Efficient image representations and features

Michael Dorr[a], Eleonora Vig[b], and Erhardt Barth[c]

[a]Schepens Eye Research Institute, Harvard Medical School, 20 Staniford St, Boston, MA, USA
[b]Center for Brain Science, Harvard University, 52 Oxford St, Cambridge, MA, USA
[c]Institute for Neuro- and Bioinformatics, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany

ABSTRACT

Interdisciplinary research in human vision and electronic imaging has greatly contributed to the current state of the art in imaging technologies. Image compression and image quality are prominent examples and the progress made in these areas relies on a better understanding of what natural images are and how they are perceived by the human visual system. A key research question has been: given the (statistical) properties of natural images, what are the most efficient and perceptually relevant image representations, what are the most prominent and descriptive features of images and videos?

We give an overview of how these topics have evolved over the 25 years of HVEI conferences and how they have influenced the current state of the art. There are a number of striking parallels between human vision and electronic imaging. The retina does lateral inhibition, one of the early coders was using a Laplacian pyramid; primary visual cortical areas have orientation- and frequency-selective neurons, the current JPEG standard defines similar wavelet transforms; the brain uses a sparse code, engineers are currently excited about sparse coding and compressed sensing. Some of this has indeed happened at the HVEI conferences and we would like to distill that.

Keywords: image coding, sparse coding, predictive coding, intrinsic dimension, nonlinear filters, visual perception, unsupervised learning, gain control, divisive normalization

1. INTRODUCTION

Given the 25^{th} anniversary of the HVEI conference, we here give an overview of contributions that the conference has made to the topic of efficient image representations. Of course, we do this with a particular perspective and theoretical background that is introduced in the next section.

We all know that, no matter how long we wait, a display with randomly selected colors will never look like an image taken in the real world. In other words, if we treat images as vectors that represent points in a high-dimensional space, this space of all possible images is more or less empty and the natural images all sit in a very small subspace. Perceptual issues matter and thus the space of, for example, high-quality images is even smaller. Although it would make sense to stick to these subspaces (many people would like to delete everything outside the space of nice images), we have, even after 25 years of HVEI, rather few ideas of how to do that in a strict sense. Efficient representations are one way of dealing with the problem: step by step we would like to reduce the redundancies in images. Some redundancies are obvious and can easily be removed, e.g., by de-correlation; others require trickier computations, but most redundancies are still hidden. So, we are looking forward to another 25 years.

Further author information: (Send correspondence to E.B.)
E.B.: E-mail: barth@inb.uni-luebeck.de, Telephone: +49 451 5005503
M.D.: E-mail: michael.dorr@schepens.harvard.edu
E.V.: E-mail: vig@fas.harvard.edu

Human Vision and Electronic Imaging XVIII, edited by Bernice E. Rogowitz, Thrasyvoulos N. Pappas, Huib de Ridder,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 8651, 86510R • © 2013 SPIE-IS&T
CCC code: 0277-786X/13/$18 • doi: 10.1117/12.2002300

2. THEORETICAL BACKGROUND

2.1 Linear filters

Fourier methods, linear filter banks, and wavelets have been used extensively to extract useful image representations and to construct models of visual representations. In terms of the notation that we would like to use here, linear transforms are just a rotation of the coordinate frame:

$$\vec{x}_{LF} = \mathbf{W}\vec{x} \tag{1}$$

where $\vec{x} \in \Re^N$ is the vector that represents the image, or an image block, $\mathbf{W}$ is a rotation matrix, and $\vec{x}_{LF}$ is the filtered image, or image block.

2.2 Learned representations and sparse coding

As an important alternative, representations (that could be linear) can be learned based on a representative set of sample data $\vec{x}_i, i = 1, ..., p$. For the learning procedure, optimization criteria are required. One useful strategy is to maximize the efficiency of the representation, i.e., to represent the data as compactly as possible. A classical approach for increasing the efficiency of the representation is to reduce the dimension of the data vector, i.e., we look for a mapping $\vec{x} \in \Re^N \longmapsto \vec{y} \in \Re^M, \qquad M < N$. Possible criteria are (i) minimal loss of information (generic criterion), or (ii) better pattern-recognition performance (task-specific criterion). For biological systems, criteria such as a reduced neural activity can be important. Ideally, the vector $\vec{y}$ contains the features of the data $\vec{x}$ that are the most relevant (for a more or less specific set of tasks). However, properties of the data vector $\vec{y}$ other than the reduced dimension may be of interest, and the challenge is to find generic criteria for data representation that turn out to be useful for various tasks.

Principal component analysis (PCA) and whitening The PCA is defined as $\vec{x}_{PCA} = \mathbf{U}^T\vec{x}$, where the matrix $\mathbf{U}$ is the matrix that diagonalizes the covariance matrix $\mathbf{C} = \langle \vec{x}_i \vec{x}_i^T \rangle$. The PCA minimizes the reconstruction error when the dimension of $\vec{x}_{PCA}$ is reduced by projection (dropping the $N - M$ components of $\mathbf{U}^T\vec{x}$ that correspond to the smallest eigenvalues of $\mathbf{C}$). Moreover, the PCA maximizes the mutual information between the data $\vec{x}$ and the representation $\vec{y}$. Whitening is usually used after PCA to normalize the variances to unity along all principal directions:

$$\vec{x}_{wht} = \mathbf{S}^{-1}\mathbf{U}^T\vec{x}. \tag{2}$$

The diagonal matrix $\mathbf{S}^2 = \mathbf{U}^T\mathbf{C}\mathbf{U}$ contains the variances in the directions of the principal components and $\mathbf{S}$ the standard deviations. In total, the PCA rotates the signal to the main axes and these are then scaled by $\mathbf{S}^{-1}$ such that the variance is equal in all main directions. After whitening the components of the signal $\vec{x}_{wht}$ are uncorrelated and have unit variances.

Independent Component Analysis (ICA) The ICA is defined by the transformation

$$\vec{x}_{ICA} = \mathbf{V}\vec{x}_{wht} \tag{3}$$

where $\mathbf{V}$ is chosen such as to minimize the multi-information $I(\mathbf{V}\vec{x}_{wht})$. Alternative implementations aim at maximizing the kurtosis or the negentropy of $\mathbf{V}\vec{x}_{wht}$. The intuition is that all linear transforms on the whitened data will preserve the property of decorrelation. Out of all the decorrelating transforms, the ICA picks the one which minimizes the multi-information (thus minimizes statistical dependency since the multi-information is defined as the Kulback-Leibler divergence between the joint distribution and the product of its marginals) or maximizes the kurtosis.

Linear version of sparse coding A sparse representation is one where only few components of the data vector are different from zero. To obtain a sparse representation we can proceed similar to the ICA

$$\vec{x}_{SC} = \mathbf{W}\vec{x} = \mathbf{V}\vec{x}_{wht} \tag{4}$$

and determine $\mathbf{W}$ or $\mathbf{V}$ such as to maximize the number of zero components in $\vec{x}_{SC}$. Since the ICA is usually computed by maximizing the kurtosis, ICA and SC yield similar results (the sparser the representation the higher the kurtosis).

Sparse coding as introduced by Olshausen and Field[21,22] The principle of sparse coding is implemented as a minimization problem with the cost function $E = -E_{preserveinformation} - \lambda E_{sparseness}$, where the individual energy terms are the mean reconstruction error $E_{preserveinformation} = -||\mathbf{W}^{-1}\vec{x}_{SC} - \vec{x}||^2$ and a sparseness term that favors small values for the components of $\vec{x}_{SC}$. The goal is to learn the representation $\vec{x}_{SC}$ and the basis functions, i.e., the rows of $\mathbf{W}^{-1}$ by solving the nested optimization problem:

$$\min_{\mathbf{W}^{-1}}(\min_{\vec{x}_{SC}} E). \tag{5}$$

Note that while the reconstructed signal $\mathbf{W}^{-1}\vec{x}_{SC}$ is obtained from the representation $\vec{x}_{SC}$ by a linear transform (data are represented as weighted sums of basis functions), the representation $\vec{x}_{SC}$ is, in general, not a linear transform of the data $\vec{x}$. Rather, the representation $\vec{x}_{SC}$ is obtained for every new data point $\vec{x}$ by minimization of the criterion E. So, this is quite different from the previous strategies in case of the PCA and the ICA: the available data set was there used to derive one (linear) transform, which could then be used to represent any new data point. Now, the learning procedure is applied not only to some initial data set but to every new data point that needs to be represented. Also note that the set of basis functions can be over-complete, i.e., one can have more basis functions than dimensions of the input space. Moreover, the basis functions need not be orthogonal.

2.3 Intrinsic dimension and nonlinear filters

Let an image be modeled by a function $f : \mathbf{R}^2 \to \mathbf{R}$. Given an (open) region Ω, for all $(u, v) \in \Omega$, either (a) $f(u, v) =$ constant; or (b) $f(u, v) = g(au + bv)$, for some g, a, b; or (c) f varies along all directions. The image f is said to locally have intrinsic dimension $0, 1$, or 2, respectively ($i0D$, $i1D$, $i2D$ for short).

In terms of the initial problem definition, we now search for a $\vec{y} = F(\vec{x})$ that is equal to zero if $\vec{x}$, now defined as the vector that contains all the pixel values $f(u, v)$ in Ω, is $i0D$ or $i1D$. $\vec{y}$ should be different from zero if $\vec{x}$ is $i2D$. We call such a transformation an $i2D$ transform or $i2D$ operator.

Since it has been shown that images are fully determined by the $i2D$ regions,[2] we know that $i2D$ transforms can be made such that the difference between $\vec{x}$ and the reconstructed data $G(\vec{y})$ is small. The concept can easily be extended to more dimensions, such that, in case of videos, the intrinsic dimension inD varies from $n = 0, ..., 3$.

2.4 A missing link

Why should a nonlinear $i2D$ transform be better than the linear transformations discussed above? One can easily show that linear transformations cannot be $i2D$ transforms.[3,4] Therefore any linear transform will fail to create zero output coefficients for all $i0D$ and $i1D$ inputs although the coefficients could be zero without loss of information. Differential geometry provides a nice framework for understanding why information is concentrated at (curved) $i2D$ regions of an image[2–4] and a number of applications have confirmed this fact.[5,6]

Moreover, it has been shown that $i0D$ and $i1D$ regions are frequent in natural images.[7] Therefore, an $i2D$ transform can produce, without loss of information, a higher degree of sparsification than a linear transform. Recent results by Olshausen[8] (see also this volume), and Labusch et al.[9] show that sparse-coding algorithms do indeed lead to nonlinear $i2D$ operators, especially if the basis is allowed to be over-complete. However, the authors of these papers hardly mention the $i2D$ operators that result from learning sparse codes. This is most likely due to the missing theoretical link between the geometric view (information and curvature) on one side and the statistical view and learning theory on the other. Maybe the next 25 years will fix that.

2.5 Gain control, Radial Gaussianisation (RG), and Divisive Normalization (DN)

Lyu and Simoncelli[10] have shown that in case of natural images the pdf is often better modeled as elliptical than as factorial (the latter is the ICA model) and have introduced RG as a method that can find representations with independent components in case of elliptical symmetric densities (ESD), i.e., densities for which the points of ct. probability are ellipsoids. Linear transforms cannot make the components of signals with ESD more independent after whitening because $p(\mathbf{V}\vec{x}_{wht}) = p(\vec{x}_{wht})$ and therefore $I(\mathbf{V}\vec{x}_{wht}) = I(\vec{x}_{wht})$. In this case, RG, defined as

$$\vec{x}_{rg} = g(||\vec{x}_{wht}||)\frac{\vec{x}_{wht}}{||\vec{x}_{wht}||} \tag{6}$$

where $g(.)$ is chosen such that $\vec{x}_{rg}$ is Gaussian, can be shown to increase statistical independence of band-pass filtered image pixels significantly more than the ICA does. Moreover, it can be shown that DN is a good approximation to RG and is related to cortical gain control.

3. OVERVIEW OF HVEI CONTRIBUTIONS

The HVEI conference has been open to controversial ideas and interdisciplinary approaches that may have been inhibited in the classical vision and image engineering communities. We here give a selective overview of HVEI contributions that are related to our topic. There will most likely be some overlap with related "theme papers" but such overlap should hopefully glue things together.

3.1 Image quality

Historically, early signal processing work was mainly concerned with faithful transmission of images over potentially noisy, error-introducing channels, but the practical importance of knowing where the good images are in the space of natural images was quickly realized. However, it turned out that the classical definition of faithful reproduction, which looks at pixel-wise differences, is only a very weak predictor of human judgement of image quality.[11] Some image transformations, such as a small translation, may be unnoticeable even (or especially) if the entire image is affected, while others, such as compression artifacts of block-based coders, may be highly noticeable. Therefore, image quality assessment methods attempt to capture the perceptual effect of image transmission, compression, or transformation to another representation; this is the dual problem of deciding what is the perceptually relevant information that needs to maintained in an efficient image representation.

While learned and sparse representations or the intrinsic dimension have not been used widely yet in the context of image quality assessment, many related topics have been addressed at HVEI.

Specific models have been presented at HVEI that quantify compression artifacts introduced by specific coders.[12] For the more general class of arbitrary distortions, however, the typical approach is to first use a linear bank of filters of different scales and orientations that resembles the early stages of human visual processing (e.g. the cortex transform[13]) and to measure the errors introduced in each subband.[14] Here, nonlinearities in the metric proved to be essential to mimic human quality judgements.[11]

For natural images, the histograms of wavelet coefficients can be efficiently fitted with a two-parameter generalized Gaussian density (GGD) model.[15] Because these histograms are sensitive to many types of distortions, Wang and Simoncelli[16] presented an efficient design to use an image as its own quality reference during potentially noisy transmission. The (few) GGD parameters of the original image can be transmitted separately through a protected, ancillary channel, and then serve as a reference to the GGD parameters of the transmitted image.

Because of the subjective nature of image quality assessment, often lengthy and expensive experiments with human observers are needed to evaluate new algorithms. In order to alleviate this need, Wang and Simoncelli[17] proposed a new method to assess image quality metrics with fewer human judgements and based on existing metrics. Starting from images with a known distortion level and two competing image metrics, the key idea was to synthesize new images such that the response of one metric is varied while the second metric remains fixed, i.e., to perform a search for those image space regions where the metrics diverge; human judgements then only need to be made for these informative images.

Further insight into human quality judgement revealed that certain image regions are perceptually more important than others. Moorthy and Bovik[18] showed that weighing image quality metrics with the saliency of the underlying image region improved metric performance.

Two common practical applications that are trying to improve image quality are superresolution and denoising. However, one issue with learning on large training sets of "good" patches is that it quickly becomes intractable to store all possible image patches in a high-dimensional space. Therefore, vector quantization techniques are often employed to reduce the number of samples. Li and Adelson[31] described a more efficient approach that gives a nested partition of the learning space and represents the mapping from one image to another parametrically as a function of partition bin.

3.2 Perceptual relevance and learned representations

Attneave and Barlow's redundancy reduction hypothesis postulates that the goal of the computations of the early stages of visual processing is to provide statistical independence. As such, the independent components of natural images give us a set of filters optimized for such statistical independence: the ICA coefficients are the least redundant that are achievable with linear transformations. However, linearity turns out to be a strong constraint and it is not clear what the real gain is from reduction of higher-order dependencies using a linear model.

In order to empirically test the perceptual significance of different coding bases, Bethge et al.[28] developed an efficient psychophysical procedure where subjects were asked to predict missing pixels or the missing ICA or DCT basis for natural image patches. Perhaps surprisingly, missing ICA coefficients were significantly better predictable than DCT coefficients, i.e., the ICA basis exhibited higher perceptual dependencies than the DCT basis despite its lower statistical dependencies.

The problem of unsupervised learning of an invariant representation can be split into two subtasks: dimension reduction and a unified representation of invariant structure and equivariant appearance. Bethge et al.[30] argued that the popular temporal stability (slowness) principle only tackles the first task but does not provide a unique basis for representation, and they learned rotation-invariant steerable bases in two-dimensional subspaces rather than one-dimensional eigenspaces instead.

3.3 Sparse coding

In 1996 Olshausen and Field published two papers at HVEI[20,21] before their highly influential Nature paper.[22] The first paper[20] relates to earlier work on efficient coding that was analyzing the statistics of natural images. Here, the focus is on the $1/f$ fall-off of the amplitude spectra, which is related to the sparseness of structure in images, as opposed to self-similarity. Moreover, the main properties of cortical neurons (localization, frequency selectivity, and orientation) are related to the principle of sparse coding: the steeper the fall-off of the amplitude spectrum, the fewer the number of active neurons tuned to high frequencies.

The second paper[21] focused on the learning of sparse representations and proposed to describe images with a sparse set of learned basis functions drawn from an overcomplete dictionary. In a generative framework, images are modeled as linear superposition of some basis learned from natural images under the constraint of maximizing the sparseness of the representation. The learned bases are localized, oriented bandpass functions, and thus resemble the receptive fields of cortical cells.

A later paper,[23] presented at the 2007 HVEI Special Session on Natural-Image Statistics,[23,24,26,28,30,34,38] extended the principle of sparse coding to the coding of motion signals. Now bilinear, generative image models are considered, since previous sparse codes, based on linear superposition only, cannot optimally encode objects that move or change due to other sources of variation. The idea is that, similar to the visual coding in "what" and "where" streams, the representations of object shape can be separated from object transformations and learned independently. In more general terms, representations are now learned in different subspaces.

Despite the great success of sparse-coding principles, initially there was some skepticism, especially in the engineering communities, because the utility of the sparse representations has been mainly demonstrated by pointing at the resemblance with cortical visual processing. Meanwhile, the technical literature has brought a number of extensions and applications of sparse coding (see, for example, the IEEE Transactions on Selected

Topics in Signal Processing[1]). However, to our knowledge, the first application of sparse coding to a technical problem has been presented at HVEI.[24] Currently, supervised machine-learning techniques, such as the SVM, are well understood and performance gains are hard to obtain by optimizing decision making. Instead, performance varies more with the features that are used to represent the data and good features are often found by heuristics and extensive feature selection. Given that sparse coding delivers optimal features for natural images, the same coding principle should provide good features for other particular data sets. This idea has been applied to the highly competitive MNIST benchmark of hand-written digit recognition. The sparse features have nicely captured the characteristics of the data and, in combination with a maximum rule and a SVM, have led to state-of-the art recognition performance. In fact, the algorithm performed best among all methods that do not make use of additional specific knowledge about the data base.[24] Meanwhile, deep-learning networks are the champions on MNIST and they are also based on extensive, unsupervised learning of (often sparse) features that are specific to the particular data set.

3.4 Gain control

Gain control has not been a hot topic at HVEI, but it has been shown that inclusion of Divisive Normalization (DN) can better predict human recognition performance, in this case the detection of nodules in radiographs.[25] Within the 2007 special session on image statistics, however, Lyu and Simoncelli[26] presented an invertible DN transform with applications to local contrast enhancement and compression. A few years earlier, Valerio and Navarro[27] already showed that an invertible DN stage can remove higher-order statistical dependencies beyond what is possible with linear filters. Both the linear wavelet decomposition stage and the nonlinear DN stage reduced the mutual information by a factor of six each. Two further HVEI papers deal with DN in a different context[35,38] and are discussed in Section 3.5.

3.5 Intrinsic dimension and nonlinear filters

The first paper on intrinsic dimension, and the limits of linear filters in dealing with it, appeared at HVEI.[4] This comprehensive paper has a broad scope and treats quite a number of relevant topics, such as (a) the concept of intrinsic dimension and (b) the limits of linear filters, (c) the related differential-geometry framework, (d) related biological phenomena such as end-stopping and bug detectors, (e) the generalized $i2D$ detector equation based on $\Sigma\Pi$ structures and (f) the compensation equation, (g) the activity distribution of $i2D$ detectors over oriented filters, (h) $i2D$-detectors as AND operations on oriented filters, and (i) the topological invariance of integrated $i2D$ activities.

In a tour de force (with 32 pages and 93 citations, where else could it have been published at that time?) an invited HVEI paper[29] extends the framework of the original paper[4] and brings together many different aspects of previous and later work. It starts by showing the limits of the traditional approaches to predictive coding and then develops a comprehensive research program to determine (a) the structure of the multivariate pdf underlying natural scenes, (b) nonlinear transforms to exploit this statistical structure, and (c) to what extent biological systems actually apply such transforms. Sparse coding, inD operators, gain control and their potential to deal with (b) are discussed. Moreover, higher-order statistics of natural images are analyzed with polyspectra.

In a later, again quite comprehensive, HVEI paper,[32] the authors further expand on many of the original ideas.[4] The focus is on the $\Sigma\Pi$ structures, the compensation equation, the AND operations on oriented filters, and the extension of these concepts to a theory of $i2D$ selectivity based on nonlinear Volterra-Wiener systems.[33] Possible combinations of oriented filters and the resulting tuning properties of the $i2D$ operators are analyzed in detail. Furthermore, the $i2D$ operators are related to higher-order statistical dependencies and the polyspectra of natural images.

Since the 1990 HVEI paper,[4] features that are based on the distribution of activity over directional channels, e.g., HOG and SIFT, have become quite popular. A later HVEI paper[34] extends work on Geometric Texton Theory which describes image structure as histograms over a visual feature vocabulary defined geometrically. The paper aims at defining basic image features by partitioning the filter response space of six Gaussian derivative filters.

Intrinsic dimension, $i2D$ detectors, and related features have been confined to nonlinear operations on oriented filters, and thus to cortical structures, until a further HVEI paper[35] showed that simple lateral inhibition, and

thus retinal structures, lead to $i2D$ selectivity when implemented in a deep or in a recurrent network with simple nonlinearities (ON/OFF rectification). A detailed retinal model was presented, which, a few years later, has been used to model the rather complex topological sensitivity of the bug detector in the frog's eye and human topological sensitivity attributed to early processing.[36] One important message was that cortical end-stopping could, in theory, stem from retinal processing. This is an interesting result, also since further evidence for nonlinear, $i2D$-specific processing has been found by using classification images in a vernier acuity task.[37]

The computational power of such linear-non-linear (LNL) structures has been further analyzed about ten years later[38] and it has been shown how LNL networks can reduce the statistical dependencies in natural images. The idea is to learn the linear part of the LNL sandwich, e.g., by PCA for decorrelation, then apply a simple nonlinearity that would introduce new correlations (since the nonlinearity can map higher-order dependencies to second-order dependencies), which can then be removed in the next linear stage. As in the early paper,[35] subtractive inhibition is extended to divisive inhibition and related to gain control and DN (see Sections 2.5 and 3.4). The LNL ideas have been meanwhile confirmed by the success of deep-learning strategies, where a number of layers are learned by unsupervised methods (often leading to sparse representations) and only the final layer is trained for making decisions.[39] However, a unified theoretical framework of LNL sandwiches and their relation to efficient representations, intrinsic dimension, sparse coding, and deep learning is still missing - see Section 2.4. Note, however, that the simple HVEI 1998 INROG (Iterated Nonlinear Ratio of Gaussians)[35] had it all: deep structure with increasingly sparse $i2D$ features and divisive normalization.

4. CONCLUSIONS

We have seen that, from a geometric perspective, the goal of efficient representation is to find the tiny submanifolds in which the natural images, or a particular set of images, lie. The statistical view is that one obtains more efficient representations the more redundancies are removed. From a practical perspective, a representation is efficient if it helps to solve a particular problem. Often, perceptual issues matter.

Sparse coding has been one of the major developments in this area and has convincingly shown how biologically-inspired research can have a major impact on the mathematical and technical sciences. In the beginning, people were telling us that sparse coding might be good for the brain because it saves sugar, but why should it be good for signal processing? Well, now they know. Other nonlinear extensions such as Gaussianization are also becoming popular in signal processing.

The concept of intrinsic dimension has similar origins and is related to sparse coding, as we have seen. People have used corner detectors before and have spent more bits on features that activate more than just one orientation channel, without knowing that such $i2D$ features are unique.[2] More explicit use of the framework has been made recently, for example, by Vig et al.,[5] who showed that a predictor based on $i3D$ features outperformed more complex, state-of-the-art saliency models. The inD saliency model was also useful for human activity recognition, which is still a very challenging real-world problem. Performance of several state-of-the-art action recognition algorithms was significantly improved by a preprocessing step that restricted action recognition to sparse saliency video regions only.[6]

There are many conferences that involve image representations. So, what makes HVEI special? We would like to mention only two aspects. First, HVEI has always been open to new ideas. Today, many conferences are proud of their high rejection rates, but this, in the end, may lead to insider events, which often do not last for 25 years. Second, while interdisciplinarity now is often used as a buzzword only, HVEI has been truly interdisciplinary from the beginning. If, as an imaging engineer, you are not really interested in visual perception, you can talk about SNR but not about image quality and lossy compression. Moreover, some of the innovative approaches to image representation, such as most of the ones reviewed here, came out of interdisciplinary studies of the visual system.

ACKNOWLEDGMENTS

M.D. was supported by NIH grants EY018664 and EY019281. E.V. was supported by the Postdoc-Programme of the German Academic Exchange Service (DAAD, D/11/41189).

REFERENCES

[1] "Special issue on adaptive sparse representation of data and applications in signal and image processing," *IEEE Transactions on Selected Topics in Signal Processing* **5**(5) (2011).

[2] Mota, C. and Barth, E., "On the uniqueness of curvature features," in [*Dynamische Perzeption*], Baratoff, G. and Neumann, H., eds., *Proceedings in Artificial Intelligence* **9**, 175–178, Infix Verlag, Köln (2000).

[3] Zetzsche, C. and Barth, E., "Fundamental limits of linear filters in the visual processing of two-dimensional signals," *Vision Research* **30**, 1111–1117 (1990).

[4] Zetzsche, C. and Barth, E., "Image surface predicates and the neural encoding of two-dimensional signal variation," in [*Human Vision and Electronic Imaging: Models, Methods, and Applications*], Rogowitz, B. E. and Allebach, J. P., eds., *Proc. SPIE* **1249**, 160–177 (1990).

[5] Vig, E., Dorr, M., Martinetz, T., and Barth, E., "Intrinsic dimensionality predicts the saliency of natural dynamic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(6), 1080–1091 (2012).

[6] Vig, E., Dorr, M., and Cox, D., "Space-variant descriptor sampling for action recognition based on saliency and eye movements," in [*LNCS 7578, Proceedings of the European Conference on Computer Vision*], 84–97, Springer, Springer, Firenze, Italy (2012).

[7] Zetzsche, C., Barth, E., and Wegmann, B., "The importance of intrinsically two-dimensional image features in biological vision and picture coding," in [*Digital Images and Human Vision*], Watson, A. B., ed., 109–38, MIT Press (Oct. 1993).

[8] Olshausen, B., Cadieu, C., and Warland, D., "Learning real and complex overcomplete representations from the statistics of natural images," in [*Wavelets XIII*], V.K. Goyal, M. Papadakis, D. v. d. V., ed., *Proc. SPIE* **7446** (2009).

[9] Labusch, K., Barth, E., and Martinetz, T., "Sparse Coding Neural Gas: Learning of Overcomplete Data Representations," *Neurocomputing* **72**(7-9), 1547–1555 (2009).

[10] Lyu, S. and Simoncelli, E., "Nonlinear extraction of independent components of natural images using radial gaussianization," *Neural Comput.* **21**(6), 1485–1519 (2009).

[11] Chen, J. and Pappas, T. N., "Perceptual coders and perceptual metrics," in [*Human Vision and Electronic Imaging VI*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **4299**, 150–162 (2001).

[12] Cheng, H. and Lubin, J., "Reference-free objective quality metrics for mpeg-coded video," in [*Human Vision and Electronic Imaging X*], Rogowitz, B. E., Pappas, T. N., and Daly, S. J., eds., *Proc. SPIE* **5666**, 160–167 (2005).

[13] Watson, A. B., "Efficiency of a model human image code," *J Opt Soc Am* **4**(12), 2401–17 (1987).

[14] Daly, S. J., Feng, X., and Speigle, J. M., "Practical applications that require some of the more advanced features of current visual models," in [*Human Vision and Electronic Imaging VII*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **4662**, 70–83 (2002).

[15] Buccigrossi, R. W. and Simoncelli, E. P., "Image compression via joint statistical characterization in the wavelet domain," *IEEE Transactions on Image Processing* **8**, 1688–1701 (1999).

[16] Wang, Z. and Simoncelli, E. P., "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in [*Human Vision and Electronic Imaging X*], Rogowitz, B. E., Pappas, T. N., and Daly, S. J., eds., *Proc. SPIE* **5666**, 149–159 (2005).

[17] Wang, Z. and Simoncelli, E. P., "Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics," in [*Human Vision and Electronic Imaging IX*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **5292**, 99–108 (2004).

[18] Moorthy, A. K. and Bovik, A. C., "Perceptually significant spatial pooling techniques for image quality assessment," in [*Human Vision and Electronic Imaging XIV*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **7240**, 724012–724012–11 (2009).

[19] Rajashekar, U., Wang, Z., and Simoncelli, E. P., "Perceptual quality assessment of color images using adaptive signal representation," in [*Human Vision and Electronic Imaging XV*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **7527**, 75271L–75271L–9 (2010).

[20] Field, D. J., Olshausen, B. A., and Brady, N., "Wavelets, blur, and the sources of variability in the amplitude spectra of natural scenes," in [*Human Vision and Electronic Imaging*], Rogowitz, B. E. and Allebach, J. P., eds., *Proc. SPIE* **2657**, 108–119 (1996).

[21] Olshausen, B. A. and Field, D. J., "Learning efficient linear codes for natural images: the roles of sparseness, overcompleteness, and statistical independence," in [*Human Vision and Electronic Imaging*], Rogowitz, B. E. and Allebach, J. P., eds., *Proc. SPIE* **2657**, 132–138 (1996).
[22] Olshausen, B. A. and Field, D. J., "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature* **381**, 607–609 (1996).
[23] Olshausen, B. A., Cadieu, C., Culpepper, J., and Warland, D. K., "Bilinear models of natural images," in [*Human Vision and Electronic Imaging XII*], Rogowitz, B. E., Pappas, T. N., and Daly, S. J., eds., *Proc. SPIE* **6492**, 649206–649206–10 (2007).
[24] Labusch, K., Siewert, U., Martinetz, T., and Barth, E., "Learning optimal features for visual pattern recognition," in [*Human Vision and Electronic Imaging XII*], Rogowitz, B. E., Pappas, T. N., and Daly, S. J., eds., *Proc. SPIE* **6492**, 64920B–64920B–8 (2007).
[25] Luo, T. and Mou, X., "Divisive normalization in channelized hotelling observer," in [*Human Vision and Electronic Imaging XV*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **7527**, 75271K–75271K–10 (2010).
[26] Lyu, S. and Simoncelli, E. P., "Statistically and perceptually motivated nonlinear image representation," in [*Human Vision and Electronic Imaging XII*], Rogowitz, B. E., Pappas, T. N., and Daly, S. J., eds., *Proc. SPIE* **6492**, 649207–649207–15 (2007).
[27] Valerio, R. and Navarro, R., "Nonlinear image representation with statistical independent features: efficient implementation and applications," in [*Human Vision and Electronic Imaging VIII*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **5007**, 352–363 (2003).
[28] Bethge, M., Wiecki, T. V., and Wichmann, F. A., "The independent components of natural images are perceptually dependent," in [*Human Vision and Electronic Imaging XII*], Rogowitz, B. E., Pappas, T. N., and Daly, S. J., eds., *Proc. SPIE* **6492**, 64920A–64920A–12 (2007).
[29] Zetzsche, C. and Krieger, G., "Nonlinear neurons and higher-order statistics: new approaches to human vision and digital image processing," in [*Human Vision and Electronic Imaging IV*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE*, 2–33 (1999).
[30] Bethge, M., Gerwinn, S., and Macke, J. H., "Unsupervised learning of a steerable basis for invariant image representations," in [*Human Vision and Electronic Imaging XII*], Rogowitz, B. E., Pappas, T. N., and Daly, S. J., eds., *Proc. SPIE* **6492**, 64920C–64920C–12 (2007).
[31] Li, Y. and Adelson, E., "Image mapping using local and global statistics," in [*Human Vision and Electronic Imaging XIII*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **6806**, 680614–680614–11 (2008).
[32] Zetzsche, C., Krieger, G., and Mayer, G., "Nonlinear AND interactions between frequency components and the selective processing of intrinsically two-dimensional signals by cortical neurons," in [*Human Vision and Electronic Imaging VI*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **4299**, 36–68 (2001).
[33] Krieger, G., Zetzsche, C., and Barth, E., "Nonlinear image operators for the detection of local intrinsic dimensionality," in [*Proc. IEEE Workshop Nonlinear Signal and Image Processing*], 182–185 (1995).
[34] Griffin, L. D. and Lillholm, M., "Feature category systems for 2nd order local image structure induced by natural image statistics and otherwise," in [*Human Vision and Electronic Imaging XII*], Rogowitz, B. E., Pappas, T. N., and Daly, S. J., eds., *Proc. SPIE* **6492**, 649209–649209–11 (2007).
[35] Barth, E. and Zetzsche, C., "Endstopped operators based on iterated nonlinear center-surround inhibition," in [*Human Vision and Electronic Imaging III*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **3299**, 67–78 (1998).
[36] Barth, E., Ferraro, M., and Zetsche, C., "Global topological properties of images derived from local curvature features," in [*Visual Form 2001. Lecture Notes in Computer Science*], Arcelli, C., Cordella, L. P., and di Baja, G. S., eds., 285–294 (2001).
[37] Barth, E., Beard, B. L., and Ahumada, Jr, A. J., "Nonlinear features in vernier acuity," in [*Human Vision and Electronic Imaging IV*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **3644**(8), 88–96 (1999).
[38] Zetzsche, C. and Nuding, U., "Nonlinear encoding in multilayer LNL systems optimized for the representation of natural images," in [*Human Vision and Electronic Imaging XII*], Rogowitz, B. E., Pappas, T. N., and Daly, S. J., eds., *Proc. SPIE* **6492**, 649204–649204–22 (2007).
[39] Bengio, Y., "Learning deep architectures for AI," *Foundations and Trends in Machine Learning* **2**(1), 1–127 (2009). Also published as a book. Now Publishers, 2009.