MICA: A Multilinear ICA Decomposition for Natural Scene Modeling

Raghu G. Raj, *Student Member, IEEE*, and Alan C. Bovik, *Fellow, IEEE*

***Abstract*—We refine the classical independent component analysis (ICA) decomposition using a multilinear expansion of the probability density function of the source statistics. In particular, we introduce a specific nonlinear system that allows us to elegantly capture the statistical dependences between the responses of the multilinear ICA (MICA) filters. The resulting multilinear probability density is analytically tractable and does not require Monte Carlo simulations to estimate the model parameters. We demonstrate the MICA model on natural image textures and envision that the new model will prove useful for analyzing nonstationarity natural images using natural scene statistics models.**

***Index Terms*—Independent components, multilinear independent component analysis (MICA), natural scene statistics (NSS), nonlinear modeling.**

I. Introduction

THE construction of accurate prior models of natural image source data is essential to many applications, such as low-level vision, for which unsupervised learning methods must be applied due to the inherent lack of labeled training sets. Such prior models give a framework in which to correctly interpret the data, thereby serving as the basis for subsequent analysis viewed from different levels of abstraction. There are a variety of classical unsupervised methods that exist for this purpose, including principle component analysis (PCA), independent component analysis (ICA), and multidimensional scaling (MDS) [1].

Among these classic tools, ICA has several important and distinguishing characteristics. Denote the probability of the source that we are modeling by $P(X)$, where X is a random vector whose realizations have dimensionality d. The goal of ICA is to factor the probability distribution of the source into a product of distributions: $P(X) = \prod_{i=1}^{d} p(z_i)$, where $\{z_i = X * \phi_i\}_{i=1}^{d}$ are filtered responses of the source. The filters $\{\phi_i\}_{i=1}^{d}$ are the ICA filters of the source. Statistical algorithms for computing the ICA filters have been the subject of intense study over the past decade [2], most of which involve the construction of different cost functions (usually variations of the maximum likelihood cost function).

The independent directions that emerge from an ICA decomposition can be fruitfully utilized by reducing the $d-$ dimensional problem into d independent 1-D problems. Furthermore, ICA decompositions of data having heavy tailed marginals (as is for example observed in NSS applications) tend to favor sparse representations [12]. Sparse representations are useful for many applications that seek to efficiently represent and process the data.

However, in spite of these potential advantages, in reality, the statistics of most real-world sources, such as natural image patches, cannot be strictly factored into a simple product. As a result, the so-called independent components contain significant mutual dependencies between them [3]. Accordingly, prior work has attempted to more completely capture statistical image structure by accounting for the dependences (either directly or indirectly) between the ICA components [3], [15], [16].

In this paper, we approach this problem from the perspective of refining the classic ICA model such that the dependencies between pseudo-independent components are captured using a multilinear representation of $P(X)$

$$P(X) = \frac{1}{Z} g(J) \prod_{i=1}^{d} p(z_i)$$

where $g : J = [z_1, \ldots, z_d] \to R$ and $Z \in R$ is a normalizing constant. We call the resulting model the *multilinear ICA (MICA) decomposition* of the distribution $P(X)$. Of all possible multilinear expansions of this form that could describe the source distribution, we seek the one that makes the representation of the source as sparse as possible, i.e., which minimizes the contribution of $g(J)$. Naturally we are interested in closed form approximations for such a $g(J)$. The multilinear form thus obtained retains all the attractive properties of the ICA decomposition, and at the same time lumps the interactions of the filtered responses into the function $g(J)$. Of course, when $g(J)$ is separable with respect to the filter responses, this reduces to the classical ICA representation.

The success of our proposed method depends upon the accuracy of the numerical approximation of $g(J)$. Analytical methods of approximating $g(J)$ using Taylor expansions seem formidable. Further it is necessary to estimate Z which, in general, requires tedious Monte Carlo simulations.

In Section II, we introduce a nonlinear system model that enables us to circumvent the above issues. We call the resulting refinement of ICA the *Multilinear ICA (MICA) Model*. We successfully deploy the new method to model natural scene textures in Section III, and demonstrate advantages relative to classical ICA. We conclude in Section IV with a discussion of possible

Manuscript received August 8, 2007; revised December 3, 2007. This work was supported by the National Science Foundation under Grant 0427372. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lina J. Karam.

The authors are with the Center for Perceptual Systems, Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA (e-mail: rraj@ece.utexas.edu; raghu.g.raj@gmail.com; bovik@ece.utexas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2007.916158

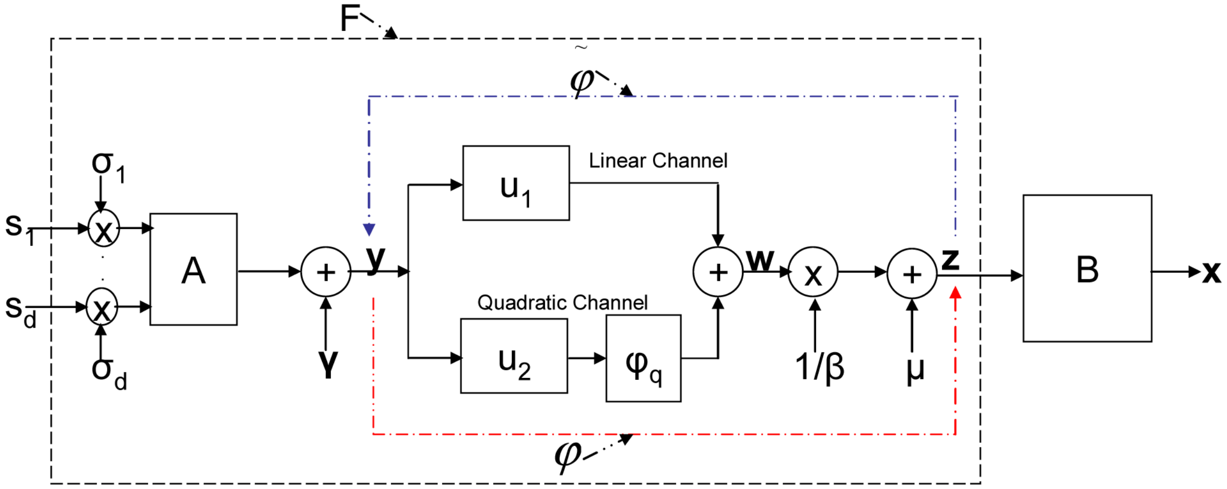


Fig. 1. Nonlinear system model of the multilinear structure of source statistics derived from natural scene models.

applications of the MICA model together with some open problems.

II. Multilinear MICA Model

A. *Overview and Parameter Description*

Consider the classical ICA model where the observation vector is modeled: $x = Bz$, where $x = [x_1, \ldots, x_d]^T \in R^d, z = [z_1, \ldots, z_d]^T \in R^d, d$ is the intrinsic dimensionality of the data, and $B \in R^{dxd}$ is a full-rank matrix. The goal of ICA is to find a matrix B such that the resulting components of z are independent random variables.

However, for many real-world sources, such as natural images, such an ideal decomposition is not possible and so the components of z will contain residual dependencies. Our aim is to explicitly capture these dependencies. In doing so, we must first recognize that z cannot be further decomposed as a combination of independent sources via another full-rank matrix! It is possible, however, that z can be decomposed with respect to an under-complete linear model, but this requires knowledge of the subspace dimensionality.

An alternate view which we explore in this paper is that, given knowledge of the intrinsic dimensionality d, the residual dependencies can be captured via nonlinear combinations of independent sources. The choice of the nonlinearity, as well as of the source distribution, must be as simple as possible, and yet must successfully account for the probabilistic structure of the observed natural image sources. To simplify matters further, we first concern ourselves only with modeling unimodal distributions which, as shown in Section III, appears to be well-suited to many natural image textures. Later on, we suggest how to extend this to multimodal cases via mixtures of MICA models. We first focus on the complete basis case (i.e., where B is a full-rank matrix). Later, we will demonstrate how these ideas can be extended to the under-complete case in a straightforward manner.

Perhaps the simplest nonlinear system that one can hypothesize for natural image source modeling is a quadratic channel. In our experiments with natural image textures, we found that the hybrid linear-quadratic model (stimulated by i.i.d. Gaussian sources) shown in Fig. 1 can successfully account for the probabilistic structure of natural image patches. We now describe this nonlinear system in detail.

The observable image source data that we are modeling is $x \in R^d$. $B \in R^{dxd}$ is a full-rank matrix initially chosen as the matrix associated with the classical ICA decomposition of x which will be re-estimated in subsequent iterations. The system F in Fig. 1 models the residual interaction between the components of $z \in R^d$. It consists of a core nonlinearity φ preceded by a linear system $y = As + \gamma$, where $y = [y_1, \ldots, y_d]^T \in R^d, \gamma = [\gamma_1, \ldots, \gamma_d]^T \in R^d$, and $s = [s_1, \ldots, s_d]^T \in R^d$ are *i.i.d.* Gaussian: $s_i \sim \aleph(0,1)$. The density of the ith Gaussian channel is denoted $q(s_i)$. The Gaussian channel variances are $\sigma = [\sigma_1, \ldots, \sigma_d]^T \in R^d, \mu = [\mu_1, \ldots, \mu_d]^T \in R^d$ is an additive mean adjusting vector, and $\beta = [\beta_1, \ldots, \beta_d] \in R^d$ is a multiplicative vector that is applied (component-wise) to all channels, and which determines the effective nonlinearity of the channels. Finally, $C = [C_{i,j}] = [C_1^T, \ldots, C_d^T] = A^{-1} \in R^{dxd}$ is an invertible linear transformation of the *i.i.d.* Gaussian sources that determines the interaction of the Gaussian sources.

B. *Structure of the MICA Distribution*

The nonlinearity φ consists of complementary linear and quadratic channels. Operators u_1 and u_2 are *complementary limiters*: $u_1(y_i) + u_2(y_i) = 1, u_1(y_i), u_2(y_i) \geqslant 0$ for $1 \leq i \leq d$. A simple choice of limiters which we have found to be useful for modeling natural image textures (see Section III), are the complementary step functions

$$u_1(y_i) = u(y_i + 1) - u(y_i - 1), u_2(y_i) = 1 - u_1(y_i)$$

where $u(x)$ is the unit step function.

From this, we obtain

$$\varphi(y) = yu_1(y) + \varphi_q(y)u_2(y)$$

where $\varphi_q(y) = y^2\mathrm{sgn}(y)$ (throughout, operations on y are applied component-wise). The function ϕ is plotted in Fig. 4.

For this choice of (u_1, u_2)

$$\begin{aligned}\tilde{\varphi}&[\beta(z-\mu)]\\ &\equiv \varphi^{-1}[\beta(z-\mu)]\\ &= \beta(z-\mu)u_1[\beta(z-\mu)]\\ &\quad + \varphi_q^{-1}[\beta(z-\mu)]\, u_2[\beta(z-\mu)]\end{aligned}$$

where

$$\varphi_q^{-1}[\beta(z-\mu)] = \{\sqrt{|\beta_i(z_i-\mu_i)|}\mathrm{sgn}[\beta_i(z_i-\mu_i)]\}_{i=1}^{d}.$$

Since the nonlinearity is invertible, system F is also: $s = F^{-1}[(z-\mu)] = C\{\tilde{\varphi}[\beta(z-\mu)] - \gamma\}$.

The distribution of $\tilde{s}$ then has the following form (where throughout $|\cdot|$ is the matrix determinant):

$$\begin{aligned}P(x) &= \frac{1}{|B|}p(z)\\ &= \frac{1}{|B|}\cdot\frac{1}{|J(F)|}\prod\nolimits_{k=1}^{d} q\left\{F_k^{-1}[\beta(z-\mu)]\right\}\end{aligned}$$

where $z = B^{-1}x$ and $q(s_k)$ is the kth Gaussian source channel. Expanding $p(z)$ yields the MICA model

$$p(z) = \frac{K}{|J(F)|}g(J)\prod\nolimits_{k=1}^{d} p((z_i-\mu_i)\beta_i) \qquad (1)$$

where

$$p[(z_i-\mu_i)\beta_i] = K_i\exp(-a_i[\tilde{\varphi}[\beta_i(z_i-\mu_i)] - c_i]^2).$$

K_i is a normalizing constant, $a_i = (1/2)\sum_{k=1}^{d}(C_{k,i}^2/\sigma_k^2)$, and

$$c_i = \gamma_i + \frac{\sum_{j\neq i}\sum_{k=1}^{d}\frac{C_{k,j}C_{k,i}}{\sigma_k^2}\gamma_j}{\sum_{k=1}^{d}\frac{C_{k,i}^2}{\sigma_k^2}}.$$

Also

$$g(J) = \exp[-\sum\nolimits_{i\neq j} G_{i,j}\varphi(\beta_i(z_i-\mu_i))\varphi(\beta_j(z_j-\mu_j))]$$

where

$$K = \frac{\exp\left[-\sum_{k=1}^{d}\frac{(C_k^T\gamma)^2}{2\sigma_k^2}\right]}{(2\pi)^{d/2}\prod_{k=1}^{d}\sigma_k K_k}$$

and

$$G_{i,j} = \sum\nolimits_{k=1}^{d}\frac{C_{k,i}C_{k,j}}{\sigma_k^2}.$$

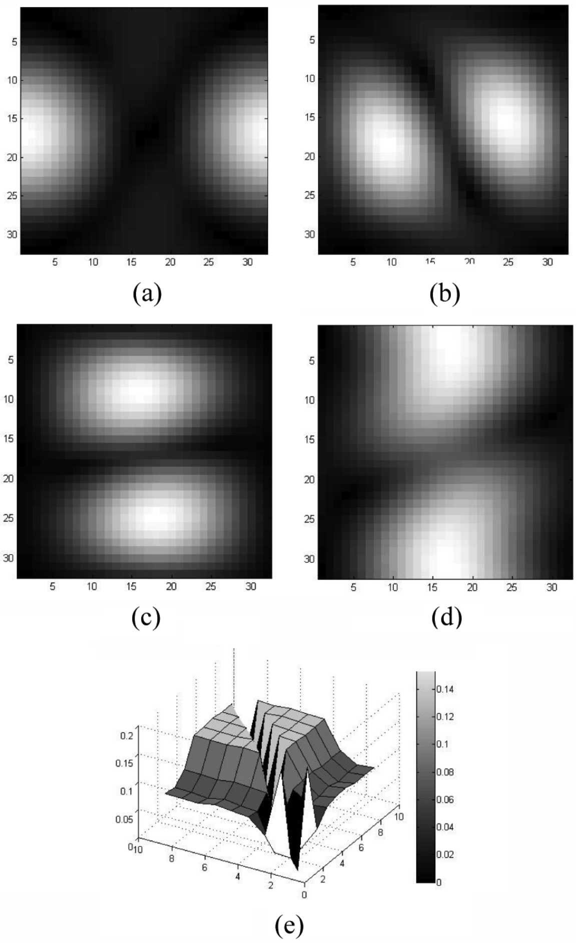


Fig. 2. (a)–(d) Examples of frequency responses of MICA filters corresponding to the Gravel texture; (e) magnitude $|G|$ of the MICA interaction matrix for the Gravel texture. The larger the magnitude of $G_{i,j}$ the greater the statistical dependency between the corresponding MICA components.

In (1), $J(F)$ is the Jacobian of the transformation F, for which the following theorem (proved in the Appendix) yields a closed form expression.

Theorem 1: The Jacobian of the transformation F is

$$J(F) = \frac{1}{|C|}\prod\nolimits_{k=1}^{d}\frac{\psi\left[\beta_k(z_k-\mu_k)\right]}{\beta_k}$$

where

$$\begin{aligned}\psi[\beta_k(z_k-\mu_k)] &= u_1[\beta(z_k-\mu_k)]\\ &\quad +2|\varphi_q^{-1}[\beta_k(z_k-\mu_k)]|u_2[\beta_k(z_k-\mu_k)].\end{aligned}$$

♣

The *MICA interaction matrix* $[G_{i,j}]$ captures interactions between the MICA components. In particular, when $[G_{i,j}]_{i\neq j} = 0$, the MICA components are independent. Fig. 2(a)–(d) shows the frequency response of a few of the MICA filters (derived from the matrix B in Fig. 1) of the Gravel texture as described in more detail in Section III. These MICA filters exhibit bandpass like behavior and generally, there will be overlap among the spectra between the various MICA components. The overlapping of spectra, however, does not by itself indicate the degree of dependence between the MICA filters. The latter is captured more accurately by the MICA interaction matrix which, for the Gravel texture, is shown in Fig. 2(e). In particular, the greater the value of $G_{i,j}$, the greater the degree of statistical dependency between the corresponding MICA filters.

The parameter β determines the degree of nonlinearity in the system which can be qualitatively understood as follows. When training the MICA model (given the filtered data z), β determines the extent to which w is scaled inside the unit interval and consequently determines (after σ is adjusted as a part of the MICA optimization) the extent to which the linear channel of the system is active. Thus, β determines the tradeoff between the linear and quadratic models when determining the optimal tail and peak behaviors of the MICA distribution. Once the above parameters have been adjusted, the vector μ is chosen to optimally adjust the mean of MICA. Finally, γ determines the skew of the marginal distributions by asymmetrically assigning the nonlinearity within the effective domain of the distribution.

C. Parameter Estimation for the MICA Model

We estimate the optimal parameters of the MICA model (1) by employing a steepest gradient algorithm with respect to the log-likelihood function

$$\begin{aligned}\log[p(z)] = {} & \log\left(\frac{\mathrm{abs}\,(|C|)}{(2\pi)^{d/2}}\prod_{k=1}^{d}\frac{\beta_k}{\sigma_k}\right) \\ & - \log\left(\prod_{i=1}^{d}|\psi(\beta_i(z_i-\mu_i))|\right) - \sum_{k=1}^{d}\frac{\left(C_k^T\gamma\right)^2}{2\sigma_k^2} \\ & - \frac{1}{2}\sum_{i=1}^{d}\left(\sum_{k=1}^{d}\frac{C_{k,i}^2}{\sigma_k^2}\right)\{\tilde{\varphi}\,[\beta_i(z_i-\mu_i)]\}^2 \\ & + \sum_{i=1}^{d}\left(\sum_{k=1}^{d}\frac{C_k^T}{\sigma_k^2}C_{k,i}\right)\widetilde{\varphi}[\beta(z_i-\mu_i)] \\ & - \sum_{i\neq j}G_{i,j}\tilde{\phi}[\beta(z_i-\mu_i)]\tilde{\phi}[\beta_j(z_j-\mu_j)]. \end{aligned} \qquad (2)$$

From (2), the gradient of the log-likelihood function with respect to the different parameters can be computed in a straightforward manner

$$\begin{aligned}\frac{\partial\log[p(z)]}{\partial C_{m,n}} & \\ = {} & \frac{(-1)^{m+n}|C^{m,n}|}{\mathrm{abs}(|C|)}\mathrm{sgn}|C| - \left(\frac{C_m^T\gamma}{\sigma_m^2}\right)\gamma_n \\ & - \frac{C_{m,n}}{\sigma_m^2}\{\tilde{\varphi}[\beta_n(z_n-\mu_n)]\}^2 \\ & + \tilde{\varphi}[\beta_n(z_n-\mu_n)]\left[\frac{C_m^T\gamma}{\sigma_m^2}+\frac{C_{m,n}\gamma_n}{\sigma_m^2}\right] \\ & - \sum_{i\neq n}\frac{C_{m,i}}{\sigma_m^2}\tilde{\varphi}[\beta_i(z_i-\mu_i)]\tilde{\varphi}[\beta_n(z_n-\mu_n)] \end{aligned} \qquad (3)$$

$$\begin{aligned}\frac{\partial\log[p(z)]}{\partial\sigma_m} & \\ = {} & -\frac{1}{\sigma_m}+\frac{\left(C_m^T\gamma\right)^2}{\sigma_m^3} \\ & + \frac{1}{\sigma_m^3}\sum_{i=1}^{d}C_{m,i}^2\{\tilde{\varphi}\,[\beta_i(z_i-\mu_i)]\}^2 \\ & - \frac{2}{\sigma_m^3}\sum_{i=1}^{d}\left(C_m^T\gamma\right)C_{m,i}\tilde{\varphi}[\beta_i(z_i-\mu_i)] \\ & + \frac{2}{\sigma_m^3}\textstyle\sum_{i\neq j}C_{m,i}C_{m,j}\tilde{\varphi}[\beta_i(z_i-\mu_i)]\tilde{\varphi}[\beta_j(z_j-\mu_j)] \end{aligned} \qquad (4)$$

$$\begin{aligned}\frac{\partial\log(p(z))}{\partial\gamma_m} & \\ = {} & -\sum_{k=1}^{d}\frac{C_k^T\gamma}{\sigma_k^2}C_{k,m} \\ & + \sum_{i=1}^{d}\tilde{\varphi}[\beta_i(z_i-\mu_i)]\left(\sum_{k=1}^{d}\frac{C_{k,m}C_{k,i}}{\sigma_k^2}\right) \end{aligned} \qquad (5)$$

$$\begin{aligned}\frac{\partial\log[p(z)]}{\partial\mu_m} = {} & \frac{\beta_m\hat{u}_2(\beta_m(z_m-\mu_m))}{\varphi_q^{-1}(\beta_m(z_m-\mu_m))}\frac{1}{|\psi(\beta_m(z_m-\mu_m))|} \\ & +\beta_m\left[u_1[\beta_m(z_m-u_m)]+\frac{u_2[\beta_m(z_m-u_m)]}{2\sqrt{|\beta_m(z_m-u_m)|}}\right] \\ & \begin{bmatrix}\sum_{k=1}^{d}\frac{C_{k,m}^2}{\sigma_k^2}\tilde{\varphi}(\beta_m(z_m-\mu_m)) \\ -\sum_{k=1}^{d}\frac{C_k^T\gamma}{\sigma_k^2}C_{k,m} \\ +\sum_{i\neq m}\left(\sum_{k=1}^{d}\frac{C_{k,i}C_{k,m}}{\sigma_k^2}\right)\tilde{\varphi}(\beta_i(z_i-\mu_i))\end{bmatrix} \end{aligned} \qquad (6)$$

where $\hat{u}_2(z)=u_2(z)+[\delta(z-1)-\delta(z+1)]$ such that δ is the impulse function, and

$$\begin{aligned}\frac{\partial\log[p(z)]}{\partial\beta_m} = {} & \frac{1}{\beta_i}-\frac{\hat{u}_2(\beta_m(z_m-\mu_m))}{2{\varphi_q}^{-1}[\beta_m(z_m-\mu_m)]}\frac{1}{|\psi\,[\beta_m(z_m-\mu_m)]|} \\ & -\left[u_1[(\beta_m(z_m-\mu_m)]+\frac{u_2[\beta_m(z_m-\mu_m)]}{2\sqrt{|\beta_m(z_m-\mu_m)|}}\right] \\ & \begin{bmatrix}+\tilde{\varphi}[\beta_m(z_m-\mu_m)]\sum_{k=1}^{d}\frac{C_{k,m}^2}{\sigma_k^2} \\ -\sum_{k=1}^{d}\frac{C_k^T\gamma}{\sigma_k^2}C_{k,m} \\ +\sum_{i\neq m}G_{i,m}\tilde{\varphi}[\beta_i(z_i-\mu_i)]\end{bmatrix}(z_m-\mu_m) \end{aligned} \qquad (7)$$

where $C^{m,n}$ is the co-factor matrix of C with respect to (m,n), and $C_{m,n}$ is the (m,n)th entry of C.

Our goal is to obtain a multilinear expansion of $P(\tilde{s})$ corresponding to a sparse representation of the source. This can be accomplished by initializing B with the matrix associated with the classical ICA decomposition of source x. Then $z=B^{-1}x$. A gradient descent algorithm then obtains the optimum parameters

High-level MICA Algorithm:

(1) Acquire data samples *x*
(2) Compute B-matrix in Fig. 1 (Initialization using Comon's algorithm [5], Gabor bases etc.; updation using Eq. (8))
(3) Compute the optimal MICA parameters using Eqs. (2)-(7)
(4) Repeat steps (2)-(3) as needed
(5) The optimal MICA parameters thus computed furnish the Multilinear model in Eq. 1

Fig. 3. High-level MICA algorithm.

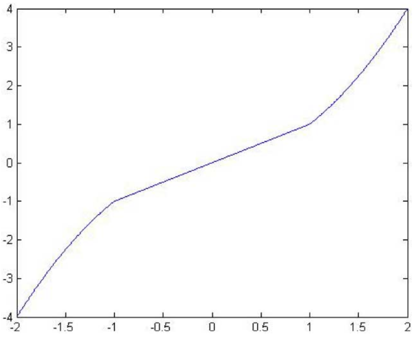


Fig. 4. Function ϕ: Linear in the unit interval and quadratic outside.

$C = A^{-1}, \sigma, \gamma$, μ and β using the above expressions. A multilinear expansion of $P(x)$ is obtained as in (1), the structure of which is specified by these parameters. The estimate of B can be further refined by fixing the parameters, then invoking a gradient descent algorithm. Let $D = B^{-1}$; then $P(x) = \mathrm{abs}(|D|)p(z)$. The gradient of $\log[P(x)]$ with respect to $D_{m,n}$ (the (m,n)th entry of D) is shown in (8) at the bottom of the page, where $\hat{u}_2(z) = u_2(z) + [\delta(z-1) - \delta(z+1)]$ such that δ is the impulse function.

Once B is computed, the two-step process of estimating $(C, \sigma, \gamma, \mu, \beta)$ followed by re-estimating B may be performed until a desired level of accuracy is achieved. However, in our simulations, we find that a single estimate of $(C, \sigma, \gamma, \mu, \beta)$, without subsequent re-estimation of B generally outperforms classical ICA modeling on natural image textures (using the KLD as a measure of performance) as shown in Section III below. The high-level MICA algorithm thus described is summarized in Fig. 3.

We have found it convenient to heuristically estimate β instead of employing (7). Our heuristic for estimating β is described and motivated as follows. Consider the case where the distribution of the ith data channel is heavy-tailed (high-kurtosis). Modeling the ith channel histogram by a Laplacian distribution

$$f(z|b_i) = \frac{1}{2b_i} \exp\left(-\frac{|z - \mu_i|}{b_i}\right)$$

the parameter b_i can be estimated from the data using the following closed form expression [8]

$$\hat{b}_i = \frac{1}{N} \sum\nolimits_{j=1}^{N} |z_j^i - \hat{\mu}_i|$$

where $\hat{\mu}_i$ is the sample median of the ith channel data $\{z_j^i\}_{j=1}^N$. From (1), β_i can be thought of as proportional to $1/b_i$. This yields a heuristic for the initial estimate of β_i: $\beta_i^0 = 1/(\hat{b}_i)$. Further refinements can be obtained in subsequent iterations by observing that in (1), a more accurate relationship between the b_i and β_i is as follows: $\beta_i \approx 1/(\hat{b}_i a_i)$. The kth estimate of β_i is $\beta_i^k \approx 1/(\hat{b}_i a_i^{k-1})$, where a_i^{k-1} is the $(k-1)^{\mathrm{st}}$ estimate of a_i obtained when using β_i^{k-1}. As shown in Section III, the initial estimate $\beta_{high-kurt} = \beta_i^0$ yields better performance than ICA, even without subsequent re-estimation of β_i.

For the case where the ith data channel is not heavy tailed (i.e., the low-kurtosis case), it is intuitive to emphasize the linear part of $\tilde{\varphi}$—tantamount to initializing β_i such that $\beta_i \leqslant (1/\max)_n[z_i(n)]$. In practice, $\beta_{\mathrm{low-kurt}} = \beta_i^0 = (1/\max_n[z_i(n)])$ suffices for low-kurtosis cases. As with the high kurtosis case, we find it unnecessary to update the estimate of β_i at every iteration, but instead use the initial estimate β_i^0 throughout the optimization process.

To simplify matters further, we employ a single scalar parameter β that we apply to *all* channels. To estimate β we employ a similar heuristic as above. For high-kurtosis, use $\beta_{\mathrm{high-kurt}} = 1/\hat{b}$ where $\hat{b} = (1/Nd)\sum_{i,j} |z_j^i - \hat{m}_i|$ and $\hat{m}_i$ is the sample median of $\{z_j^i\}_{i,j}$. Similarly, for low kurtosis take $\beta_{\mathrm{low-kurt}} = (1/\max_{i,j}[z_i(j)])$.

As a simple way of deciding the β estimate to be used, we measure the local sample kurtosis κ. The high-kurtosis heuristic is used when $\kappa \geq 4$, and the low-kurtosis heuristic when $\kappa \leq 4$.

$$\begin{aligned}\frac{\partial \log[P(x)]}{\partial D_{m,n}} &= \frac{(-1)^{m+n}|D^{m,n}|sign(|D|)}{abs(|D|)} - \frac{\beta_m \hat{u}_2[\beta_m(z_m - \mu_m)]x_n}{2|\Psi[\beta_m(z_m - \mu_m)]|\varphi_q^{-1}[\beta_m(z_m - \mu_m)]} \\ &\quad -\beta_m\left[u_1[\beta_m(z_m - \mu_m)] + \frac{u_2[\beta_m(z_m - \mu_m)]}{2\sqrt{|\beta_m(z_m - \mu_m)|}}\right] \begin{bmatrix} +\tilde{\varphi}(\beta_m(z_m - \mu_m))\sum_{k=1}^{d} \frac{C_{k,m}^2}{\sigma_k^2} \\ -\sum_{k=1}^{d} \frac{C_k^T \gamma}{\sigma_k^2} C_{k,m} \\ +\sum_{j \neq m} \left(\sum_{k=1}^{d} \frac{C_{k,i} C_{k,m}}{\sigma_k^2}\right) \tilde{\varphi}(\beta_i(z_i - \mu_i)) \end{bmatrix} x_n \qquad (8)\end{aligned}$$

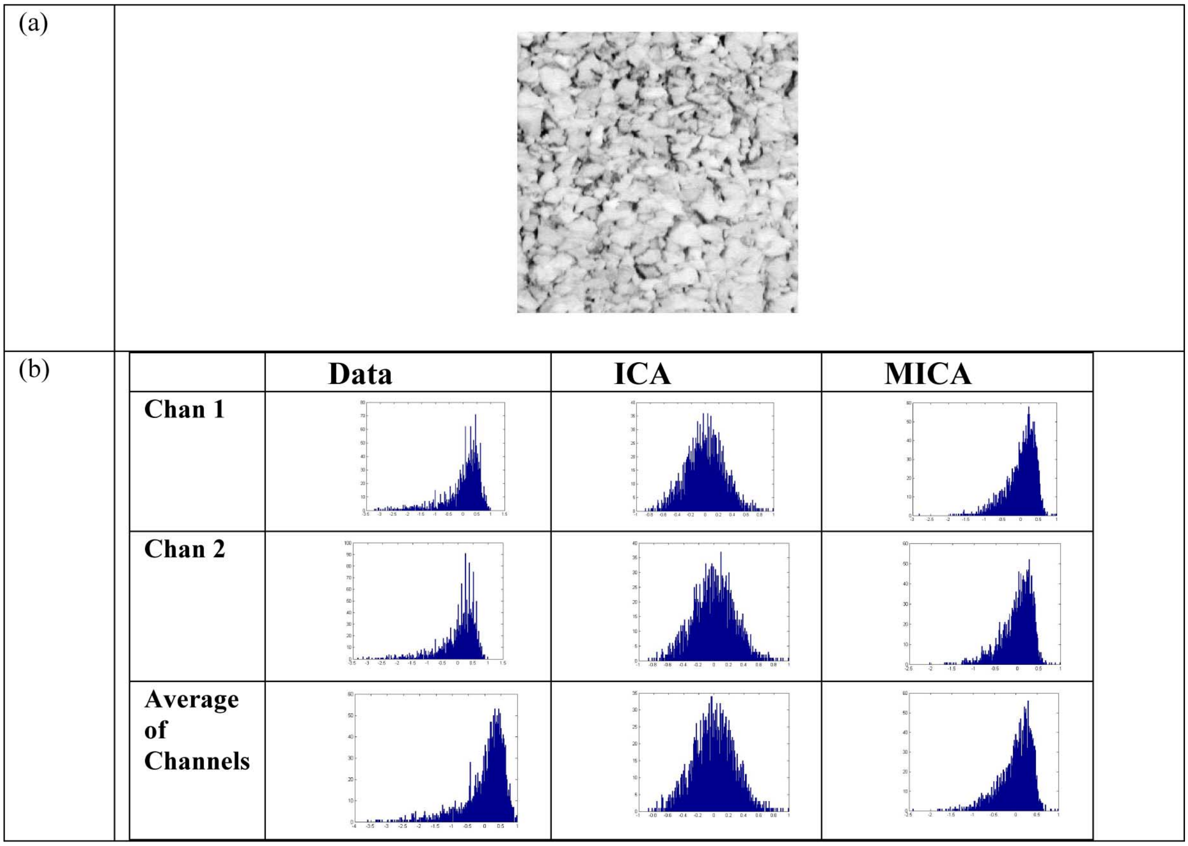


Fig. 5. (a) Gravel. (b) Channel histograms of channels and their corresponding ICA and MICA distributions. The high-kurtosis heuristic $\beta_{\text{high-kurt}}$ was used.

D. Extension to Under-Complete MICA Models

Consider the case where the observation x is modeled as follows: $x = Bz$, where $x = [x_1, \ldots, x_l]^T \in R^l$ and $z = [z_1, \ldots, z_d]^T \in R^d$, such that $d < l$, i.e., $B \in R^{lxd}$ is an under-complete matrix. Under-complete models arise in situations where dimensionality reduction is required in order to model the data in an appropriate subspace.

As before, we ask whether multilinear modeling can accurately capture the statistical dependencies between the components of z. As shown in Section III, the answer to this is affirmative. A simple way of assessing the performance of MICA for under-complete models is to first consider the corresponding complete basis case where $x = B^{\sim} z^{\sim}, x = [x_1, \ldots, x_l]^T \in R^l$ is the observed source vector as before, $\tilde{B} \in R^{lxl}$ and $\tilde{z} = [\tilde{z}_1, \ldots, \tilde{z}_l]^T \in R^l$. The matrix $\tilde{B}$ is initialized with the classic ICA matrix as described in previous sections. Now we let z constitute the d most significant ICA components of $\tilde{z} : z = z^{\sim}(1:d) = Vx$, where $V = \tilde{B}^{-1}(1:d,:)$ (assuming that the rows of $\tilde{B}$ are arranged according to the energy corresponding to the corresponding directions in the data space). We now model the components of z by the complete MICA model developed in previous sections. In order to evaluate the performance of MICA, first obtain estimates of the original source vector x by assigning the initial estimate of matrix B to be the pseudo-inverse of V,, i.e., $B = (VV^*)^{-1}V$. As in previous sections, we do not re-estimate B but just use the initial estimate along with the optimal MICA parameters computed as above.

III. Simulation Results

We define the $M \times M$ *image patch statistics* of an $N \times N$ image region to be the joint distribution of the random variables (pixel values) from $M \times M$ patches that sample the image region. In this paper we are specifically interested in modeling the $M \times M$ image patch statistics of natural scenes. We first demonstrate the performance of MICA for the case of complete basis for $M = 3$. Since, for larger patch sizes the MICA optimization algorithm becomes more computationally cumbersome, we demonstrate how under-complete MICA models can be successfully exploited to reduce complexity.

To evaluate the complete basis case, we uniformly sampled texture images obtained from the USC-SIPI Brodatz database [4] with $N_{\text{ptch}} = 2000$ patches of size $M \times M$. An ICA was then performed on the data vectors obtained from each texture using Comon's algorithm [5] to obtain the matrix B. Subsequently the parameters $(C, \sigma, \gamma, \mu, \beta)$ of the MICA model were estimated as described in Section II. The parameter β, as mentioned earlier, was estimated heuristically at the outset of the simulation and held to a constant value throughout. To limit computation time, the optimization routine for estimating (C, σ, γ, μ) was forced to terminate after only a few iterations.

Parameter initialization prior to running the optimization routine was performed as follows. Matrix C was initialized to the identity matrix, the entries of σ were initialized to 0.5, and the entries of μ were adjusted such that each of the z channels are

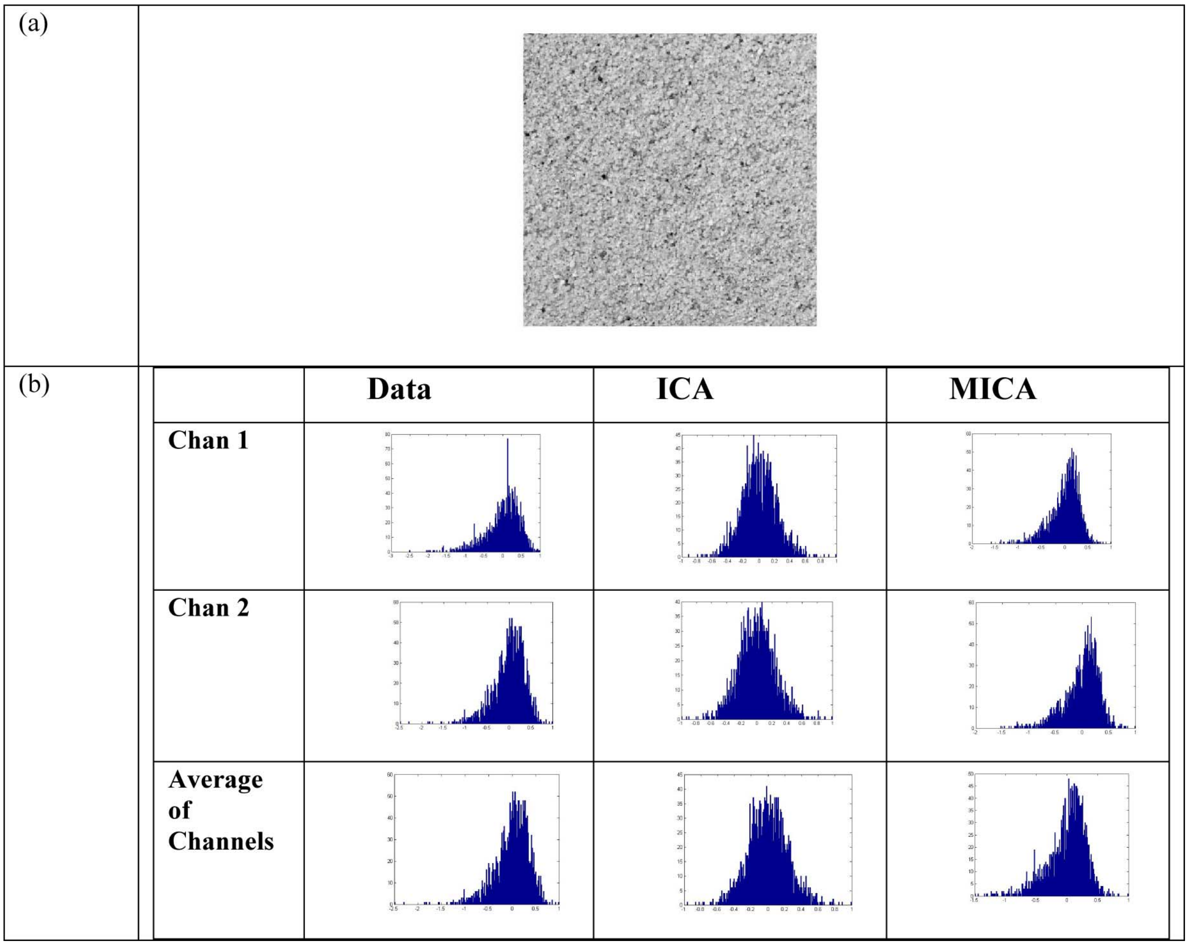


Fig. 6. (a) Sand. (b) Channel histograms of channels and their corresponding ICA and MICA distributions. The high-kurtosis heuristic $\beta_{\mathrm{high-kurt}}$ was used.

zero-mean. After running the MICA optimization routine, setting the γ parameter to the skew of the corresponding data channels gives consistently good performance. The intuitive reason for this choice of γ can be seen by considering the generative model in Fig. 1. Once all the parameters of the MICA model have been adjusted, varying γ determines the asymmetry with which samples are exposed to the linear and quadratic channels: by varying γ, we can directly control the skew of the resulting distribution.

Thereafter, for each texture, we compared the data distribution of each channel derived from test data sets (different from the training data sets) to the corresponding distribution predicted by the ICA and MICA models. In addition, the average of all the data channels was also compared with that predicted by the ICA and MICA models. Simulation of the MICA model was accomplished by generating d i.i.d. zero mean, unit-variance Gaussian channels as shown in Fig. 1, and plotting the histograms outputs of the channels when the optimal parameters (for the texture being modeled) were used. The ICA model was simulated by first computing the empirical distributions of each channel, which were then independently sampled and processed by the matrix B. The histograms of the ICA and MICA channels were then compared with the corresponding channels of the original data distributions using the Kullback–Leibler divergence (KLD) [9]. The above procedure was repeated over several trials, and the average KLD for each channel (for both ICA and MICA) computed with respect to the corresponding channels of the data distributions.

Figs. 5(a)–11(a) depict texture images taken from the Brodatz database [4]. Figs. 5(b)11(b) show the histograms of two of the channels corresponding to each of the textures, as well as both the corresponding computed ICA distributions and the corresponding computed MICA distributions. Also shown are the histograms of the data distributions when all of the data channels of the corresponding textures are averaged together as well as the corresponding computed ICA and MICA distributions. The heuristic strategy used to compute the parameter β for each of these cases is also indicated. Ideally, of course, one can compute β using the optimal derivation given earlier, but the heuristic kurtosis-based approach has proven to yield efficient, near-optimal MICA solutions.

In Figs. 5–7, it is apparent that the MICA model allows for significantly improved approximation of the original data distributions as compared to the classic ICA model. In Figs. 8–11, it is apparent that MICA does a better job in capturing the kurtosis of the channels and does a slightly better job in capturing the skew of the original data distributions; as for example in Figs. 9 and 10. Furthermore, in all cases, there is improved approximation

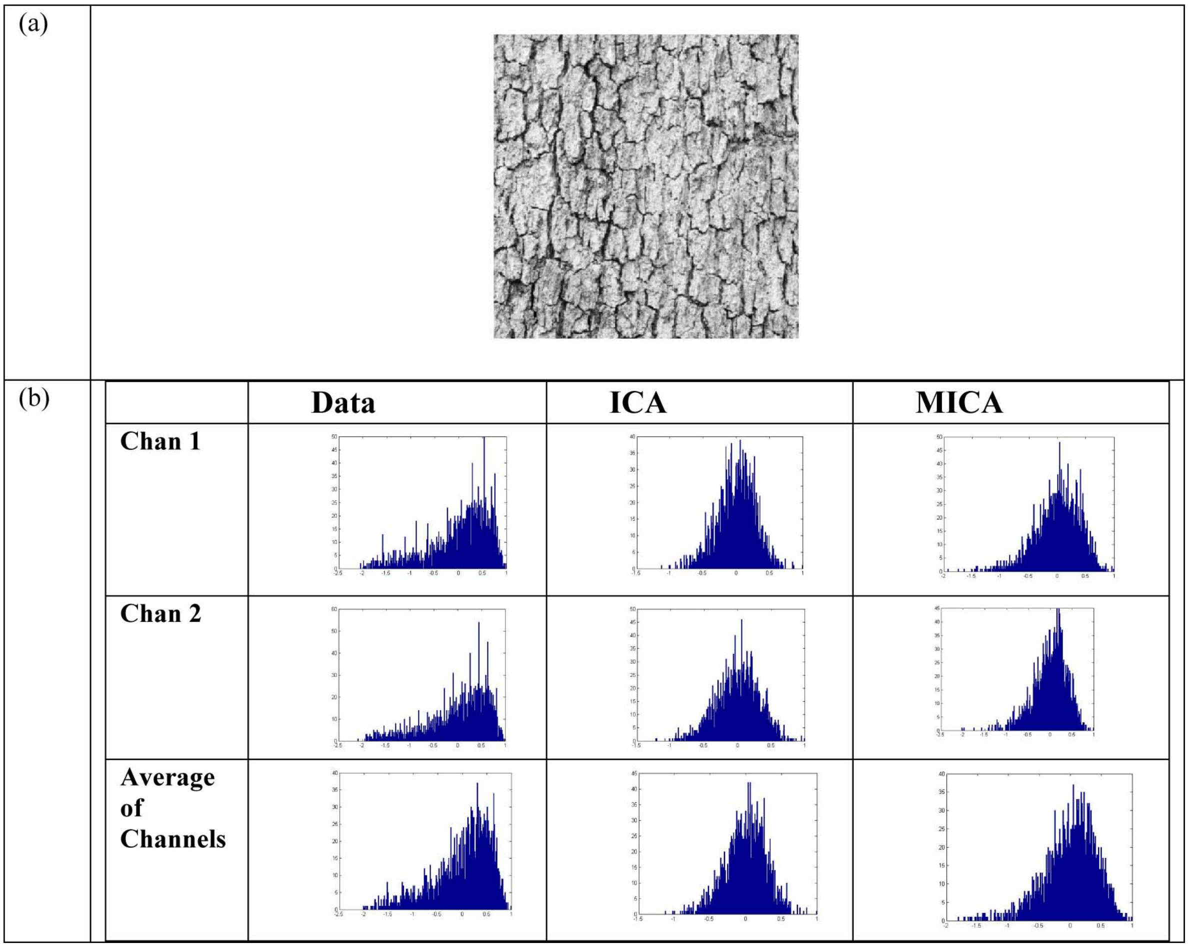


Fig. 7. (a) Bark. (b) Channel histograms of channels and their corresponding ICA and MICA distributions. The low-kurtosis heuristic $\beta_{\mathrm{low-kurt}}$ was used.

of the peak and tail behavior of the original data distributions as compared to ICA. Quantitative evaluation of the MICA model for the complete basis case is provided in Table I, where the relative improvement of MICA relative to classic ICA is measured as

$$\theta_{\mathrm{KLD}}^{\mathrm{MICA}} = \frac{\mathrm{KLD}\,(\mathrm{ICA}) - \mathrm{KLD}\,(\mathrm{MICA})}{\mathrm{KLD}\,(\mathrm{ICA})}$$

where KLD(MICA) is the KLD between the MICA channels and the corresponding original data distributions (averaged across all channels), KLD(ICA) is the corresponding average KLD for the ICA model, and $\theta_{\mathrm{KLD}}^{\mathrm{MICA}}$ is the relative improvement due to MICA over classic ICA with respect to KLD. It is apparent from Table I that the relative performance of MICA is consistently better than that of classic ICA for all natural scene textures.

Similarly, Table II quantifies the performance of the under-complete MICA model for $(M = 5, d = 9)$. The initialization of the parameters before the optimization is similar to that described before, except that the entries of μ are set to zero without subsequent setting of γ to the skew of the data channels. Furthermore, β is always chosen according to the low-kurtosis heuristic. Unlike the complete basis case, $p(z)$ and $P(x)$ are no longer related by a simple scale factor, and so the roles of the different parameters in determining $P(x)$ becomes more complicated. Nevertheless we find, as shown in Table II, that MICA consistently outperforms classical ICA using our simple approach. Under-complete models are useful when is desired to use large patch sizes to sample the image, yet make the problem computationally tractable by working in a lower dimensional subspace. These results demonstrate that the basic idea of multilinear modeling of probability distributions can be successfully extended to under-complete cases.

We further point out that a comprehensive approach to finding the optimal MICA model parameters would be to incorporate an additional simulation optimization phase where the Gaussian random vector S is generated to drive the optimization of the parameters to match the desired data distributions. Such a procedure is likely to be more efficient than a Monte-Carlo simulation approach due to the explicit knowledge of the Jacobian function involved in normalizing the resulting MICA distribution. Nevertheless we have shown that even the computationally simpler optimization approach outlined in this paper suffices to outperform classical ICA.

Tables III and IV show the relative performance of MICA for contrast images and densely sampled textured regions, respectively. Given the original image I, the corresponding contrast

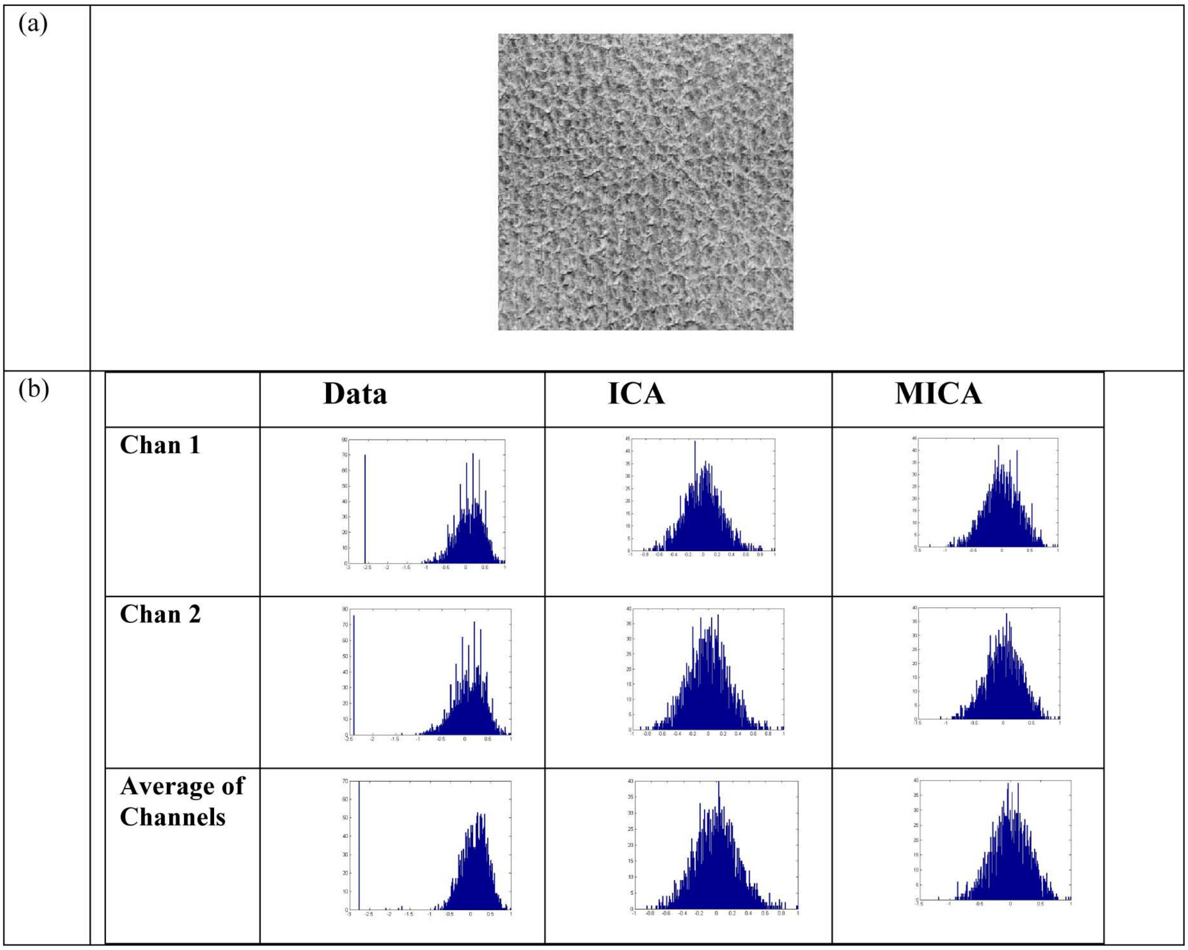


Fig. 8. (a) Pigskin. (b) Channel histograms of channels and their corresponding ICA and MICA distributions. The low-kurtosis heuristic $\beta_{\mathrm{low-kurt}}$ was used.

image $J = C(I)$ was obtained as shown in the equation at the bottom of the page, where N is chosen so that the contrast at each point in the image is computed in a 32 × 32 window [10] about $(m,n), \mu(m,n) = \sum_{i,j} w(i,j)I(m-i,n-j)$ is the local mean of image I around a 32 × 32 window about (m,n), and $\{w(i,j)\}$ is a set of raised cosine filter weights applied to the 32 × 32 window [10]. Contrast plays an important role in visual perception and is the basis for visual adaptation and other mechanisms employed by the human visual system in encoding low-level visual information [11] and for directing visual attention [13]. It is also a useful feature of image processing algorithms that seek to emulate human performance [14]. Table III shows that MICA appears to outperform classic ICA when modeling the image patch statistics of contrast images.

Finally, we also consider the situation where a 32 × 32 patch of a luminance texture image is densely sampled with $M \times M$ patches $(M = 3)$. The resulting samples (resulting in roughly $N_{\mathrm{ptch}} = 1024$ samples/channel) are used to train the MICA model, as before, and subsequently compared with the ICA model. In Table IV, MICA again outperforms classic ICA in modeling the densely sampled image patch statistics.

We emphasize that all the results obtained above are for suboptimal MICA distributions inasmuch as the parameter β was heuristically chosen, and the matrix B was not updated in subsequent iterations. Nevertheless, consistent and statistically significant improvement relative to classic ICA are obtained when modeling image patch statistics, while at the same time revealing detailed quantitative information about the statistical interactions between the ICA components. We finally point out that further improvements in the MICA model are likely possible by means of direct estimation of β, incorporation of simulation phase of optimization, further refinement of the

$$J(m,n) = \sqrt{\left(\sum_{i,j}^{N} w_{i,j}\right)^{-1} \sum_{i,j}^{N} \frac{w_{i,j}[I(m-i,n-j)-\mu(m,n)]^2}{[\mu(m,n)]^2}}$$

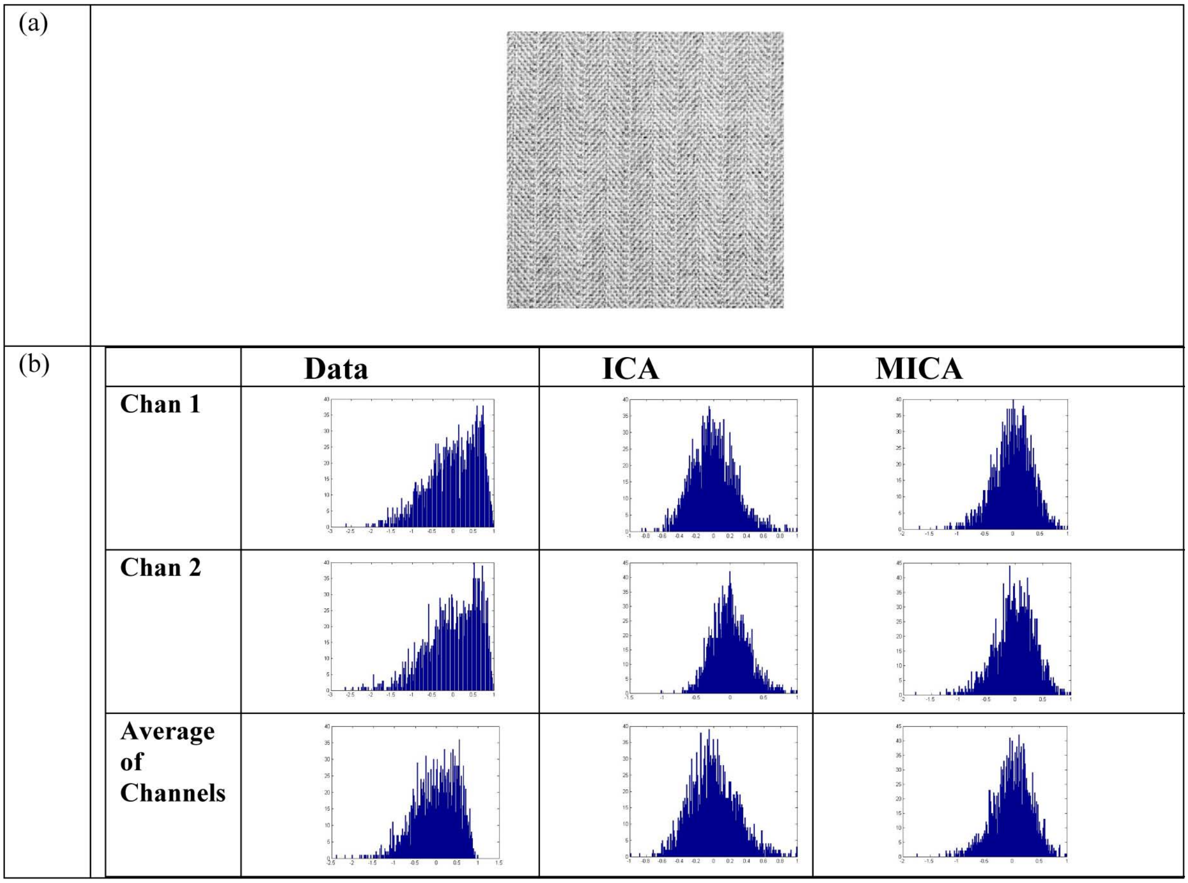


Fig. 9. (a) Herringbone. (b) Channel histograms of channels and their corresponding ICA and MICA distributions. The low-kurtosis heuristic $\beta_{\mathrm{low-kurt}}$ was used.

TABLE I
PERCENT OF IMPROVEMENT IN KLD (W.R.T. ICA) DUE TO MICA MODEL

Texture	% θ_{KLD}^{MICA}
Gravel	47.1576
Sand	55.4293
Bark	56.8961
Pigskin	45.1190
Herringbone	55.3094
Straw	50.3340
Grass	8.7352

Relative improvement with respect to classic ICA when using the *complete basis* MICA model is shown for the various textures

TABLE II
PERCENT OF IMPROVEMENT IN KLD (W.R.T. ICA) DUE TO MICA MODEL

Texture	% θ_{KLD}^{MICA}
Gravel	22.5067
Sand	43.0873
Bark	32.5062
Pigskin	14.0147
Herringbone	7.8592
Straw	18.3099
Grass	21.7529

Relative improvement with respect to classic ICA when using the *under-complete* MICA model.

matrix B, etc., which in turn will be facilitated by faster and more efficient MICA parameter estimation algorithms.

These results demonstrate the considerable promise that multilinear modeling has in capturing the image patch statistics of natural images. Such models can find important applications in image processing and computational vision.

IV. DISCUSSION

In this paper, we have developed multilinear extension of ICA with application to the modeling image patch statistics. A simple linear-quadratic nonlinearity was shown to successfully account for dependences between the pseudo-ICA components, consequently approximating the true structure of the original joint probability distribution much better than possible with simple linear ICA. The quantitative information obtained about the statistical dependences between the pseudo-ICA components, which is naturally furnished by the MICA model, can potentially be used in a variety of applications such as nonstationarity measurement in natural images [7], texture synthesis, and modeling of simple cells in visual cortex.

Apart from such applications, there are open problems that emerged from this work of which we briefly mention a few.

1) *Sparse Coding:* Consider a sparse coding problem involving the joint minimization of the MSE (i.e., mean-squared coding error with respect to $\{\phi_i\}_{i=1}^{d}$)

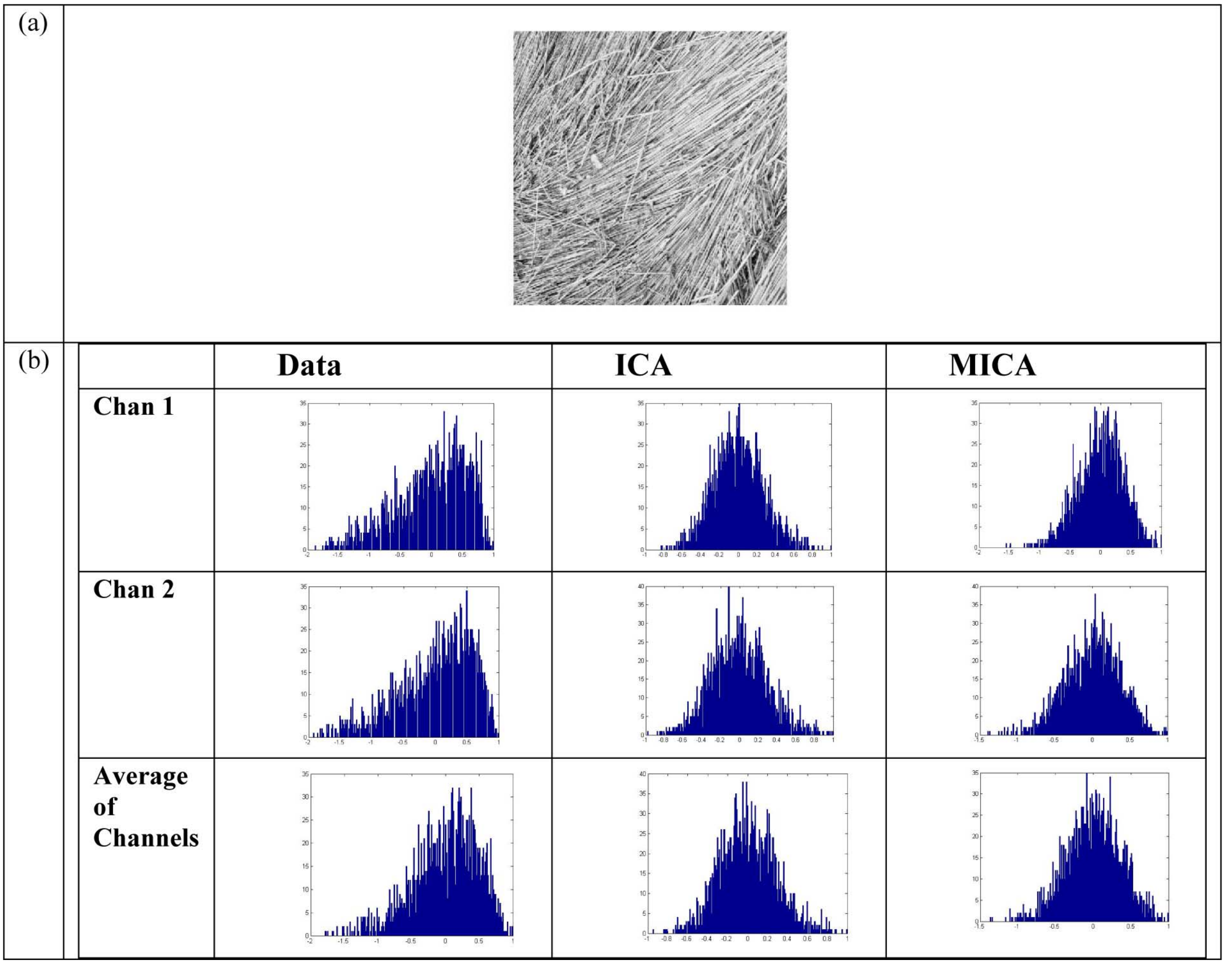


Fig. 10. (a) Straw. (b) Channel histograms of channels and their corresponding ICA and MICA distributions. The low-kurtosis heuristic $\beta_{\mathrm{low-kurt}}$ was used.

TABLE III
PERCENT OF IMPROVEMENT IN KLD (W.R.T. ICA) DUE TO MICA MODEL

Texture	% θ_{KLD}^{MICA}
Sand	11.7162
Herringbone	49.7926
Straw	14.1573

Relative improvement with respect to classic ICA when using MICA for *contrast images*

TABLE IV
PERCENT OF IMPROVEMENT IN KLD (W.R.T. ICA) DUE TO MICA MODEL

Texture	% θ_{KLD}^{MICA}
Sand	29.5017
Herringbone	48.4710
Straw	27.2106

Relative improvement with respect to classic ICA when using MICA for densely sampled texture regions

and a sparsity term induced by $g(J)$. Is there an optimum basis set that is a solution to this problem?

2) *Over-complete Models:* A first step is to address the problem of parameter estimation of a mixture of MICA models. This would have the added benefit of enabling the analysis of data from multimodal probability distributions.

3) *Nonsparse Multilinear Forms:* The basic methodology outlined here can be used to explore the original joint distribution with respect to projections on arbitrary basis; for example, the matrix B can be initialized with Gabor vectors.

Finally, there is considerable scope for improving the existing MICA model in terms of devising more efficient algorithms for parameter estimation, thus improving parameterizations of the MICA model and, thus, for unleashing the full potential of this statistical modeling methodology.

APPENDIX

Proof of Theorem 1: We prove the lemma by induction on d. For the base case $(d = 2)$ it is easily shown that

$$J(F) = \frac{1}{\beta_1 \beta_2 |C|} \prod_{k=1}^{2} \psi(\beta_k z_k).$$

Now assume by inductive hypothesis that the lemma is true for $d = 2, \ldots, N$.

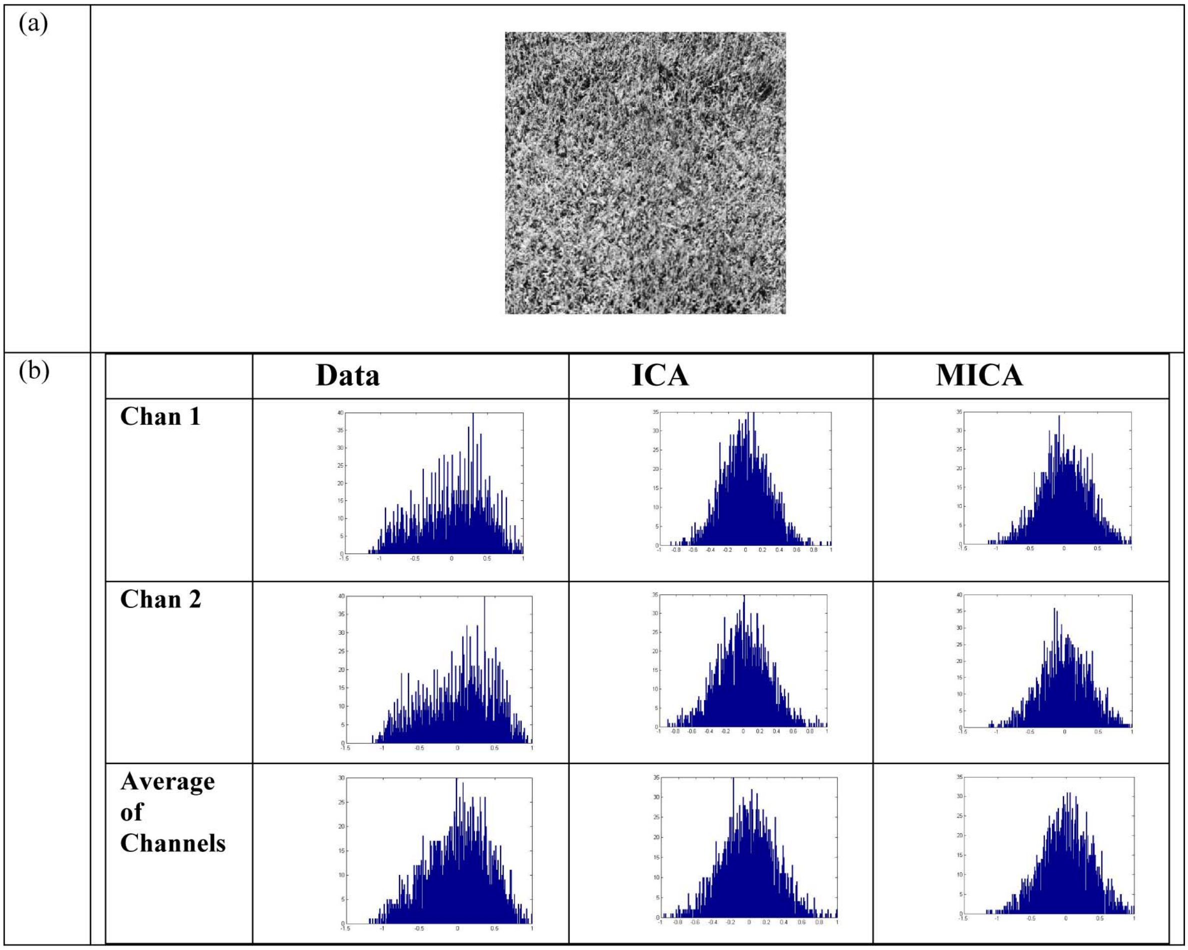


Fig. 11. (a) Grass. (b) Channel histograms of channels and their corresponding ICA and MICA distributions. The low-kurtosis heuristic $\beta_{\mathrm{low-kurt}}$ was used.

Consider the Jacobian when $d = N + 1$

$$J(F) = \begin{vmatrix} \frac{\partial F_1}{\partial s_1} & \cdot & \cdot & \frac{\partial F_1}{\partial s_{N+1}} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \frac{\partial F_{N+1}}{\partial s_1} & \cdot & \cdot & \frac{\partial F_{N+1}}{\partial s_{N+1}} \end{vmatrix}$$

$$= \begin{vmatrix} a_{1,1}\psi(z_1) & \cdot & \cdot & a_{1,N+1}\psi(z_1) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ a_{N+1,1}\psi(z_{N+1}) & \cdot & \cdot & a_{N+1,N+1}\psi(z_{N+1}) \end{vmatrix}.$$

Expand $J(F)$ with respect to the first row

$$J(F) = \left(\prod_{j=1}^{N+1} \beta_j\right)^{-1} \sum_{k=1}^{N+1} (-1)^{1+k} |C_{1,k}| a_{1,k} \psi(z_1)$$

where $C_{1,k}$ is the minor matrix of $J(F)$ with respect to $(1,k)$. Applying the inductive hypothesis yields

$$J(F) = \frac{1}{\beta^{N+1}} \prod_{k=1}^{N+1} \psi(z_k) \sum_{k=1}^{N+1} (-1)^{1+k} |A_{1,k}| a_{1,k}$$

where $A_{1,k}$ is the minor matrix of A with respect to $(1,k)$. Thus

$$J(F) = \frac{1}{\beta^{N+1}} \prod_{k=1}^{N+1} \psi(z_k)|A| = \frac{1}{\beta^{N+1}|C|} \prod_{k=1}^{N+1} \psi(z_k)$$

thereby proving the lemma for all d. ♣

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley Interscience, 2001.

[2] J.-F. Cardoso, "Blind source separation: Statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[3] A. Hyvarinen, P. O. Hoyer, and M. Inki, "Topographic independent component analysis," *Neural Comput.*, vol. 13, no. 7, pp. 1527–1558, Jul. 2001.

[4] USC Signal & Image Processing Institute Image Database vol. 1 [Online]. Available: http://sipi.usc.edu/database/database.cgi?volume=textures

[5] P. Comon, "Independent component analysis: A new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[6] R. G. Raj, A. C. Bovik, and W. S. Geisler, "Non-stationarity detection in natural images," presented at the IEEE Int. Conf. Image Process., San Antonio, TX, Sep. 2007.

[7] R. G. Raj and A. C. Bovik, "Non-stationarity measurement in natural images with applications to texture based fixation selection," *IEEE Trans. Image Process.*, to be published.

[8] N. Balakrishnan and C. R. Rao, Eds., *Handbook of Statistics 16: Order Statistics: Theory & Methods*. New York: Elsevier, 1998.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[10] R. G. Raj, W. S. Geisler, R. A. Frazor, and A. C. Bovik, "Contrast statistics for foveated visual systems: Fixation selection by minimizing contrast entropy," *J. Opt. Soc. Amer.*, vol. 22, pp. 2039–2049, Oct. 2005, A.

[11] B. A. Wandell, *Foundations of Vision*. Sunderland, MA: Sinauer, 1995.

[12] A. Hyavarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[13] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "Foveated analysis of image features at fixations," *Vision Res.*.

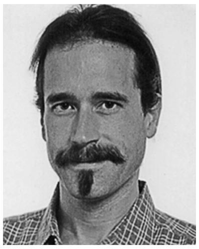
[14] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "GAFFE: A gaze-attentive fixation finding engine," *IEEE Trans. Image Process.*, to be published.

[15] U. Koster and A. Hyvarinen, "A two-layer ICA-like model estimated by score matching," in *Proc. Int. Conf. Artificial Neural Networks*, Porto, Portugal, 2007, pp. 798–807.

[16] Y. Karklin and M. S. Lewicki, "A hierarchical Bayesian model for learning non-linear statistical regularities in non-stationary natural signals," *Neural Comput.*, vol. 17, no. 2, pp. 397–423, 2005.

Raghu G. Raj (S'01) received B.S. degrees in computer science and electrical engineering from Washington University, St. Louis, MO, in 1998, the M.S. degree in electrical engineering from the University of Texas (UT) at Austin in 2000, during which period he was also a Graduate Research Assistant in the Advanced Sonar Division of the Applied Research Laboratories (ARL), UT Austin, and the Ph.D. degree in electrical engineering in the area of visual search from UT Austin in 2007.

From 2000–2004, he was with Motorola, Inc., Austin, TX, working on signal processing applications for communication systems. His areas of interest include visual search, automatic target recognition, signal/image processing, multidimensional stochastic processes, computational vision, pattern recognition, data mining, and data compression.



Alan C. Bovik (S'80–M'81–SM'89–F'96) received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from the University of Illinois, Urbana-Champaign, in 1980, 1982, and 1984, respectively.

He is currently the Curry/Cullen Trust Endowed Professor at the University of Texas, Austin, where he is the Director of the Laboratory for Image and Video Engineering (LIVE) in the Center for Perceptual Systems. His research interests include image and video processing, computational vision, digital microscopy, and modeling of biological visual perception. He has published over 450 technical articles in these areas and holds two U.S. patents. He is the author of *The Handbook of Image and Video Processing*, 2nd ed. (Elsevier Academic Press, 2005) and *Modern Image Quality Assessment* (Morgan & Claypool, 2006).

Dr. Bovik has received a number of major awards from the IEEE Signal Processing Society, including: the Education Award (2007), the Technical Achievement Award (2005), the Distinguished Lecturer Award (2000), and the Meritorious Service Award (1998). He is also a recipient of the IEEE Third Millennium Medal (2000), and a two-time Honorable Mention winner of the international Pattern Recognition Society Award for Outstanding Contribution (1988 and 1993). He is a Fellow of the Optical Society of America. He has been involved in numerous professional society activities, including: Board of Governors, IEEE Signal Processing Society, 1996–1998; Editor-in-Chief, IEEE TRANSACTIONS ON IMAGE PROCESSING, 1996–2002; Editorial Board, THE PROCEEDINGS OF THE IEEE, 1998–2004; Series Editor for *Image, Video, and Multimedia Processing*, Morgan & Claypool Publishing Company, 2003-present; and Founding General Chairman, First IEEE International Conference on Image Processing, held in Austin, TX, in November 1994. Dr. Bovik is a Registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial, and academic institutions.