pp. 53–58.

[9] R. Mantiuk, K. Kim, A. Rempel, and W. Heidrich, "HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions," in *ACM Transactions on Graphics (TOG)*, vol. 30, no. 4. ACM, 2011, p. 40.

[10] C. Hillar, R. Mehta, and K. Koepsell, "A Hopfield recurrent neural network trained on natural images performs state-of-the-art image compression," in *Image Processing (ICIP), 2014 IEEE International Conference on.* IEEE, 2014, pp. 4092–4096.

[11] J. Smith, "The information capacity of amplitude-and variance-constrained scalar gaussian channels," *Information and Control*, vol. 18, no. 3, pp. 203–219, 1971.

[12] S. Fix, "Rate distortion functions for squared error distortion measures," in *Annual Allerton Conference on Communication, Control and Computing, 16th, Monticello, Ill*, 1978, pp. 704–711.

[13] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. Info. Th.*, vol. 40, no. 6, pp. 1939–1952, 1994.

[14] J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[15] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of mathematical biology*, vol. 5, no. 4, pp. 115–133, 1943.

[16] E. Ising, "Beitrag zur Theorie des Ferromagnetismus," *Zeitschrift fur Physik*, vol. 31, pp. 253–258, 1925.

[17] J. Sohl-Dickstein, P. Battaglino, and M. DeWeese, "New method for parameter estimation in probabilistic models: minimum probability flow," *Physical Review Letters*, vol. 107, no. 22, p. 220601, 2011.

[18] C. Hillar, J. Sohl-Dickstein, and K. Koepsell, "Efficient and optimal binary Hopfield associative memory storage using minimum probability flow," in *4th Neural Information Processing Systems (NIPS) workshop on Discrete Optimization in Machine Learning (DISCML)*, 2012.

[19] J. van Hateren and A. van der Schaaf, “Independent component filters of natural images compared with simple cells in primary visual cortex,” *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 265, no. 1394, pp. 359–366, 1998.

[20] R. Mehta, S. Marzen, and C. Hillar, “Exploring discrete approaches to lossy compression schemes for natural image patches,” in *European Signal Processing Conference (EUSIPCO 2015)*, 2015.

[21] H. Gerhard, F. Wichmann, and M. Bethge, “How sensitive is the human visual system to the local statistics of natural images?” *PLoS computational biology*, vol. 9, no. 1, p. e1002873, 2013.