A DISCRETE UNIVERSAL DENOISER AND ITS APPLICATION TO BINARY IMAGES

Erik Ordentlich Gadiel Seroussi Sergio Verdú Marcelo Weinberger Tsachy Weissman

1. INTRODUCTION

In a recent work [1], the authors introduced a *discrete universal denoiser* (DUDE) for recovering a signal with finite-valued components corrupted by finite-valued, uncorrelated noise. The DUDE is asymptotically optimal and universal, in the sense of asymptotically achieving, without access to any information on the statistics of the clean signal, the same performance as the best denoiser that does have access to such information. It is also practical, and can be implemented in low complexity. In this work, we extend the definition of the DUDE to two-dimensionally indexed data, and present results of an implementation of the scheme for binary images. Section 2 presents the problem setting, definitions, and notation used throughout the paper. Section 3 describes the DUDE for two-dimensional data (this description readily extends to higher dimensions). Section 4 presents theoretical performance guarantees establishing the DUDE's asymptotic optimality. The denoiser assumes a particularly simple form for binary alphabets, which is presented in Section 5. Practical considerations in the implementation of the binary scheme are presented in Section 6, while experimental results of its application to noisy binary images are presented in Section 7. In the examples considered we find that the DUDE outperforms current popular schemes [2, 3] for binary image denosing. Finally, in Section 8 we discuss conclusions and directions for ongoing and future research.

2. PROBLEM SETTING AND NOTATION

Throughout we let $\mathcal{A}$, of size $|\mathcal{A}|=M$, denote the finite alphabet where the components of the clean, as well as those of the noise-corrupted image, take their values. We assume that the noiseless image, for which no statistical model is available, is corrupted by a *discrete memoryless channel* (DMC) characterized by a transition probability matrix $\mathbf{\Pi} = \{\Pi(a,b)\}_{a,b\in\mathcal{A}}$. That is to say that the noise components are statistically independent and the probability that the observed noisy symbol at a given location is b when the underlying clean symbol is a is given by $\Pi(a,b)$. Our assumption is that $\mathbf{\Pi}$ is invertible, which holds for channels arising in practice. We also assume a given *loss function* (fidelity criterion) $\Lambda : \mathcal{A}^2 \to [0,\infty)$, represented by a matrix $\mathbf{\Lambda} = \{\Lambda(i,j)\}_{i,j\in\mathcal{A}}$, where $\Lambda(i,j)$ is the loss incurred by estimating the symbol i with the symbol j.

The i-th component of a vector $\mathbf{u}$ will be denoted by u_i or $\mathbf{u}[i]$. For M-dimensional vectors $\mathbf{u}$ and $\mathbf{v}$, we denote by $\mathbf{u}\odot\mathbf{v}$ the component-wise product of the vectors, i.e., $(\mathbf{u}\odot\mathbf{v})[i] = \mathbf{u}[i]\mathbf{v}[i]$, $1 \le i \le M$. An *image* is a two-dimensional array over $\mathcal{A}$. Let $\mathbf{x} = \{x_i\}_{i\in\mathbb{N}^2}$ denote the (conceptually infinite) noise-free image and $\mathbf{z} = \{z_i\}_{i\in\mathbb{N}^2}$ its noise-corrupted version. Here, $\mathbb{N}$ is the set of positive integers, and we extend the indexing notations u_i and $u[i]$ to two-dimensional indices. For any $\mathcal{S} \subseteq \mathbb{N}^2$ we denote $x(\mathcal{S}) = \{x_i\}_{i\in\mathcal{S}}$ and $z(\mathcal{S}) = \{z_i\}_{i\in\mathcal{S}}$. Thus, $x(\mathcal{S})$ is a $|\mathcal{S}|$-dimensional vector with $\mathcal{A}$-valued components indexed by the elements of $\mathcal{S}$, and we denote by $\mathcal{A}^{\mathcal{S}}$ the set of all such vectors. For $m,n\in\mathbb{N}$ let $V_{m\times n}$ denote the $m\times n$ rectangle $\{(i_x,i_y)\in\mathbb{N}^2 : i_x \le m, i_y \le n\}$. To simplify notation, we shall write $x_{m\times n}$ for $x(V_{m\times n})$, $z_{m\times n}$ for $z(V_{m\times n})$, and $\mathcal{A}^{m\times n}$ for $\mathcal{A}^{V_{m\times n}}$. Finally, for $\mathcal{S}\subseteq\mathbb{N}^2$ and $i\in\mathbb{N}^2$ we let $\mathcal{S}+i = \{j+i : j\in\mathcal{S}\}$.

A $m\times n$ *image denoiser* is a mapping $\hat{X}^{m\times n} : \mathcal{A}^{m\times n}\to\mathcal{A}^{m\times n}$. For $x_{m\times n}, z_{m\times n}\in\mathcal{A}^{m\times n}$ we let $L_{\hat{X}^{m\times n}}(x_{m\times n}, z_{m\times n})$ denote the normalized denoising loss, as measured by Λ, of the image denoiser $\hat{X}^{m\times n}$ when the observed noisy image is $z_{m\times n}$ and the underlying one is $x_{m\times n}$, i.e.,

$$L_{\hat{X}^{m\times n}}(x_{m\times n}, z_{m\times n}) = \frac{1}{mn}\sum_{i\in V_{m\times n}} \Lambda(x_i, \hat{X}^{m\times n}(z_{m\times n})[i]),$$

with $\hat{X}^{m\times n}(z_{m\times n})[i]$ denoting the component of $\hat{X}^{m\times n}(z_{m\times n})$ at the i-th location.

A *neighborhood* is a subset of $\mathbb{Z}^2$ not containing the origin $(0,0)$ (the *center* of the neighborhood). If $\mathcal{S}$ is a neighborhood we shall refer to any element of $\mathcal{A}^{\mathcal{S}}$ as a $\mathcal{S}$-*context*, or simply a *context*. For $r\ge 0$ we let $\mathcal{B}_r$ denote the L_2 ball of radius r in $\mathbb{Z}^2$, centered at $(0,0)$, i.e., $\mathcal{B}_r = \{i\in\mathbb{Z}^2 : \|i\|_2 \le r\}$. $\mathcal{S}_r$ will denote the neighborhood $\mathcal{S}_r = \mathcal{B}_r \setminus (0,0)$.

For a neighborhood $\mathcal{S}\subseteq\mathbb{Z}^2$ and a $\mathcal{S}$-context $b(\mathcal{S})$ we let $\mathbf{m}(z_{m\times n}, b(\mathcal{S}))$ denote the M-dimensional column vector whose α-th component, $\mathbf{m}(z_{m\times n}, b(\mathcal{S}))[\alpha]$, $\alpha\in\mathcal{A}$, is given by

$$|\{i\in V_{m\times n} : \mathcal{S}+i\subseteq V_{m\times n}, z(\mathcal{S}+i) = b(\mathcal{S}), z_i = \alpha\}|.$$

In words, $\mathbf{m}(z_{m\times n}, b(\mathcal{S}))[\alpha]$ denotes the number of locations in the noisy image $z_{m\times n}$ where the symbol α appears (after appropriate translation) in context $b(\mathcal{S})$.

3. DESCRIPTION OF THE DENOISER

Let $\mathcal{S}\in\mathbb{Z}^2$ be a neighborhood, and let $\boldsymbol{\lambda}_a$ and $\boldsymbol{\pi}_a$ denote the ath columns of $\mathbf{\Lambda}$ and $\mathbf{\Pi}$, respectively, for all $a\in\mathcal{A}$.

S. Verdú is with the Dept. of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (verdu@princeton.edu). The other authors are with Hewlett-Packard Laboratories, Palo Alto, CA 94304 USA (tsachyw@hpl.hp.com; eord@hpl.hp.com; seroussi@hpl.hp.com; marcelo@hpl.hp.com). T. Weissman is also with the Dept. of Statistics, Stanford University, Stanford, CA 94305 USA (tsachy@stat.stanford.edu).

0-7803-7750-8/03/$17.00 ©2003 IEEE.

The $m \times n$ image denoiser $\hat{X}_{\mathcal{S}}^{m\times n}$ is defined by

$$\hat{X}_{\mathcal{S}}^{m\times n}(z_{m\times n})[i] = \arg\min_{\hat{x}\in\mathcal{A}} \mathbf{m}^T(z_{m\times n}, z(\mathcal{S}+i))\cdot\mathbf{\Pi}^{-1}\cdot(\boldsymbol{\lambda}_{\hat{x}}\odot\boldsymbol{\pi}_{z_i}), \quad (1)$$

for locations i where $\mathcal{S}+i \subseteq V_{m\times n}$. The specific definition of the denoiser output for other locations $i \in V_{m\times n}$ will be inconsequential for the validity of the theoretical results below. For concreteness, we can assume that the value of any "out of bound" sample is set to some arbitrary constant from $\mathcal{A}$, so that the values in $z(\mathcal{S}+i)$ are well defined for all $i \in V_{m\times n}$. Notice that the definition of the denoiser assumes knowledge of the channel $\mathbf{\Pi}$. Ways to address this need in practical situations where a full description of the channel might not be available are discussed in Section 6. Finally, for given m, n, we let our $m \times n$ universal image denoiser be given by

$$\hat{X}_{\mathsf{univ}}^{m\times n} = \hat{X}_{\mathcal{S}_{r(m,n)}}^{m\times n}, \quad (2)$$

where $r(m,n) = g(\min\{m,n\})$ and g is an unboundedly increasing function such that $g(t)M^{g(t)} = o(t^{1/4})$, for example, $g(t) = c\log_M t$ with $c < \frac{1}{4}$.

4. ASYMPTOTIC OPTIMALITY

In this section we present theoretical results assessing the universal asymptotic optimality of the image denoiser $\hat{X}_{\mathsf{univ}}^{m\times n}$. Analogous results for one-dimensionally indexed signals were derived and proved in [1]. The proofs of the results below are similar in nature, though they require some additional ingredients that have no analogues in the one-dimensional case and will be detailed in [4].

A *sliding window denoiser* of radius r is one that determines the denoised value at a location i as a function of $z(\mathcal{B}_r + i)$. Let $D_r(x_{m\times n}, z_{m\times n})$ denote the *r-th order denoisability* of $(x_{m\times n}, z_{m\times n})$, defined by

$$\min_{f:\mathcal{A}^{\mathcal{B}_r}\to\mathcal{A}} \left[\frac{1}{mn}\sum_{i:\mathcal{B}_r+i\in V_{m\times n}} \Lambda\left(x_i, f(z(\mathcal{B}_r+i))\right)\right]. \quad (3)$$

This definition of $D_r(x_{m\times n}, z_{m\times n})$ can be interpreted as the denoising performance of a "genie-aided" scheme, allowed to select the best sliding-window denoiser of radius $\leq r$, based on knowledge of both the noisy *and the underlying noiseless image*. Note that most image denoisers applied in practice, such as median filters morphological operators , and context-dependent spatial operators (see, e.g., [2, 3]) are sliding-window denoisers, so the r-th order denoisability is a lower bound on the performance of all such schemes (for r large enough). Our main theoretical result guarantees that the image denoiser $\hat{X}_{\mathsf{univ}}^{m\times n}$, for large images, does essentially as well as this genie-aided scheme, regardless of the underlying noiseless image. The setting for Theorem 1 below is one where the image $\mathbf{x}$ is some arbitrary, unknown, deterministic element of $\mathcal{A}^{\mathbb{N}^2}$, while $\mathbf{Z}$ is the random field resulting when $\mathbf{x}$ is corrupted by the memoryless channel $\mathbf{\Pi}$.

Theorem 1 *For all $\mathbf{x} \in \mathcal{A}^{\mathbb{N}^2}$, with probability one,*

$$L_{\hat{X}_{\mathsf{univ}}^{m\times n}}(x_{m\times n}, Z_{m\times n}) - D_{r(m,n)}(x_{m\times n}, Z_{m\times n}) \to 0 \quad (4)$$

as $m, n \to \infty$.

Consider now a fully stochastic setting where the noiseless image $\mathbf{X}$ is assumed generated by a spatially stationary and ergodic source. Letting $\mathbf{P}_{X_{m\times n}}$ and $\mathbf{P}_{\mathbf{X}}$ denote, respectively, the distributions of $X_{m\times n}$ and $\mathbf{X}$, and denoting by $\mathcal{D}_{m\times n}$ the class of all $m \times n$ image denoisers, we define

$$\mathbb{D}(\mathbf{P}_{X_{m\times n}}, \mathbf{\Pi}) = \min_{\hat{X}^{m\times n}\in\mathcal{D}_{m\times n}} E\left[L_{\hat{X}^{m\times n}}(X_{m\times n}, Z_{m\times n})\right],$$

with expectation on the right side taken with respect to $\mathbf{P}_{X_{m\times n}}$ and the channel $\mathbf{\Pi}$. We further define

$$\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}) \stackrel{\triangle}{=} \inf_{m,n\geq 1} \mathbb{D}(\mathbf{P}_{X_{m\times n}}, \mathbf{\Pi}).$$

$\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi})$ is the optimal asymptotic image denoising performance attainable with full knowledge of the statistical characterization of the image. The following result guarantees that this performance is universally attained by $\hat{X}_{\mathsf{univ}}^{m\times n}$, without any knowledge of $\mathbf{P}_{\mathbf{X}}$.

Theorem 2 *For all spatially stationary and ergodic $\mathbf{X}$,*

(a) $\limsup_{m,n\to\infty} L_{\hat{X}_{\mathsf{univ}}^{m\times n}}(X_{m\times n}, Z_{m\times n}) \leq \mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi})$, *with probability one.*

(b) $\lim_{m,n\to\infty} E\left[L_{\hat{X}_{\mathsf{univ}}^{m\times n}}(X_{m\times n}, Z_{m\times n})\right] = \mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi})$.

Remark: Note that implicit in the statements of Theorems 1 and 2 are the results that the associated limits do not depend on the way in which m and n are sent to infinity.

5. THE BINARY IMAGE DENOISER FOR A BSC

Consider the case where $\mathcal{A}=\{0,1\}$ and the image is corrupted by a *binary symmetric channel* (BSC) with parameter δ. Thus, the observed binary value at each image location is the result of flipping the clean image sample with probability δ. In this case, and assuming $\delta < 1/2$ and Hamming loss, the denoiser in (1) simplifies to

$$\hat{X}_{\mathcal{S}}^{m\times n}(z_{m\times n})[i] = \begin{cases} z_i & \frac{\mathbf{m}(z_{m\times n}, z(\mathcal{S}+i))[z_i]}{\mathbf{m}(z_{m\times n}, z(\mathcal{S}+i))[\overline{z}_i]} \geq \frac{2\delta(1-\delta)}{(1-\delta)^2+\delta^2}, \\ \overline{z}_i & \text{otherwise,} \end{cases} \quad (5)$$

$\overline{z}_i$ denoting the binary complement of z_i. Namely, for each location in the noisy image, we count, among the locations sharing the same context, how many had the same value and how many had the opposite value. If the ratio of these counts is below $\frac{2\delta(1-\delta)}{(1-\delta)^2+\delta^2}$, then z_i is deemed in error and is flipped. Otherwise, it is left unchanged.

6. IMPLEMENTATION OF THE BINARY IMAGE DENOISER

The binary version of the DUDE was implemented in software. While the results of Theorems 1 and 2 establish the asymptotic optimality of the scheme, special care must be taken in the implementation design to enable good performance with practical, finite images. In particular, the growth of the neighborhood size as a function of the "order" (e.g., the radius r in Section 3) must be kept in check, to avoid dilution of the context statistics, one aspect of the more general concept of *model cost* [5]. In our implementation, a finer sequence of neighborhoods was adopted, rather than the sequence of center-less spheres $\mathcal{S}_r$ implied by the discussion in Section 3. Specifically, we order pairs $i = (i_x, i_y) \in \mathbb{Z}^2 \backslash \{(0,0)\}$ in increasing order of $\|i\|_2$, then in increasing order of $\|i\|_\infty$, then in increasing order of $|i_y|$, and so on. The initial sequence is $(-1,0)$, $(1,0)$, $(0,-1)$, $(0,1)$, $(-1,-1)$, $(-1,1)$, $(1,-1)$, $(1,1)$, $(-2,0)$,$(2,0)$, ... The neighborhood of order k, denoted $\mathcal{N}_k$, is defined to consist of the first k elements of this sequence. Notice that neighborhoods are kept as symmetric as possible, and that the sequence $\mathcal{N}_k$ is indeed a refinement of the sequence $\mathcal{S}_r$, with $|\mathcal{N}_k| = k$.

A neighborhood order k, and a channel parameter δ are provided to the program as parameters, together with an input (presumed noisy) image $z_{m\times n}$. The denoiser runs two passes over the image. In the first pass, statistics are collected for context patterns $z_{m\times n}(\mathcal{N}_k + i)$, $i \in V_{m\times n}$. For each context pattern, counts of occurrences of the values 0 and 1 at the center of the context are collected. The second pass uses these counts, together with the criterion in (5), to denoise the image samples.

As noted, the DUDE requires the values of the parameters δ and k in order to process an image. In the theoretical setting, the channel parameter δ is assumed known, and the context order is only loosely based on the image size (see discussion following Equation (2). In practice, the channel parameter is often unknown, and the specific policy used to choose the context order significantly affects denoising performance. Moreover, even with the benefit of several runs of the denoiser, it is not clear a priori how one would search for the optimal parameters, since the original noiseless image is not available to assess the performance of the denoiser for a given parameter setting. Fortunately, experimental results show that near-optimal values for both δ and k can be heuristically, but quite accurately derived as functions of observable parameters.

Given k, a good estimate of the channel parameter δ is given by

$$\min_c \min\{\mathbf{m}(z_{m\times n}, c)[0], \mathbf{m}(z_{m\times n}, c)[1]\},$$

the minimum taken over contexts $c \in \mathcal{A}^{\mathcal{N}_k}$ that occur with "sufficient frequency" in $z_{m\times n}$. The intuition behind this heuristic is that if the image is denoisable, then some significant context must exhibit skewed statistics, where the least probable symbol has a low count, thus "exposing" the outcomes of the BSC. Notice that this estimation of δ requires running just the first pass of the denoiser.

As for k, it was observed in experiments where the original noiseless image was available as a reference, that the value of k that minimizes the distortion of the denoised image $\hat{x}_{m\times n}$ relative to the original $x_{m\times n}$, also consistently minimizes the compressibility of $\hat{x}_{m\times n}$. This compressibility can be estimated from observable data by a practical implementation of a universal lossless compression scheme. A more detailed discussion of the issues involved in choosing the best value of k for a given finite data sequence can be found in [1].

		Channel parameter δ			
Image	Scheme	0.01	0.02	0.05	0.10
Shannon 1800×2160	DUDE	0.00096 k=11	0.0018 k=12	0.0041 k=12	0.0091 k=12
	median	0.00483	0.0057	0.0082	0.0141
	morpho.	0.00270	0.0039	0.0081	0.0161
Einstein 896×1160	DUDE	0.0035 k=18	0.0075 k=14[1]	0.0181 k=12[1]	0.0391 k=12[1]
	median	0.156	0.158	0.164	0.180
	morpho.	0.149	0.151	0.163	0.193

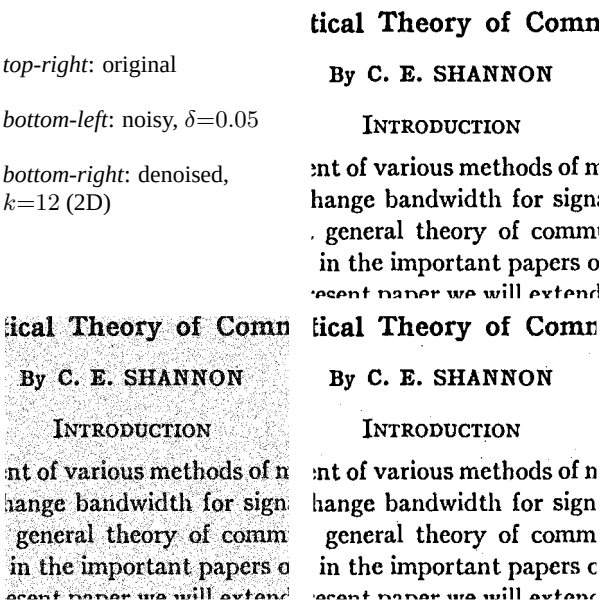
Table 1. Denoising results

The steps of empirically estimating δ and k might need to be iterated, as the estimate of one depends on the estimate of the other. In practice, however, it was observed that very few, if any, iterations are needed if one starts from a reasonable guess of the channel parameter. The best k is estimated given this guess, and from it a more accurate estimate of δ is obtained. In the majority of cases, no further iterations were needed. In some practical applications, a denoiser is offered as part of a toolkit used in an interactive environment. In those cases, the search for the best parameters δ and k could be aided by visual inspection.

7. EXPERIMENTAL RESULTS

We report on results of running the denoiser on two binary images. The first image is the first page from a scanned copy of Shannon's seminal paper [6], available in the publications data base of the IEEE Information Theory Society. The image was corrupted by running it through BSCs of various parameter values. The noisy image was then denoised with the DUDE, estimating the best parameters δ and k as outlined above. The results are shown in the upper portion of Table 1, which lists the normalized bit-error rate of the denoised image, relative to the original one. The table also shows results of denoising the same image with a 3×3 median filter [2], and a morphological filter [3] available under MATLAB. The results for the morphological filter are for the best ordering of the morphological open and close operations based on a 2×2 structural element, which was found to give the best performance. The results in the table show that the DUDE significantly outperforms the reference filters. Figure 1 shows corresponding portions of the noiseless, noisy, and DUDE-denoised images, respectively, for the experiment with $\delta = 0.05$ (the whole image is not shown due to space constraints).

[1] One-dimensional contexts of size k, consisting of $k/2$ samples to the left, and $k/2$ to the right of the denoised sample, were used in these cases to obtain the best results. While a two-dimensional context scheme obtains bit error-rates that come close to those reported, the visual quality of the denoised halftone was superior with the one-dimensional contexts.

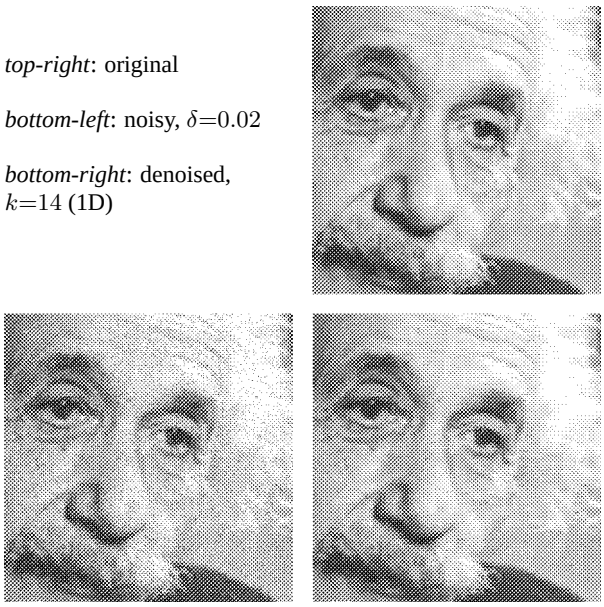


top-right: original

bottom-left: noisy, δ=0.05

bottom-right: denoised, k=12 (2D)

Fig. 1. Denoising of a scanned text page



top-right: original

bottom-left: noisy, δ=0.02

bottom-right: denoised, k=14 (1D)

Fig. 2. Denoising of a halftone image

The second image reported on is a half-toned photograph of Albert Einstein. While it is arguable whether denoising of half-tone images is a common application, these images provide good test cases for a denoiser, which has to distinguish between the random noise and the "texture" of the half-tone pattern. The results are shown in the lower part of Table 1, which shows that the DUDE is able to achieve significant denoising of the half-tone. In contrast, the more traditional algorithms fail, and, in fact, significantly amplify the distortion. Portions of the clean, noisy, and DUDE-denoised half-tone images for the experiment with $\delta = 0.02$ are shown in Figure 2. The experiments on half-tones serve to showcase the universality of the DUDE: the same algorithm that performed well on the scanned text of the first example, also performs well for the half-toned photograph, a very different type of image.

8. CONCLUSION

It is often the case that theoretically optimal schemes offer an excellent foundation for practical algorithms, but the path from theory to practice is not straightforward, and it involves judicious compromises, heuristics, and experimentation. The DUDE is no different in this respect. In this paper, we extended the theory underlying the DUDE to two-dimensionally indexed data, and reported on its implementation for binary images. It can be argued that the binary case is one whose implementation can be kept closest to the asymptotically optimal scheme. Yet, even the binary case exhibits some of the fundamental practical challenges, including the design of a feasible and efficient context model. The requirement for such a design becomes essential when extending the domain of application to larger alphabets (e.g., continuous tone images). However, similar issues have been addressed in related areas of image processing, e.g. lossless image compression (see, for instance, [7]), and significant knowledge and experience have been generated, which can be brought to bear on the discrete denoising problem. In particular, techniques for context model optimization through context quantization and aggregation, as well as incorporation of prior knowledge on the data, are likely to yield significant improvements on the DUDE's practical performance. Research is ongoing on these issues, with promising initial results.

9. REFERENCES

[1] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú and M.J. Weinberger, "Universal Discrete Denoising: Known Channel," Hewlett-Packard Laboratories Tech. Rep. HPL-2003-29, February 2003; submitted to *IEEE Trans. Inform. Theory* (manuscript available at `http://www.hpl.hp.com/infotheory`).

[2] R.C. Gonzalez and R.E. Woods, *Digital Image Processing*, Addison Wesley, New York, 1992.

[3] P. Soille, *Morphological Image Analysis: Principles and Applications*, Springer-Verlag, 1999.

[4] E. Ordentlich, G. Seroussi, S. Verdú, M.J Weinberger and T. Weissman, "Universal Discrete Denoising for Images and Mutli-Dimensional Data-Arrays," in preparation.

[5] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1998.

[6] C.E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.*, **27** (1948), pp. 379–423 and 623–656.

[7] B. Carpentieri, M.J. Weinberger, and G. Seroussi, "Lossless Compression of Continuous-tone Images," *Proceedings of the IEEE*, **88** (2000), pp. 1797–1809.