Image Denoising with Unsupervised, Information-Theoretic, Adaptive Filtering

Suyash P. Awate and Ross T. Whitaker

UUCS-04-013

School of Computing
University of Utah
Salt Lake City, UT 84112 USA

Abstract

The problem of denoising images is one of the most important and widely studied problems in image processing and computer vision. Various image filtering strategies based on linear systems, statistics, information theory, and variational calculus, have been effective, but invariably make strong assumptions about the properties of the signal and/or noise. Therefore, they lack the generality to be easily applied to new applications or diverse image collections. This paper describes a novel unsupervised, information-theoretic, adaptive filter (UINTA) that improves the predictability of pixel intensities from their neighborhoods by decreasing the joint entropy between them. In this way UINTA automatically discovers the statistical properties of the signal and can thereby reduce noise in a wide spectrum of images and applications. The paper describes the formulation required to minimize the joint entropy measure, presents several important practical considerations in estimating image-region statistics, and then presents a series of results and comparisons on both real and synthetic data.

Image Denoising with Unsupervised, Information-Theoretic, Adaptive Filtering

Suyash P. Awate and Ross T. Whitaker

Scientific Computing and Imaging Institute,

School of Computing,

University of Utah

October 24, 2004

Abstract

The problem of denoising images is one of the most important and widely studied problems in image processing and computer vision. Various image filtering strategies based on linear systems, statistics, information theory, and variational calculus, have been effective, but invariably make strong assumptions about the properties of the signal and/or noise. Therefore, they lack the generality to be easily applied to new applications or diverse image collections. This paper describes a novel unsupervised, information-theoretic, adaptive filter (UINTA) that improves the predictability of pixel intensities from their neighborhoods by decreasing the joint entropy between them. In this way UINTA automatically discovers the statistical properties of the signal and can thereby reduce noise in a wide spectrum of images and applications. The paper describes the formulation required to minimize the joint entropy measure, presents several important practical considerations in estimating image-region statistics, and then presents a series of results and comparisons on both real and synthetic data.

1 Introduction

The problem of denoising images is one of the most important and widely studied problems in image processing and computer vision. Research has led to a plethora of algorithms based on diverse strategies such as linear systems, statistics, information theory, and variational calculus. However, most of the image filtering strategies invariably make strong assumptions about the properties of the signal and/or noise. Therefore, they lack the generality to be easily applied to diverse image collections and they break down when images exhibit properties that do not adhere to the underlying assumptions. Hence, there is still a need for general image filtering algorithms/strategies that are effective for a wide spectrum of denoising tasks and are easily adaptable to new applications.

This paper describes a novel *unsupervised information-theoretic adaptive filter* (UINTA) for image denoising. UINTA denoises pixels by comparing pixel values with other pixels in the image that have similar neighborhoods. The underlying formulation relies on an information-theoretic measure of goodness combined with a nonparametric model of image statistics. The information-theoretic optimization measure relies on the entropy of the patterns of intensities in image regions. Entropy is a nonquadratic function of the image intensities, and therefore the filtering operation is nonlinear. UINTA operates without a priori knowledge of the geometric or statistical structure of the signal or noise, but relies instead on some very general observations about entropy of natural images. It does not rely on labeled examples to shape its output, and is therefore *unsupervised*. Because UINTA automatically generates a statistical representation of the image derived from the input data and constructs a filtering strategy based on that model, it is *adaptive*. Moreover, UINTA adjusts its free parameters automatically using a data-driven approach and information-theoretic metrics. Because UINTA is nonlinear, nonparametric, adaptive, and unsupervised, it can automatically reduce image noise in a wide spectrum of images and applications.

The organization of the remainder of the paper is as follows. Section 2 discusses recent

work in image filtering, including both variational and statistical formulations, and discusses their relationship to the proposed method. Section 3 describes the mathematical formulation of the proposed filtering scheme and motivates the choice of the particular information-theoretic measure based on joint entropy. Entropy optimization entails the estimation of probability densities for the associated random variables. Hence, Section 4 describes a non-parametric statistical technique for multivariate density estimation from a finite number of random samples. That section also describes the general problems associated with density estimation in high-dimensional spaces and give some reasons behind the success of UINTA in spite of these difficulties. Section 5 formulates a gradient-descent scheme to optimize the joint entropy measure and discusses several important practical challenges pertaining to statistical estimation and its application to image neighborhoods. Section 6 gives numerous results for the experiments on real and synthetic images and analyzes the behavior of UINTA on the same. Section 7 summarizes the contributions of the paper and presents ideas for further exploration.

2 Related Work

The literature on signal and image denoising is vast, and a comprehensive review is beyond the scope of this paper. This section establishes the relationship of this work to several important, relevant areas of nonlinear image filtering.

Classical approaches to image filtering, also applied to the problem of image restoration, rely mostly on the application of the theory of linear systems to images [6]. This work includes Fourier transform methods, least-squares methods, wavelet-based methods, and scale-space theory [17, 22, 6]. Although these are computationally efficient, because of their linearity, the effects of these algorithms are not local in space and therefore they have difficulties dealing with local features, such as edges. Nonlinear filtering approaches are typically based on either variational methods, which result in algorithms based on partial

differential equations (PDEs); or statistical methods, which result in nonlinear estimation problems. Nonlinear filters can overcome some of the limitations of linear filters, but they introduce some problems such as higher computational costs and the additional tuning of extra free parameters. Furthermore, most linear and nonlinear approaches enforce specific geometric or statistical assumptions on the image.

PDE-based image processing methods became widespread after the work of Perona and Malik [32], where they propose a modified version of the heat equation (calling it *anisotropic diffusion*) to include an inhomogeneous diffusivity term. The diffusivity is based on the local intensity gradient, and the result is a nonlinear diffusion equation localizing the diffusion in space. Analytical and empirical analyses show that the method smoothes images in regions of low gradient and sharpens or enhances edges. The anisotropic diffusion equation is also the first variation of an image energy that penalizes image gradients with an allowance for outliers [29, 41, 4], and therefore seeks piecewise constant solutions (in 1D—the situation is somewhat more complex in multiple dimensions but holds qualitatively). Because such variational approaches prefer certain image geometries, we refer to these local geometric configurations as *models*. In a related body of work, Rudin, Osher, and Fatemi [36] propose a restoration method that relies on the total variation prior [30], which allows grayscale discontinuities as part of a bounded-variation image model. A multitude of nonlinear PDE models have been developed for a wide variety of images and applications [35, 48], including PDE versions of the Mumford and Shah [27] variational model (which explicitly models edges) and a variety of algorithms based on level sets [30, 38, 1, 7].

Several authors have attempted to extend these PDE-based methods to more complicated image models. Vese *et al.*[45] model textured images by functional minimization and partial differential equations by decomposing images as a sum of two functions—a cartoon-like image (bounded variation) and a texture image. Weickert has proposed a *coherence enhancing* flow (not derived as a variation of an image energy), which preserves and enhances textures that exhibit a homogeneous structure tensor [49]. Several authors have proposed higher-order

flows that correspond to piecewise-linear image models [2, 43, 25]. These nonlinear PDE models have proven to be very effective, but only for particular applications where the input data is well suited to the model's underlying geometric assumptions. Moreover, the parameter tuning is a challenge because it entails fuzzy thresholds that determine which image features are preserved (or enhanced) and which are smoothed away.

The statistical approaches to nonlinear filtering fall into several classes. One class is the methods that use *robust statistics*, the most prevalent being the median filter. The median filter enforces a constant image model with an allowance for outliers; iterative applications of the median filter to 1D signals result in piecewise-flat solutions. Miller [26] has proposed robust statistics for fitting higher-order models. Bilateral filtering [44] is a robust, nonlinear filtering algorithm that replaces each pixel by the weighted average over a neighborhood with a fuzzy mechanism for excluding outliers. Like many of the variational approaches, these statistical methods are essentially mechanisms for fitting simple geometric models to local image neighborhoods in a robust way.

Another class of statistical methods for image processing rely on stochastic image models described by Markov random fields (MRFs), as proposed by Geman and Geman [16]. The Markov property for images is based on the assumption of spatial dependency or predictability of the image intensities—it implies that the probability of a pixel having a particular intensity depends only on the intensities of its spatial neighbors. In [16] they describe an algorithm that relies on Gibbs sampling to modify pixel intensities. Gibbs sampling, assuming the knowledge of the conditional distributions of the pixel intensities given the intensities of their neighbors, generates a Markov chain of pixel intensities which converges (point wise) to the desired denoised image. These conditional probabilities for image neighborhood configurations (called *cliques*) play a similar role to the image energy in the variational approaches. For instance, MRF image models often include extra parameters (*hidden* parameters) that explicitly model intensity edges, allowing these models to achieve piecewise-constant solutions. Thus, these conditional probabilities encode a set of probabilistic assumptions about

the geometric properties of the signal (noiseless image). The method in this paper also exploits the Markov property of the images, but in a different context. Rather than imposing a particular model on the image, UINTA *estimates* the relevant conditional probability density functions (PDFs) from the input data and updates pixel intensities to decrease the randomness of these conditional PDFs.

Figure 1 shows the results [1] of filtering on the *Lena* image using some of the prevalent nonlinear techniques, demonstrating their typical characteristics. Perona and Malik's diffusion (Figure 1(c)) eliminates the noise on the cheeks but introduces spurious edges near the nose and the lips. Bilateral filtering [44] (Figure 1(d)), which is essentially an integral form of anisotropic diffusion [3], tends to smooth away fine textures resulting in their elimination,

[1] In printed copies, it may be difficult to notice/distinguish subtle features/differences in many of the images in this paper. Please refer to the electronic copy, whenever in doubt.

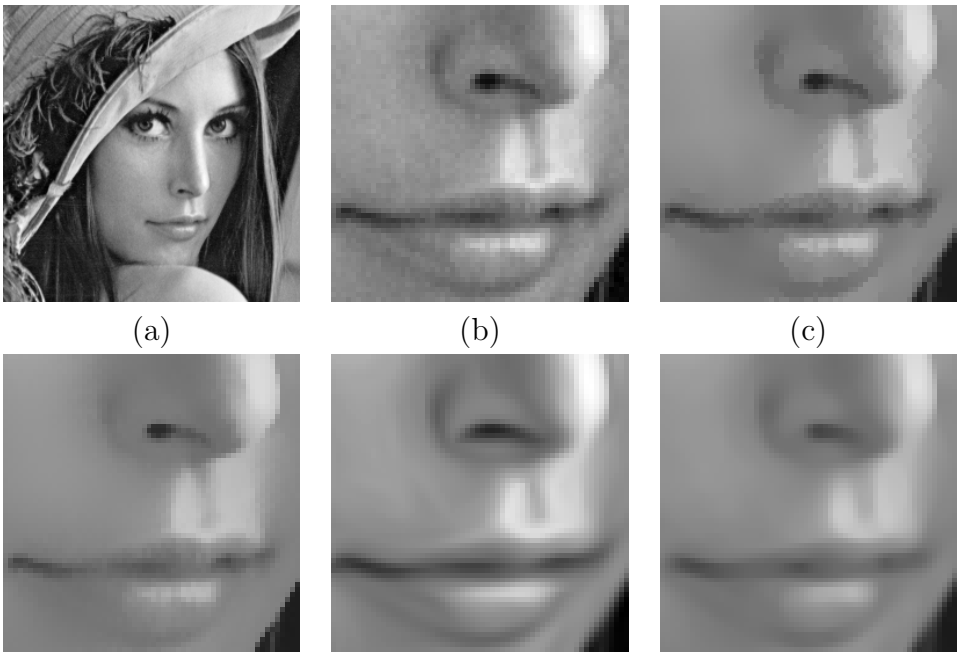


Figure 1: (a) Noisy *Lena* image (256×256, grayscale values range:0-255). (b) A magnified portion of the noisy image. Results with (c) Anisotropic diffusion (K=0.5 grayscale values, 20 iterations) (d) Bilateral filtering (σ_{domain}=3 pixels, σ_{range}=12 grayscale values) (e) Coherence enhancing diffusion (σ=0.1 pixels, ρ=2 pixels, α=0.0001, C=0.0001, 15 iterations) (f) Curvature flow (time step=0.2, 8 iterations)

e.g. on the lips. Both of these algorithm entail two free parameters (scale and contrast), and require significant tuning. The coherence enhancing diffusion (Figure 1(e)) forces specific elongated shapes in images, as seen in the enlarged nostril and the lips' curves. On the other hand, curvature flow [38, 30], which is very similar to the total variation strategy of [36], tends to shrink features by rounding them off (Figure 1(f)). The *Lena* image, which appears to be a very typical grayscale photograph, does not adhere very well to the basic geometric models underlying these algorithms.

Recently, researchers have begun to analyze the statistics of natural images in terms of local neighborhoods, and are drawing conclusions that are consistent with MRF models of images. For instance, Huang *et al.*[21] analyze the statistical properties of the intensity and range values of natural images. These include single pixel statistics, two-point statistics and derivative statistics. They found that the mutual information between the intensities of two adjacent pixels in natural images is rather large and attributed this to the presence of spatial correlation in the images. Lee *et al.*[24] and Silva *et al.*[11] analyze the statistics of 3×3 high-contrast patches in optical images, in the corresponding high-dimensional spaces, and find the the data to be concentrated in clusters and low-dimensional manifolds exhibiting a nontrivial topology. The work in this paper also relies on the hypothesis that natural images exhibit some regularity in neighborhood structure, but UINTA discovers this regularity for each image individually in a nonparametric manner.

The literature shows several statistically-based image processing algorithms that do rely on information theory. For instance, the Maximum Entropy Method (MEM) [22, 33] is used in astronomy to deal with the particular nature of blurred, speckled images that are prevalent in that field. MEM entails maximizing the entropy of the image-intensity distribution with a fidelity constraint. MEM is related to the standard image enhancement technique of histogram equalization—i.e. a flat histogram maximizes entropy. MEM, by increasing the entropy, tries to restore the fine details that are lost to blurring (during imaging). Conventional image filtering, e.g. through low-pass transforms, typically *decreases* the entropy

of signals [39]. Indeed MEM-processed images sometimes appear too noisy because of the method's failure to model the regularity of local neighborhoods. This drawback has led to the development of the intrinsic correlation function or preblur function [5], which explicitly imposes spatial correlation, and hence smoothness, in MEM image processing.

Another information-theoretic processing method is the *mean-shift* algorithm [15, 40, 8, 9], which moves the samples uphill on a PDF associated with the data. This process produces a steady state in which all of the data have values corresponding to the nearest local maximum of the PDF (assuming appropriate windowing strategies). The mean-shift procedure, thus, can be said to be a *mode seeking* process. However, the mean-shift algorithm operates only on image intensities (be they scalar or vector valued) and does not account for neighborhood structure in images. Thus, mean shift resembles a kind of image-driven thresholding process (particularly in the algorithm proposed by [9], in which the density estimate is static as the algorithm iterates). This paper shows the mathematical relationship between the mean-shift procedure and entropy reduction and thereby formulates UINTA as a generalization of the mean-shift algorithm, which incorporates image neighborhoods to reduce the entropy of the associated conditional PDFs.

Several other bodies of research also relate to the proposed method. One area uses image coding/compression as a mechanism for denoising [28]. The idea is that effective lossy image compression looses irrelevant information (noise) while maintaining the most meaningful aspects of the data (signal). Often, lossy compression algorithms seek to decrease the degree of randomness in images. This strategy raises the compression problem to the difficulty of the denoising problem. In practice, however, lossy compression algorithms used for denoising depend on particular image decompositions (e.g. wavelets) and rely on particular assumptions about the importance of various wavelet coefficients in the actual signal. However, these approaches do share a common strategy with the proposed work, which is that denoising is a process that *increases* the redundancy in the image as measured by image statistics.

Another related body of research is by Weissman *et al.*, [50], who propose the DUDE

algorithm. DUDE addresses the problem of denoising data sequences generated by a discrete source and received over a discrete, memoryless channel. DUDE assumes that the source and the received data sequences take values from a finite population of symbols and that the transition probabilities over the channel are known. However, DUDE assumes no knowledge of the statistics of the source and yet performs (asymptotically) as well as any denoiser (e.g., one that knows the source distribution), thereby making DUDE *universal* and *optimal*. DUDE assigns image values based on the similarity of neighborhoods gathered from image statistics, which resembles the construction of conditional probabilities in the proposed method. However, the DUDE approach is limited to discrete-valued signals whereas the proposed method addresses continuous-valued signals, such as those associated with grayscale images. While the DUDE algorithm is demonstrably effective for removing *replacement noise*, it is less effective in case of additive noise.

The literature dealing with *texture synthesis* also sheds some light on the proposed method. Recent texture synthesis algorithms rely on image statistics from an input image to construct novel images that bear a qualitative resemblance to the input [47, 14]. Given a texture image, a new image with similar texture is generated by marching sequentially through the new image and inserting pixel values in the new image by finding the neighborhoods in the input image that best match the current neighborhood in the new image. This is a different application and these algorithms do not rely on information-theoretic formulations, but they demonstrate the power of neighborhood statistics, and mode-seeking processing in capturing essential aspects of image structure.

3 Joint Entropy Based Image Filtering

This Section describes the formulation of UINTA. It begins with an overview of the notation and a review of information-theoretic measures. It shows how these measures are applied to image neighborhoods, concluding with a high-level algorithmic description of UINTA.

3.1 Information-Theoretic Measures

The notation in the paper is as follows. An uppercase letter, e.g. X, denotes a random variable (RV), which may be scalar/vector-valued, as necessary. A lowercase letter, e.g. x, denotes the value of a particular sample from the sample space of X. $p(x)$ denotes the probability density function (PDF) for X. Applying a function $f(\cdot)$ on X yields a new RV, $f(X)$. When the function is the PDF $p(x)$ itself, we refer to the new RV as $p(X)$.

The Shannon *entropy* of a RV measures the information, or uncertainty, associated with the RV [39, 10]. For a continuous RV X the *differential entropy* $h(X)$ is

$$h(X) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx = -E_p[\log p(X)] \tag{1}$$

where E_p denotes the expectation of the RV with X drawn from the PDF $p(x)$.

The *conditional entropy* between two continuous RVs X and Y measures the uncertainty remaining in X after observing Y. It is defined as the weighted average of the entropies associated with the conditional PDFs [10, 18].

$$h(X|Y) = \int_{-\infty}^{\infty} p(y) h(X|Y = y) dy = h(X, Y) - h(Y), \tag{2}$$

where $h(X, Y)$ is the *joint entropy* of the RVs X and Y.

Besides entropy, several other measures exist that measure the relative information content *between* variables. *Mutual information*, $I(X, Y) = h(X) + h(Y) - h(X, Y)$, quantifies the uncertainty in X that is *resolved* after observing Y. For a set of RVs, $Z_1, \ldots, Z_n$, the *multiinformation*, $M(Z_1, \ldots, Z_n) = \sum_i h(Z_i) - h(Z_1, \ldots, Z_n)$, generalizes mutual information and quantifies all combinations of functional dependencies [42].

All these information-theoretic quantities are potentially interesting for processing images based on neighborhood statistics. However, each measure has a distinct effect on the marginal and joint PDFs of the image neighborhoods, as discussed in Section 3.2.

3.2 Neighborhood Entropy

The following discussion assumes a 2D image, but the formulation extends directly to higher dimensional images. A 2D grayscale image is a function $I: S \mapsto \mathbb{R}$, which assigns a real value to each element of a domain, $S \subset \{\mathbb{I} \times \mathbb{I}\}$, where typically $S = (1, \ldots, n) \times (1, \ldots, m)$. We call each point in this domain as a pixel. A *region* $r(s) \subset S$ centered at a pixel location s is the ordered set (i.e. vector) of pixels $\{t: |t - s|_\infty \leq d\}$, where the set ordering is based on the values of the two spatial coordinates of pixels t, and d denotes the region size. Thus $r(s)$ is an $n = (2d + 1)^2$ dimensional vector comprising the locations of the pixels in the region centered at s. We write the vector $r(s) = (s + o_1, \ldots, s + o_n)$ where $o_1, \ldots, o_n$ correspond to the offsets from the center pixel to its neighbors, and $o_c = 0$ is the offset of the center pixel.

The UINTA algorithm filters images by increasing the predictability of pixel intensities from the neighborhood intensities. This requires an entropy measure for intensities in image regions. We first define a continuous RV $X : S \mapsto \mathbb{R}$ that maps each pixel in the image to its intensity. The statistics of X are the grayscale statistics of the image. Thus $x(s) = I(s)$, and every intensity in the image is seen as a *realization* of the RV X. Next we define the random vector $Y = (X(s + o_1), \ldots, X(s + o_{c-1}), X(s + o_{c+1}), \ldots, X(s + o_n))$ which captures the statistics of intensities in pixel neighborhoods. Thus $y(s) = (I(s + o_1), \ldots, I(s + o_{c-1}), I(s + o_{c+1}), \ldots, I(s + o_n))$. Let $Z = (Z_1, \ldots, Z_n) \equiv (X, Y)$ be a random vector taking values as $z(s) = (I(s+o_1), \ldots, I(s+o_n))$ representing pixel intensities in the region $r(s)$ centered at s. We call the space in which the vectors $z(s)$ lie as the *feature space*. The proposed algorithm relies on the statistical relationship between the intensity of each pixel and the intensities in a set of nearby pixels defined by its neighborhood. The strategy, for each pixel-neighborhood pair $(X = x(s), Y = y(s))$ from the image, is to reduce the entropy $h(X|Y = y(s))$ of the conditional PDF by manipulating the value of each center pixel $x(s)$.

The UINTA algorithm employs a gradient descent to minimize entropies of the conditional PDFs. In principle, the gradients of $h(X|Y)$ have components corresponding to both the

center pixel, $x(s)$, and the neighborhood, $y(s)$, and thus the entire neighborhood, $(x(s), y(s))$, would be updated for a gradient descent scheme. In practice we update only the center pixel of each neighborhood. That is, we project the gradient onto the direction associated with the center pixel. Given this projection, UINTA is a reweighted gradient descent on either the joint entropy, $h(X, Y)$, or the conditional entropy, $h(X|Y)$—they are equivalent for this particular descent strategy.

This choice of entropy as a measure of goodness follows from several observations. First, the addition of any sort of noise to the image increases the joint entropy because the sum of two random variables corresponds to a convolution of their PDFs in the probability space, which necessarily increases entropy [39]. Thus, any kind of denoising algorithm (for additive noise) must decrease entropy. Of course, continuing entropy reduction by filtering forever might not be *optimal* and might also eliminate some of the normal variability in the signal (noiseless image). However, UINTA is motivated by the observation that noiseless images tend to have very low entropy relative to their noisy counterparts. Thus, UINTA, as an entropy reducing filter, first affects the noise (in the noisy image) substantially more than the signal. Second, among the various measures of information content, the proposed entropy measure, $h(X|Y = y(s))$, makes sense for several reasons. For an image, $h(X|Y = y(s))$ is low when the center pixels, $x(s)$, are predictable from their neighborhoods, $y(s)$. However, $h(X|Y = y(s))$ will also be low when the image by itself is predictable, e.g. an image with a constant intensity. Although maximizing mutual information ($I(X, Y)$) and multiinformation ($M(Z_1, \ldots, Z_n)$) penalizes joint entropy, it rewards higher entropy among the individual RVs. This tendency towards higher entropy stems from the term $h(X)$ (in $I(X, Y)$) and $\sum_i h(Z_i)$ (in $M(Z_1, \ldots, Z_n)$). Thus, these information measures tend to perform denoising *and* contrast enhancement simultaneously. For many images, enhancement and denoising are not compatible goals. Furthermore, additive noise can, in some cases, increase the multiinformation measure.

3.3 High-Level Structure of the UINTA Algorithm

The high-level structure of the UINTA algorithm is as follows.

1. The noisy input image, namely I, consists of a set of intensities $x(s)$. These values form the initial values of a sequence of images $I^0, I^1, I^2, \ldots$.

2. Using the image I^m, construct another image (with same dimensions), V^m, composed of the intensity vectors, $z^m(s)$, of length $n = (2d+1)^2$.

3. For each pixel $z^m(s) \equiv (x^m(s), y^m(s))$ in V^m, estimate the PDF $p(x|Y = y^m(s))$ and compute the derivative of $h(X|Y = y^m(s))$ with respect to $x^m(s)$.

4. Construct a new image I^{m+1} consisting of pixel values $x^{m+1}(s)$ using finite forward differences on the gradient descent: $x^{m+1}(s) = x^m(s) - \lambda \partial h / \partial x^m(s)$.

5. Based on a suitable stopping criterion, terminate, or go to Step 2. (Appendix B discusses more about stopping criteria.)

The algorithm includes two important parameters. The first is the size of the image neighborhoods (the parameter d in the previous discussion). Typically, values of 1 or 2 suffice (giving regions of sizes 3×3 or 5×5 and feature spaces of dimensions 9 or 25, respectively). However, as we shall see in later sections, more complex or noisy images may require larger neighborhoods for reliable denoising. The second free parameter is in the stopping criterion. For most natural images we would not expect the steady states of the UINTA filter to be an acceptable result for a denoising task—that is, we expect some degree of variation in neighborhoods. However, the algorithm is consistent with several conventional techniques for enforcing fidelity to the input data; such mechanisms inevitably introduce an additional parameter.

Although each step of the UINTA algorithm operating on a single pixel (Step 4 above) is merely a gradient descent on the center pixel, the interactions from one iteration to the

next are quite complicated. The updates on the center-pixel intensities in Step 4 affect, in the next iteration, not only the center pixels but also the neighborhoods. This is because the image-regions, $r(s)\forall s$, overlap and the set of pixels that form the centers of regions is the same as that which form the neighborhoods. Thus, UINTA filtering consists of two kinds of processes. One is the first-order *optimization process*, which computes updates for pixels based on their neighborhoods. The other second-order process causes updates of the neighborhoods based on the role of those pixels as centers in the previous iteration. The result of the filtering can be seen as the quest for the steady state of the combination of these two processes.

4 Nonparametric Multivariate Density Estimation

Entropy optimization entails the estimation of the PDFs of the RVs involved. For a $(2d+1)\times(2d+1)$ pixel neighborhood one must perform density estimation in a $(2d+1)^2$-dimensional space. This introduces the challenge of high-dimensional, scattered-data interpolation, even for modest sized image neighborhoods ($d = 2$ yields a 25D space). High-dimensional spaces are notoriously challenging for data analysis (regarded as the *the curse of dimensionality* [40, 37]) because they are so sparsely populated. Despite theoretical arguments suggesting that density estimation beyond a few dimensions is impractical, the empirical evidence from the statistics literature is more optimistic [37]. The results in this paper confirm that observation.

One of the advantages for UINTA is that the random vector $Z \equiv (X, Y)$ is *sphered*, by definition [37]. A sphered random vector is one for which the marginal PDFs of each individual RV have the same mean and variance. For UINTA, each marginal PDF is simply the grayscale intensity PDF, $p(x)$, of the image. The literature shows that sphered RVs lend themselves to more accurate density estimates [37, 40]. Also, UINTA relies on the neighborhoods in natural images having a lower-dimensional topology in the multi-dimensional feature space [24, 11]. This is also a general property for multivariate data [37]. Therefore,

locally (in the feature space) the PDFs of images are lower dimensional objects that lend themselves to better density estimation.

The literature shows Parzen windows as an effective nonparametric density estimation technique [31, 13]. The Parzen-window density estimate $p(z)$, in an n-dimensional space, is

$$p(z) \approx \frac{1}{|A|} \sum_{s_j \in A} G(z - z_j, \Psi) \tag{3}$$

where $|A|$ denotes the cardinality of the set A, and z_j is a shorthand for $z(s_j)$. The set A is a randomly selected subset of the sample space. UINTA chooses $G(z, \Psi)$ as the n-dimensional Gaussian

$$G(z, \Psi) = \frac{1}{(2\pi)^{n/2}|\Psi|^{1/2}} exp(-\frac{1}{2} z^T \Psi^{-1} z) \tag{4}$$

where Ψ is the $n \times n$ covariance matrix. The Gaussian kernel is not a unique choice, but it suits this application for several reasons: it is smooth, relatively efficient to compute (approximated by a look-up table), and entails a small number of free parameters. Having no a priori information on the structure of the PDFs, we choose an isotropic Gaussian of standard deviation σ, i.e. $\Psi = \sigma^2 I$, where I is the $n \times n$ identity matrix. A proper choice of the parameters σ and $|A|$, that determine the quality of the density estimate, is critical to UINTA's success. Section 5.2 discusses a strategy for computing the optimal parameter values.

4.1 A Stochastic Approximation for Entropy

Equation 1 gives entropy as an expectation of a RV. The approximation for entropy follows from the result that the sample mean converges, *almost surely*, to the expectation as the number of samples tends to infinity [46, 12]. Thus

$$E_p[\log p(Z)] \approx \frac{1}{|B|} \sum_{s_i \in B} \log p(z_i) \tag{5}$$

Equations 1, 3, and 5, give

$$h(Z) \approx -\frac{1}{|B|} \sum_{s_i \in B} \log \left[\frac{1}{|A|} \sum_{s_j \in A} G(z_i - z_j, \Psi) \right] \tag{6}$$

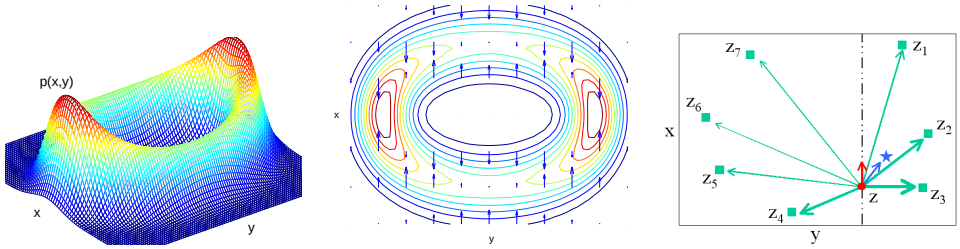
The set A, which generates the density estimate $p(z_i)$, should not contain the point s_i itself—because this biases the entropy estimates. The set B consists of the samples that will *move* at each iteration, and therefore it must consist of every sample in the image, i.e. $B = S$. The samples in set A are, typically, a small fraction of those in B, chosen at random. The relatively small cardinality of A has two important implications. First, it significantly reduces the computational cost for the entropy estimation, from $O(|B|^2)$ to $O(|A||B|)$. Second, because A is different for each element of B for each iteration, the entropy estimate, $h(Z)$, is stochastic. Hence, a gradient descent entropy optimization technique results in a stochastic-gradient algorithm [23, 19]. The stochastic-gradient effectively overcomes the effects of spurious local maxima introduced in the Parzen-window density estimate using finitely many samples [13, 46]. Thus, the proposed entropy-estimation scheme is important not only for computational efficiency but also for effective entropy minimization.

5 Conditional Entropy Minimization

Entropy minimization in UINTA relies on the derivative of the entropy with respect to the center-pixel value of the samples in B. Each pixel intensity in the image undergoes a gradient descent, based on the entropy of the conditional PDF estimated from A. The gradient descent for $x_i \equiv x(s_i)$ for each $s_i \in B$ is

$$\frac{\partial x_i}{\partial t} = -\frac{\partial h(X|Y = y_i)}{\partial x_i} \approx \frac{1}{|B|} \frac{\partial \log p(x_i|y_i)}{\partial x_i} = \frac{1}{|B|} \frac{\partial \log p(z_i)}{\partial x_i} \tag{7}$$

$$= -\frac{1}{|B|} \frac{\partial z_i}{\partial x_i} \sum_{s_j \in A} \frac{G(z_i - z_j, \Psi)}{\sum_{s_k \in A} G(z_i - z_k, \Psi)} \Psi^{-1}(z_i - z_j) \tag{8}$$



(a) (b) (c)

Figure 2: (a) An example 2D PDF, $p(x, y)$, on feature space, $< x, y >$. (b) A contour plot of the PDF depicts the forces (vertical blue arrows) that reduce entropy of the conditional PDFs $p(x|Y = y(s))$, as per Equation 7. (c) Attractive forces (green arrows: width $\equiv$ force magnitude) act on a sample (z:red circle) towards other samples (z_j:green squares) in the set A, as per Equation 8. The resultant force (blue arrow) acts towards the weighted mean (blue star), and the sample z moves based on its projection (vertical red arrow).

where $\partial z_i / \partial x_i$ is a projection operator that projects an n-dimensional vector z_i onto the dimension associated with the center pixel intensity x_i. Figure 2(a) gives an example PDF $p(x, y)$. Figure 2(b) depicts the forces that lead to the entropy reduction of the conditional PDF $p(x|Y = y(s))$, as seen in Equation 7. Figure 2(c) depicts the attractive forces acting on a sample towards samples in the set A, as seen in Equation 8, and the projection of the resultant force that actually moves the sample.

If we choose a $\sigma > 0$, the entropy for a finite set of samples is always bounded. Because we perform a (projected) gradient descent on a bounded energy function, the process converges (for sufficiently small time steps). Indeed, analysis of simple examples shows the existence of nontrivial steady states (e.g. an image which is a discrete sampling of a linear function $f(x, y)$). Empirical evidence, using real and synthetic images in Section 6, shows that the filtering algorithm does sometimes converge to interesting results. However, for many applications, convergence is not the goal; as with many other iterative filtering strategies, several iterations of the gradient descent are sufficient for acceptable denoising.

5.1 Relationship To the Mean-Shift Procedure

The mean-shift procedure [15, 40, 8, 9] moves each sample in a feature space to a weighted average of other samples using a weighting scheme that is similar to Parzen windowing. This can also be viewed as moving samples uphill on a PDF. Comaniciu and Meer [9] propose an iterative mean-shift algorithm for image intensities (where the PDF does not change with iterations) that provides a mechanism for image segmentation. Each grayscale (or vector) pixel intensity is drawn toward a local maximum in the grayscale (or vector-valued) histogram.

This section shows how UINTA relates to the mean-shift procedure. We begin by establishing the relationship between the mean-shift procedure and gradient-descent entropy minimization. Consider, as an example, a gradient descent on the entropy of the grayscale pixel intensities. This gives

$$\frac{\partial x_i}{\partial t} = -\lambda \frac{\partial h(X)}{\partial x_i} \approx -\frac{\lambda}{|B|} \sum_{s_j \in A} \frac{G(x_i - x_j, \Psi)}{\sum_{s_k \in A} G(x_i - x_k, \Psi)} \Psi^{-1}(x_i - x_j) \tag{9}$$

Finite forward differences $x^{m+1}(s) = x^m(s) - \lambda \partial h / \partial x^m(s)$ with a time step $\lambda = |B|\sigma^2$ give

$$\begin{aligned} x_i^{m+1} &= x_i^m + \left[\frac{\sum_{s_j \in A} x_j^m G(x_i^m - x_j^m, \Psi)}{\sum_{s_k \in A} G(x_i^m - x_k^m, \Psi)} - x_i^m \right] \\ &= \sum_{s_j \in A} x_j^m W(x_i^m, x_j^m, A, \Psi) = \sum_{s_j \in A} W_j x_j^m \end{aligned} \tag{10}$$

Thus each new pixel value is a weighted average of a selection of pixel values from the previous iteration with weights $W_j > 0$ such that $\sum_j W_j = 1$. This is exactly the mean-shift update proposed by Fukunaga [15]. Note that here the PDFs on which the samples climb get updated after every iteration. Thus the mean-shift algorithm is a gradient descent on the entropy associated with the grayscale intensities of an image. We observe that samples $x(s)$ are being attracted to every other sample, with a weighting term that diminishes with the distance between the two samples. The UINTA updates have the same form, except that the

weights are influenced not only by the distances/similarities between intensities $x(s)$ but also by the distances/similarities between the neighborhoods $y(s)$. That is, pixels in the image with similar neighborhoods have a relatively larger impact on the weighted means that drive the updates of the center pixels.

5.2 Implementation Issues

The UINTA algorithm as presented in previous sections presents a number of significant engineering questions which are crucial to effectiveness of the algorithm. This section discusses some of these issues and the proposed solutions at a somewhat high level; the Appendix covers these issues in more detail.

The first issue is the selection of the scale or size of the Parzen window (the standard deviation of the Gaussian). The Parzen-window density estimate, using a finite number of samples, shows a great deal of sensitivity for different values of σ. Thus, the particular choice of the standard deviation σ, and thereby Ψ, for the Gaussian in Equation 3, is a crucial factor that determines the behavior of the entire process of entropy optimization. Furthermore, this choice is related to the sample size $|A|$ in the stochastic approximation. For a particular choice of $|A|$, we propose to use the σ that minimizes the joint entropy, which we will call the *optimal* scale for a data set. This can be determined automatically at each iteration in the UINTA processing. Our experiments show that for sufficiently large $|A|$ the entropy estimates and optimal scale are virtually constant, and thus $|A|$ can also be generated directly from the input data.

The second issue is the choice of stopping criterion. For this we refer to the vast literature on nonlinear methods, which presents an array of choices, which are discussed in more detail in the Appendix. For some of the simpler examples in this paper steady-state results are quite good. However, for more complicated images (e.g. real-world images) we choose the number of iterations empirically based on subjective impressions of the quality of the results.

Another issue is the shape of the image neighborhoods. The square neighborhoods described in Section 3.2 show anisotropic artifacts, and favor features that are aligned with the cardinal directions. To obtain isotropic filtering results we use a metric in the feature space that controls the influence of each neighborhood pixel so that the resulting *mask* is more rotationally symmetric. In this way directions in the feature space corresponding to corners of neighborhood collapse so that they do not influence the filtering. A similar strategy enables us to handle image boundaries without distorting the statistics of the image. That is, pixels at image boundaries rely on the statistics in lower-dimensional subspaces corresponding to the set of neighborhood pixels lying within the input image.

Finally, the issue of sample selection for the set A also influences the behavior of the filter. While a uniform sampling over the image produces acceptable results for many images, our experiments have shown that the processing of some images (particularly those with spatially inhomogeneous image statistics) benefit from a *biased* sampling strategy, which estimates the PDF using samples that lie nearby the pixel being processed. We have found that a Gaussian distribution (centered at the pixel in question) works quite well, and that the results are not particularly sensitive to the size or scale of this sampling function.

6 Experiments and Results

This Section gives the results of UINTA filtering on real and synthetic images and analyzes the behavior of UINTA on the same. Parzen windowing in all of the examples uses, unless otherwise stated, a uniform random sampling of the image domain with 500 samples (i.e. $|A| = 500$), as explained in Appendix E. The noise in the synthetic-image examples is, unless otherwise stated, additive, zero mean, independent, and Gaussian. Furthermore, the amount of noise is high enough that thresholding the noisy image can never yield the noiseless image. Note that UINTA, in its formulation, does *not* assume any particular distribution on the noise. Because of interactions of neighborhoods from one iteration to the next, the

time step $\lambda = |B|\sigma^2$ can lead to oscillations in the results. We have found that a time step of $\lambda = |B|\sigma^2/\sqrt{n}$ alleviates this effect. The size of the Parzen window σ is recomputed after each iteration to minimize the entropy of the processed image. The UINTA implementation in this paper relies on the Insight Toolkit (ITK) [20].

Figure 3 shows the result of 11 iterations of UINTA on the *Lena* image with spatially local sampling (explained in Appendix E). The algorithm preserves and enhances fine structures, such as strands of hair or feathers in the hat, while removing random noise. The results are noticeably better than any of those obtained using other methods shown in Figure 1. A relatively small number of iterations produce subjectively good results for this image—further processing oversimplifies the image and removes significant details.

The fingerprint image in Figure 4 shows another example of the structure-enhancing tendencies of UINTA. UINTA enhances the contrast of the light and dark lines without

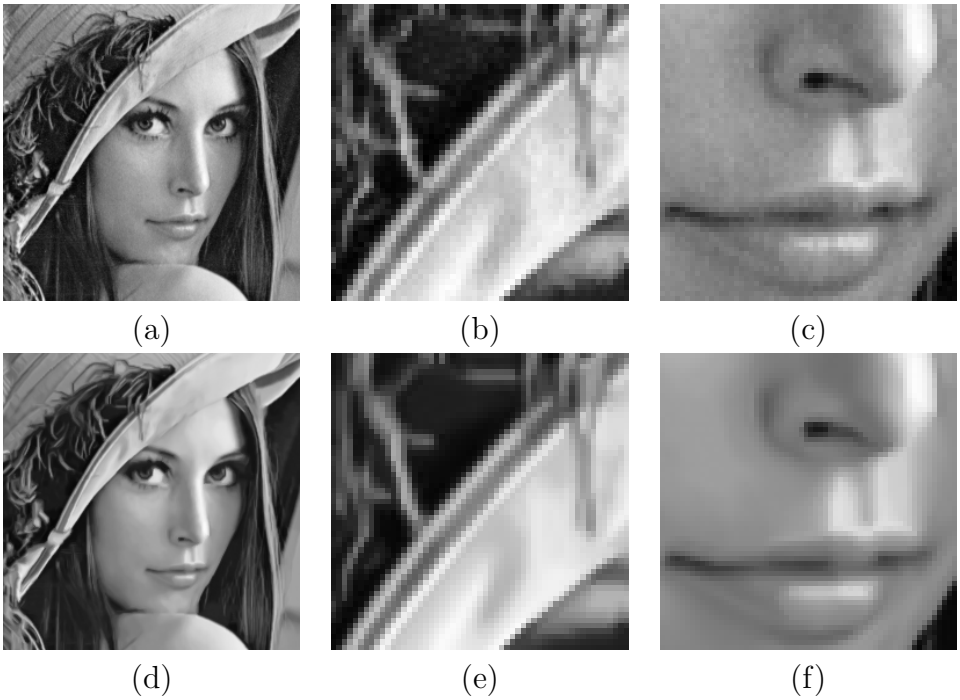


Figure 3: (a) Noisy image (256×256). (d) Filtered image. (b)-(c) and (e)-(f) show magnified portions of the noisy and filtered images, respectively. The intensities in each top-bottom pair are scaled to reflect the same contrast.

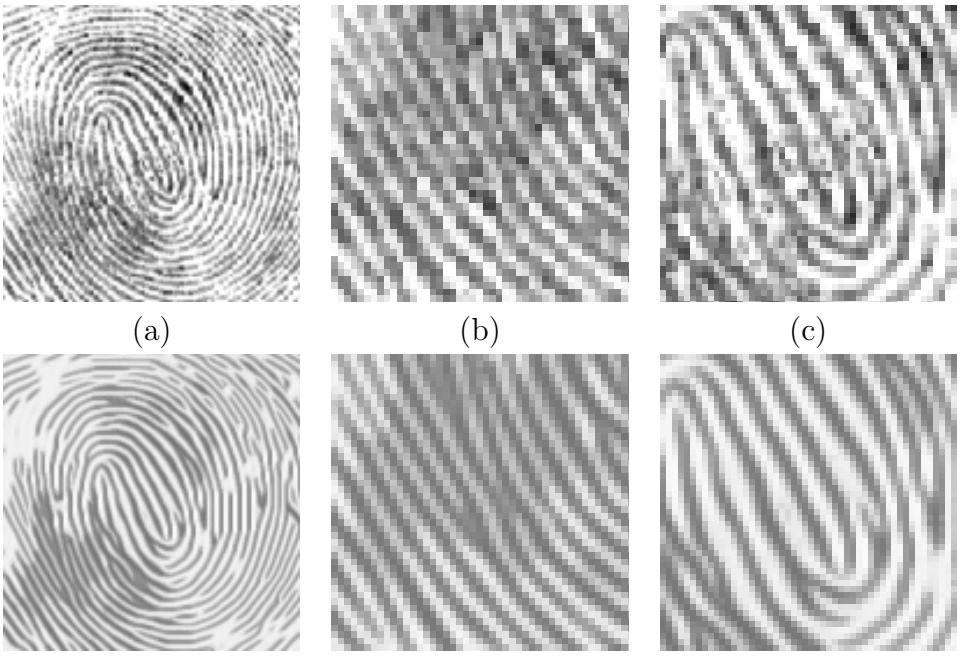


Figure 4: (a) Noisy image (128×128). (d) Filtered image. (b)-(c) and (e)-(f) show magnified portions of the noisy and filtered images, respectively.

significant shrinkage. UINTA performs a kind of multidimensional classification of image neighborhoods—therefore features in the the top-left are lost because they resemble background more than ridges. Figure 5 presents the results of other denoising strategies for visual comparison with UINTA. We see that the piece-wise smooth image models associated with anisotropic smoothing, bilateral filtering, and curvature flow (Figure 5(a)-(c)) are clearly inappropriate for the this image. A mean-shift procedure (Figure 5(d)) on image intensities (with the PDF not changing with iterations) yields a thresholded image retaining most of the noise. Weickert's coherence enhancing filter [49] (which is as well suited to this image as virtually any other) does not succeed in retaining or enhancing the light-dark contrast boundaries, and yet it forces elongated structures to grow or connect (Figure 5(e)). Thus, UINTA (Figure 5(f)) appears to remain more faithful to the underlying data.

Figure 6 shows the results of processing an MRI image of a human head for 8 iterations. This example employs the local sampling strategy (explained in Appendix E), and shows

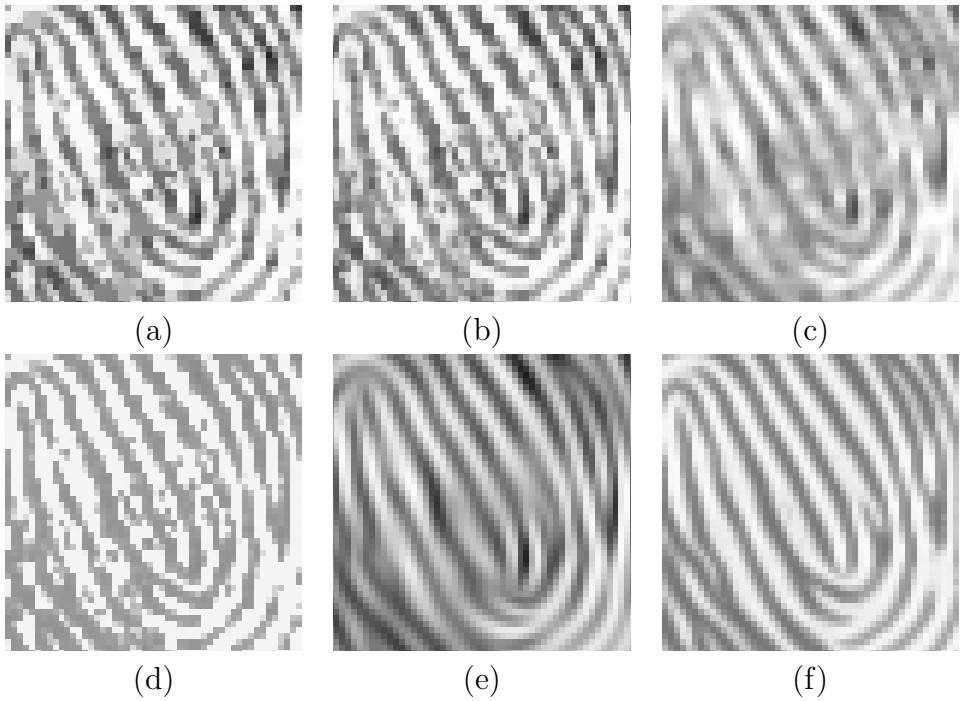


Figure 5: Noisy image grayscale values range:0-255. Results with (a) Anisotropic diffusion (K=0.45 grayscale values, 99 iterations) (b) Bilateral filtering (σ_{domain}=3 pixels, σ_{range}=15 grayscale values) (c) Curvature flow (time step=0.2, 5 iterations) (d) Mean shift on intensities (σ_{range}=5 grayscale values, 99 iterations) (e) Coherence enhancing diffusion (σ=0.1 pixels, ρ=2 pixels, α=0.0001, C=0.0001, 15 iterations) (f) UINTA (8 iterations)

the ability of UINTA to adapt to a variety of grayscale features. It enhances structure while removing noise, without imposing a piecewise constant intensity profile. As with the *Lena* example, more iterations tend to slowly erode important features.

Figures 7 gives a denoising example involving large amounts of noise. The checks are 4×4 pixels in size and the UINTA neighborhoods are 5×5 pixels. UINTA restores all of the edges and the corners and the image boundaries show no signs of artifacts. Figure 10(a) shows that the RMS error (root of the mean squared difference between pixel intensities in the filtered image and the noiseless image) decreases by 90 percent. Figure 13(c) shows that the joint entropy and the Parzen window size, σ, decrease monotonically as the filtering progresses. For this example, a multi-threaded implementation takes roughly 1 hour for 100 iterations with a pair of Intel®Xeon$^{\text{TM}}$2.66 GHz Pentium 4 processors (shared memory).

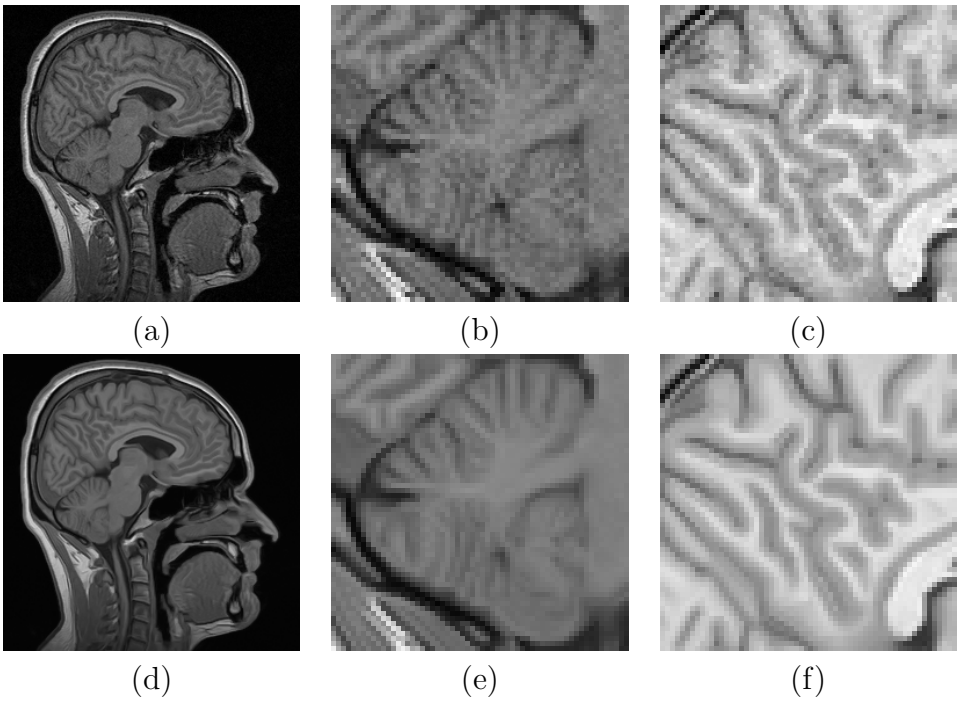


Figure 6: (a) Noisy image (256×256). (d) Filtered image. (b)-(c) and (e)-(f) show magnified portions of the noisy and filtered images, respectively. The intensities in each top-bottom pair are scaled to reflect the same contrast.

Figure 8 shows the results of applying UINTA to another corrupted binary image. The smoothness of the resulting circle boundary demonstrates the effectiveness of UINTA in preserving rotational invariance, as explained in Appendix C. The edges of the square are also well restored, but, unlike the checkerboard example, the corners are rounded. The presence of many similar corners in the checkerboard image form a well defined pattern in feature space, whereas the corners of the square are unique features (in that image) and hence UINTA treats them more like noise—those points in the feature space are attracted to more prevalent (less curved) features. Figure 9 shows the application of 15 iterations of UINTA to an image of hand-drawn curves (with noise). UINTA learns the pattern of black-on-white curves and forces the image to adhere to this pattern. However, UINTA does make mistakes when curves become too close, exhibit a very sharp bend, or when the noise introduces ambiguous gaps.

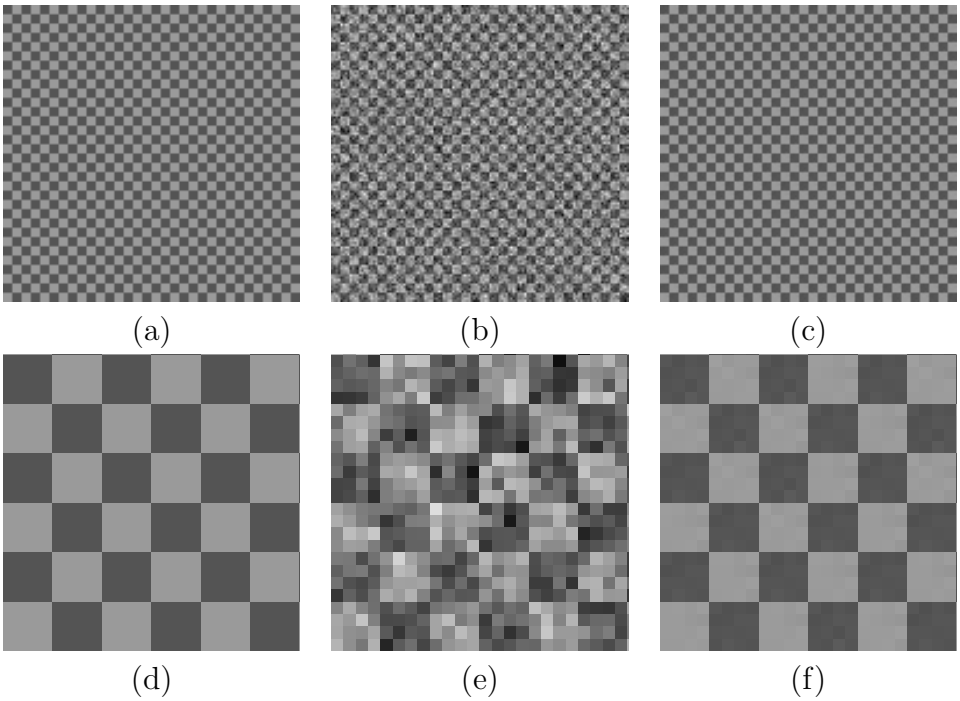


Figure 7: (a) Noiseless image (128×128). (b) Noisy image. (c) Filtered image (steady state). (d)-(f) show the magnified bottom-left corner of images (a)-(c) respectively.

To study the effect of the neighborhood size on UINTA we performed filtering with different neighborhood sizes on the checkerboard image from Figure 7, but with significantly more noise. Figure 10(b) shows comparisons of UINTA's performance with three different region sizes, i.e. 3×3, 5×5 and 7×7, that reflect the advantage of larger neighborhoods. For higher levels of noise, we find that larger neighborhoods are able to better discern patterns

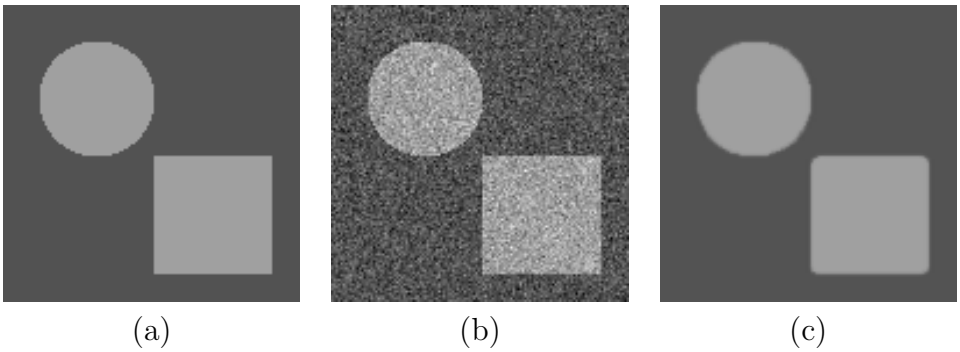


Figure 8: (a) Noiseless image (128×128). (b) Noisy image. (c) Filtered image (steady state).

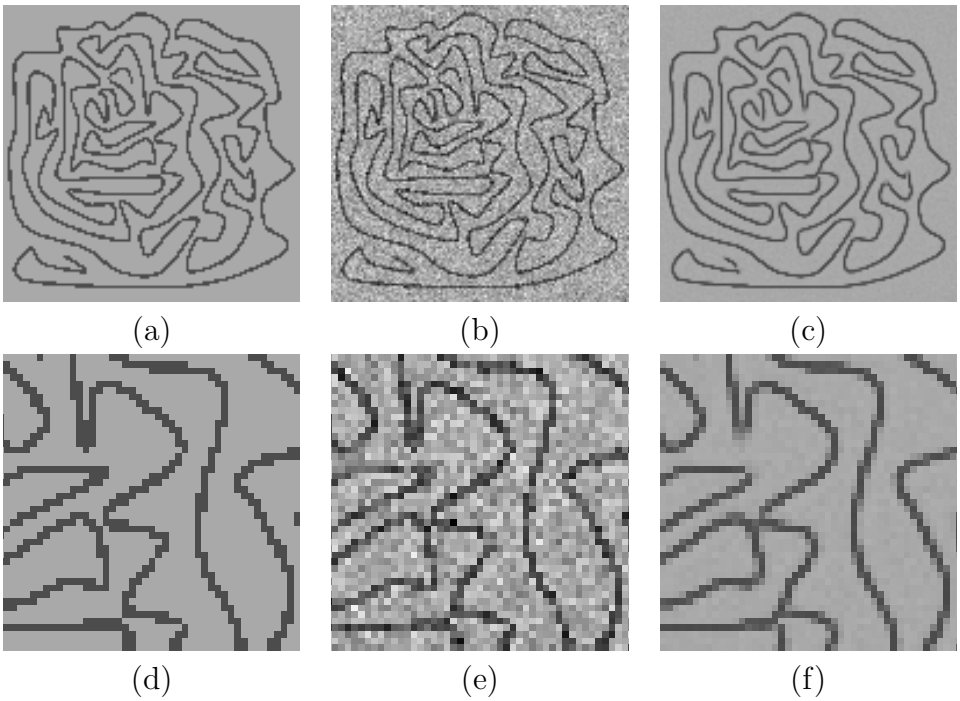
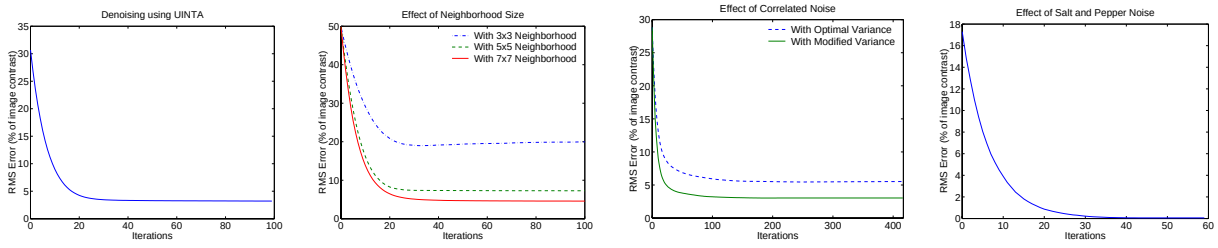


Figure 9: (a) Original image (128×128). (b) Noisy image. (c) Filtered image. (d)-(f) show the magnified portions of images (a)-(c) respectively.

in image regions and yield superior denoised images.

The UINTA algorithm is effective for removing various kinds of noise. UINTA filtering of the checkerboard image with correlated noise (gotten by applying a low-pass filter to zero-mean, independent, Gaussian noise) shows a significant improvement in RMS error



(a) (b) (c) (d)

Figure 10: RMS errors while denoising the checkerboard, expressed as a percentage of image contrast in the noiseless binary checkerboard image (difference between the two intensities), (a) for the example in Figure 7, (b) with significantly more noise and using different neighborhood sizes, (c) with correlated noise, and (d) with salt and pepper noise.

(see Figure 10(c)), but the reduction in RMS error is not as good as in the examples with uncorrelated noise. Correlated noise modifies the PDF $p(x, y)$ by creating more local maxima and thereby *fractures* the manifold associated with the original data. The strategy for automatically choosing the Parzen window size, σ, together with the joint entropy minimization scheme, is unable to remove these new feature-space structures. We can verify this by artificially increasing the size of the Parzen window. Multiplying the *optimal* σ by 2 gives a lower RMS error (see Figure 10(c)). This example brings out a weakness in the UINTA filter and our choice of a single isotropic Parzen window—an area of possible improvement. Experiments also show that UINTA is also very effective at removing replacement noise (e.g. salt and pepper noise), a somewhat easier denoising task (see Figure 10(d)).

Figure 11 shows the results of many iterations of UINTA on the hand-drawn image of Figure 9(a). UINTA has no explicit model of geometry and yet it gradually smooths out the kinks in these curves producing progressively simpler geometric structures. The joint entropy of straighter curves is lower, because of reduced variability in the associated image neighborhoods. The result is similar to that of curvature-reducing geometric flows [38, 30]. This example shows how statistics of image neighborhoods describe fundamental aspects of image geometry.

(a) (b) (c)

Figure 11: (a) Original image (128×128). (b) and (c) show filtered images after 100 and 200 iterations, respectively.

7 Conclusions and Discussion

UINTA is a novel, unsupervised, information-theoretic, adaptive filter that improves the predictability of pixel intensities from the intensities in the neighborhoods by decreasing the joint entropy. UINTA can preserve and enhance structures in a way that resembles many nonlinear, variational filters, but does so without any explicit geometric model. Because it is nonparametric, it can adapt to the statistics of the input image, and therefore it applies quite readily to new applications with very little parameter tuning.

The stochastic gradient-descent algorithm for minimizing joint entropy entails density estimation in high-dimensional spaces, and relies on Parzen windowing with automatic parameter selection. In order to be effective for image processing the UINTA algorithm operates with a feature-space metric that preserves rotational symmetry (see Appendix C) and allows for boundary conditions (see Appendix D). The UINTA algorithm is a generalization of the mean-shift classification algorithm [8] that conditions the distribution based on the pixel neighborhood. Results show that the statistics of image neighborhoods are sufficiently regular for reliable image denoising.

Despite these promising results, this paper presents only a preliminary implementation that could benefit from some engineering advances. For instance, the method of density estimation with single-scale isotropic Parzen-window kernels is clearly insufficient for all situations, and it is reasonable that kernels be chosen adaptively to accommodate the signal and/or noise. The computation times for the implementation are impractical for most applications, and improving the computational scheme is an important area of future work.

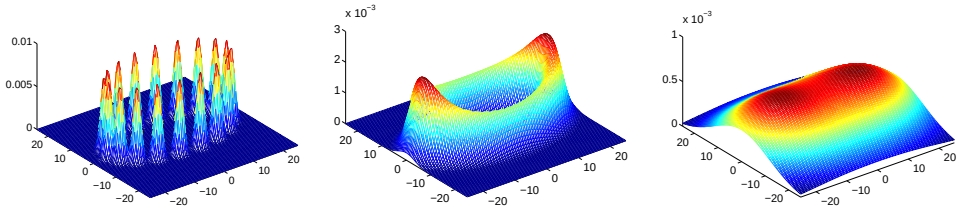
The implications of the empirical results in this paper are significant. They show that it is possible to construct nonparametric density estimations in the very high dimensional spaces of image neighborhoods. These results also suggest that the statistical structure in these spaces capture important geometric properties of images. The UINTA formulation also generalizes in several different ways. All of the mathematics, statistics, and engineering

in this paper are appropriate for higher-dimensional image domains and vector-valued data. The challenge is the increase in computation time, which is already quite significant. The same scheme could easily apply to other image representations, such as image pyramids, wavelets, or local geometric features.

Appendix

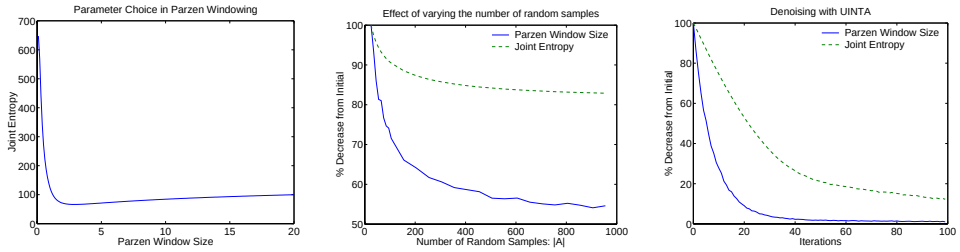
A Automatic Scale Selection for Parzen Windowing

Figure 12 shows that Parzen windowing, using a finite number of samples, is very sensitive to the value of σ [13]. Many algorithms/applications with low dimensional features spaces (e.g. 2 or 3) operate by manually tuning the scale parameter. However, because UINTA relies on a sparsely populated high dimensional space, it is very difficult to manually find values for σ that properly "connect" the data without excessively smoothing the PDF. Also, UINTA being iterative and dynamic, the best scale parameter changes every iteration. UINTA finds σ via a data-driven approach. Because the goal is to minimize joint entropy, a logical choice is to choose a value for σ that minimizes the same. Figure 13(a) confirms the existence of a unique minimum. Figure 13(b) shows that the choice of σ is *not* sensitive to the value of $|A|$ for sufficiently large $|A|$, thereby enabling UINTA to automatically fix $|A|$ to an appropriate value before the filtering begins.



(a) (b) (c)

Figure 12: (a), (b), and (c) show the drastic changes in the Parzen-window density estimates using isotropic Gaussians with $\sigma = 1$, $\sigma = 3$, and $\sigma = 10$, respectively.



(a) (b) (c)

Figure 13: (a) $h(X,Y)$ vs σ (for the *Lena* image in Figure 3(a), $|A| = 500$). (b) $h(X,Y)$ and σ (for the *Lena* image), are almost unaffected for $|A| > 500$. To give smoother curves, each measurement, for a particular $|A|$, is averaged over 5 different random sets A. (c) $h(X,Y)$ and σ (for the checkerboard denoising example in Figure 7).

We have implemented both differential (Newton's method) and discrete (Fibonacci search [34]) methods, and both offer acceptable results. Figure 13(c) depicts the decreasing trend for σ as the filtering progresses, which is common to every example and is consistent with UINTA's entropy-reducing action bringing samples closer in the feature space.

B Stopping Criteria

Like many iterative filtering strategies, the steady states of UINTA can produce simple images that do not adequately reflect important structures in the input image. There are many options for stopping criteria. One possibility is to use an information-theoretic choice based on $h(X,Y)$ or σ, both of which quantify the complexity in image neighborhoods. Figure 13(c) shows that both decrease monotonically. Hence, a stopping rule could be based on their absolute values, values relative to the input, or relative change between iterations.

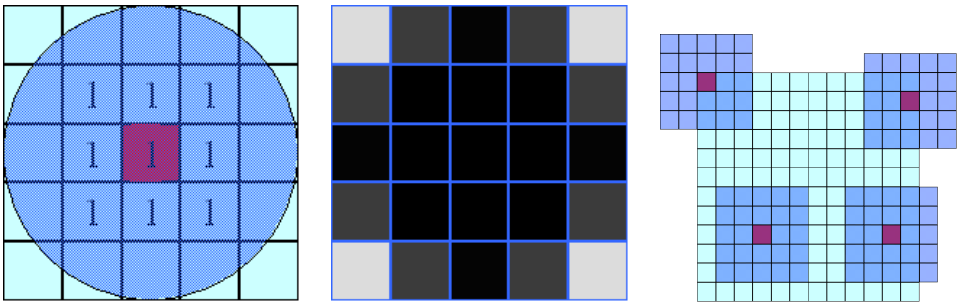
Another approach is to rely on the knowledge of the noise level in the input image. In this case UINTA could terminate when the residual (RMS difference between input and output) equals the noise level. Lastly, because UINTA is a filtering process on the image, termination can be based on visual inspection of images in the filtered image sequence.

An alternative is to use UINTA as part of a reconstruction process where entropy mini-

mization acts as a prior that is combined with an image-data or fidelity term. In this case UINTA would run to a steady state, and rather than a stopping criterion one must choose the relative weights of the prior and data, a so-called *meta parameter*. If the noise level is known, one can avoid the meta parameter and treat residual magnitude as a constraint [36].

C Rotational Invariance

Rotational invariance does not follow from UINTA's formulation, because the samples are taken on a rectilinear grid. Square neighborhoods generate results with artifacts exhibiting preferences for grid-aligned features. A solution is to weight the intensities, making neighborhoods more isotropic. UINTA incorporates such fuzzy weights by using an anisotropic feature-space distance metric, $\|z\|_M = \sqrt{z^T M z}$, where M is a diagonal matrix. The diagonal elements, $m_1, \ldots, m_n$, are the appropriate weights on the influence of the neighbors on the center pixel. To have a sphered RV Z, (that aids in density estimation; Section 4), we require the weights to be somewhat homogeneous. Figure 14(a)-(b) shows a disk-shaped mask that achieves this balance. The intensities near the center are unchanged ($m_i = 1$) while the intensities near the corners are weighted by the fraction ($m_i < 1$) of the pixel area overlapping with a hypothetical circle touching all four sides of the square neighborhood. The proposed



(a) (b) (c)

Figure 14: (a) Preserving rotational invariance via a disc-shaped mask for a 5×5 neighborhood. Intensities at pixels marked 1 remain unchanged. Other intensities are weighted by the fraction of the pixel area overlapping the circle. (b) The resulting weights (in shades of gray: black $\equiv 1$, white $\equiv 0$). (c) Anisotropic neighborhoods at boundaries.

isotropic mask is a grayscale version of the DUDE [50] strategy of using a binary disc-shaped mask for discrete (half-toned) images. Note that scaling the center-pixel intensity more than its neighbors leads to an elongated space (X', Y') along X'—in the limit, when all neighbors are weighted zero, leading to a thresholding as in the mean-shift algorithm [9].

D Anisotropic Neighborhoods at Boundaries

Typical image boundary conditions, e.g. replicating pixels or toroidal topologies, can produce neighborhoods distorting the feature-space statistics. UINTA handles boundary neighborhoods, using a strategy similar to that in Appendix C, by collapsing the feature space along the dimensions corresponding to the neighbors falling outside the image. UINTA crops the square regions crossing image boundaries and processes them in the lower-dimensional subspace. This strategy results in important modifications in two stages of UINTA. First, the cropped intensity vectors take part in a mean-shift process reducing entropies of the conditional PDFs in the particular subspace where they reside. Second, UINTA chooses the Parzen window size, σ, based only on the regions lying completely in the image interior.

E Selecting Random Samples

Parzen-window density estimation entails the selection of a set of samples belonging to the density in question, as seen in Section 4. Nominally, this set comprises random samples drawn from a uniform PDF on the sample space. The strategy works well if the image statistics are more or less uniform over the domain, e.g. the fingerprint image in Figure 4. However, it fails if the statistics of different parts of the image are diverse, e.g. the *Lena* image in Figure 3. This is because distant parts, for such an image, produce samples lying in distant regions of the feature space. To alleviate this problem we estimate $p(z(s_i))$, by selecting random samples s_j in a way that favors nearby points in the image domain—using

a Gaussian distribution centered at s_i with a relatively small standard deviation (10 pixels). This strategy is appropriate for images having locally consistent neighborhood statistics.

References

[1] D. Adalsteinsson and J. Sethian. A fast level set method for propagating interfaces. *J. Comput. Phys.*, 118(2):269–277, 1995.

[2] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Trans. Image Processing*, 10(8):1200–1211, 2001.

[3] D. Barash. A fundamental relationship between bilateral filtering, adaptive smoothing, and the nonlinear diffusion equation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(6):844–847, 2002.

[4] M. J. Black, G. Sapiro, D. Marimont, and D. Heeger. Robust anisotropic diffusion. *IEEE Trans. Image Processing*, 7(3):421–432, 1998.

[5] B. Buck and V. Macaulay, editors. *Maximum Entropy in Action*. Clarendon Press, Oxford, 1991.

[6] K. R. Castleman. *Digital image processing*. Prentice Hall Press, 1996.

[7] T. Chan, J. Shen, and L. Vese. Variational pde models in image processing, 2003.

[8] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):790–799, 1995.

[9] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[11] V. de Silva and G. Carlsson. Topological estimation using witness complexes. *Eurographics Symposium on Point-Based Graphics*, 2004.

[12] E. R. Dougherty. *Random Processes for Image and Signal Processing*. Wiley, 1998.

[13] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, 2001.

[14] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the Int. Conf. Computer Vision*, page 1033, 1999.

[15] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Info. Theory*, 21(1):32–40, 1975.

[16] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6:721–741, 1984.

[17] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice Hall, 2001.

[18] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1994.

[19] S. Haykin, editor. *Unsupervised Adaptive Filtering*. Wiley, 2000.

[20] http://www.itk.org. NLM Insight Segmentation and Registration Toolkit.

[21] J. Huang and D. Mumford. Statistics of natural images and models. *Proceedings of the Int. Conf. Computer Vision*, 1:541–547, 1999.

[22] A. K. Jain. *Fundamentals of digital image processing*. Prentice-Hall, Inc., 1989.

[23] B. Kosko. *Neural Networks and Fuzzy Systems*. Prentice-Hall, Inc., 1992.

[24] A. B. Lee, K. S. Pedersen, and D. Mumford. The nonlinear statistics of high-contrast patches in natural images. *Int. J. Comput. Vision*, 54(1-3):83–103, 2003.

[25] M. Lysaker, S. Osher, and X. Tai. Noise removal using smoothed normals and surface fitting. *UCLA Technical Report*, 2003.

[26] J. Miller and C. Stewart. Muse: Robust surface fitting using unbiased scale estimates. In *Proc. of the Conf. Computer Vision and Pattern Recog.*, pages 300–306, June 1996.

[27] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Com. of Pure and Applied Math.*, 42:577–685, 1989.

[28] B. K. Natarajan. Filtering random noise from deterministic signals via data compression. *IEEE Trans. Signal Processing*, 43:2595–2605, 1995.

[29] K. N. Nordstrom. Biased anisotropic diffusion: a unified regularization and diffusion approach to edge detection. *Image Vision Comput.*, 8(4):318–327, 1990.

[30] S. Osher and R. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*. Springer, 2003.

[31] E. Parzen. On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[32] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(7):629–639, July 1990.

[33] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C*. Cambridge Univ. Press, 1992.

[34] S. S. Rao. *Engineering Optimization, Theory and Practice*. Wiley, 1996.

[35] B. M. Romeny, editor. *Geometry-Driven Diffusion in Computer Vision*. Kluwer Academic Publishers, 1994.

[36] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1-4):259–268, 1992.

[37] D. W. Scott. *Multivariate Density Estimation*. Wiley, 1992.

[38] J. Sethian. *Level Set Methods and Fast Marching Methods*. Cambridge Univ. Press, 1999.

[39] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

[40] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[41] W. Snyder, Y. Han, G. Bilbro, R. Whitaker, and S. Pizer. Image relaxation: Restoration and feature extraction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(6):620–624, 1995.

[42] M. Studeny and J. Vejnarova. The multiinformation function as a tool for measuring stochastic dependence. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 261–297. Kluwer Academic Publishers, 1998.

[43] T. Tasdizen, R. Whitaker, P. Burchard, and S. Osher. Geometric surface processing via normal maps. *ACM Trans. on Graphics*, 2003.

[44] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the Sixth Int. Conf. Computer Vision*, page 839. IEEE Computer Society, 1998.

[45] L. A. Vese and S. J. Osher. Modeling textures with total variation minimization and oscillating patterns in image processing. *J. Sci. Comput.*, 19:553–572, 2003.

[46] P. Viola and W. M. Wells, III. Alignment by maximization of mutual information. In *Proceedings of the Fifth Int. Conf. Computer Vision*, pages 16–23, 1995.

[47] L. Wei and M. Levoy. Order-independent texture synthesis. *Stanford University Computer Science Department Technical Report TR-2002-01*, 2002.

[48] J. Weickert. *Anisotropic Diffusion in Image Processing*. Teubner-Verlag, 1998.

[49] J. Weickert. Coherence-enhancing diffusion filtering. *Int. J. Computer Vision*, 31:111–127, April 1999.

[50] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. Weinberger. Universal discrete denoising: Known channel. *HP Labs Technical Report HPL-2003-29*, 2003.