THE DUDE FRAMEWORK FOR CONTINUOUS TONE IMAGE DENOISING

Giovanni Motta   Erik Ordentlich   Ignacio Ramírez   Gadiel Seroussi   Marcelo J. Weinberger

1. INTRODUCTION

The *discrete universal denoiser* (DUDE), introduced in [1], aims at recovering a signal with finite-valued components corrupted by finite-valued, memoryless noise. The DUDE is 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 [2], the definition of the DUDE was extended to two-dimensionally indexed data, and an implementation of the scheme for binary images was presented, which outperforms other known schemes for denoising this type of data. Although the asymptotic results of [1] apply to any finite alphabet, it was observed in [2] that extending the results to continuous tone images (or, in general, to other types of data over large alphabets) presented significant challenges.

In this extended summary, we describe how these challenges can be addressed. As in lossless image compression (see, e.g., the survey [3]), a key component of the DUDE framework is the determination of a probability distribution for samples of the input (noisy) image, conditioned on their contexts. Thus, we can leverage from tools developed and tested in the context of lossless compression for determining such distributions, together with tools that are specific to the assumptions of the denoising application. These tools combine with the DUDE principles into a framework that yields powerful and practical denoisers for continuous tone images corrupted by a variety of noise processes.

Section 2 reviews the basic concepts, notations, and results from [1] and [2]. Section 3 discusses the mentioned challenges in applying the DUDE framework to continuous tone images, and the tools used to address these challenges. In Section 4 we describe experiments performed with the resulting denoisers, comparing whenever possible with other denoisers from the literature. We show that denoisers based on the DUDE framework approach and in some cases surpass state-of-the art denoising performance also on continuous tone images.

G. Motta is with Bitfone Corp. 32451 Golden Lantern, Ste. 301, Laguna Niguel, CA 92677 (gim@ieee.org). I. Ramírez is with Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay (nacho@fing.edu.uy). The other authors are with Hewlett-Packard Laboratories, Palo Alto, CA 94304 USA, ([eord,seroussi,marcelo]@hpl.hp.com). The work of G. Motta and I. Ramírez was done at Hewlett-Packard Laboratories.

2. BACKGROUND

Throughout we let $\mathcal{A}$, of size $|\mathcal{A}|$, denote the finite alphabet where the components of the clean, as well as those of the noisy image, take their values. The i-th component of a vector $\mathbf{u}$ will be denoted by u_i. We extend the notation to images, by allowing i to run over pairs $(i_h, i_v) \in \mathbb{N}^2$, where $\mathbb{N}$ denotes the set of positive integers. Let $\mathbf{x} = \{x_i\}_{i \in \mathbb{N}^2}$ denote the (conceptually infinite) clean image and $\mathbf{z} = \{z_i\}_{i \in \mathbb{N}^2}$ its noisy version. For $m, n \in \mathbb{N}$, we denote by $x^{m \times n}$ the array $(x_{(i_h, i_v)}), 1 \le i_h \le m, 1 \le i_v \le n$. We denote by $V_{m \times n}$ the set of valid indices in a $m \times n$ image.

The DUDE algorithm defined in [2] is parameterized by a stochastic *channel transition probability matrix* $\mathbf{\Pi} = \{\Pi(a, b)\}_{a, b \in \mathcal{A}}$, a *loss matrix* $\mathbf{\Lambda} = \{\Lambda(a, b)\}_{a, b \in \mathcal{A}}$ and a *neighborhood*. An entry $\Pi(a, b)$ of $\mathbf{\Pi}$ is interpreted as the probability that the observed noisy symbol at a given location is b when the underlying clean symbol is a. These probabilities should model the actual degradation process or channel giving rise to the noisy image. We assume that $\mathbf{\Pi}$ is invertible (this assumption is relaxed in [1]; many channels used in practice satisfy the condition in principle, although some present numerical problems, as discussed in Section 3). An entry $\Lambda(a, b)$ of $\mathbf{\Lambda}$ is interpreted as the loss incurred by estimating the symbol a with the symbol b. The loss matrix should reflect the fidelity criterion by which denoising performance is evaluated relative to the clean image. Finally, a neighborhood is a subset of $\mathbb{Z}^2$ not containing the origin $(0, 0)$ (the *center* of the neighborhood). Examples of neighborhoods used in practice include $(2k+1) \times (2k+1)$ squares, or circles of radius k, centered at $(0, 0)$, for $k \ge 0$. For a neighborhood $\mathcal{S}$, we let $\mathcal{S} + i = \{j + i : j \in \mathcal{S}\}$, using vector addition for 2D indices, and $x(\mathcal{S} + i) = (x_{j+i})_{j \in \mathcal{S}}$. Thus, $x(\mathcal{S} + i)$ is a $|\mathcal{S}|$-dimensional vector with $\mathcal{A}$-valued components indexed by the elements of $\mathcal{S} + i$. Such a vector $x(\mathcal{S} + i)$ for a generic index i will be referred to as the *context* of the image sample x_i at index i.

Given matrices $\mathbf{\Pi}$ and $\mathbf{\Lambda}$, and a neighborhood $\mathcal{S}$, a denoiser $\hat{X}_{\mathcal{S}}^{m \times n}$ (the DUDE) is defined in [2] by means of the computations described informally in Fig. 1 (details can be found in [2]). All the statistics needed in Steps 1-2 of Fig. 1 are generated in one pass through the image, in which conditional counts of symbols are collected for each observed context. The computation in Step 2 is performed for each observed context, while the computation associated with Step 3 is performed in a second pass through the image.

Universality in semi-stochastic and stochastic settings,

For each index $i \in V_{m \times n}$:

1. Determine the empirical distribution of *noisy* symbols z_j whose context $z(\mathcal{S}+j)$ is identical to $z(\mathcal{S}+i)$.

2. Determine a posterior distribution of *clean* symbols x_j whose corresponding *noisy* symbol is z_i with context $z(\mathcal{S}+i)$. This distribution is computed from the empirical distribution in Step 1, using the inverse of the matrix $\mathbf{\Pi}$.

3. Using the loss matrix $\mathbf{\Lambda}$, produce a denoised value $\hat{x}_i = \hat{X}_{\mathcal{S}}^{m \times n}(z^{m \times n})[i]$ such that the expectation of the loss $\Lambda(x_i, \hat{x}_i)$ with respect to the conditional distribution determined in Step 2 is minimized.

Fig. 1. DUDE outline.

in which the noisy image is generated by a discrete memoryless channel with transition probability matrix $\mathbf{\Pi}$, is established for the scheme of Fig. 1 by considering a sequence of distinct neighborhoods $\{\mathcal{S}_k\}_{k \geq 0}$ and letting k grow at an appropriate rate with respect to the dimensions m, n of the image (see [1] and [2]).

3. CONTINUOUS-TONE IMAGE DENOISING

A crucial component of the DUDE is the determination of a conditional distribution of noisy samples z_i given their context $z(\mathcal{S}+i)$ (Step 1 of Fig. 1), which is obtained by collecting empirical counts of occurrences of symbols in the observed contexts. Determining conditional distributions of samples given their context is also a key operation in lossless data compression and, in the case of universal compression, the number of conditioning contexts plays a fundamental role in determinining the convergence of the code length to the entropy. The code length includes, either explicitly or implicitly, a *model cost* [4], which is proportional to the number of free statistical parameters in the model. The model cost reflects the price paid for learning the statistics of the data: if there are many parameters to learn, many data samples will be required to accumulate significant statistics for each parameter (hence, the problem is sometimes described as one of "sparse statistics"). Model cost is particularly significant for context models over large alphabets, as the size of the alphabet generally impacts both the number of potential contexts and the number of parameters per context. The other component of the code length, a modeling component, is determined by the degree to which elements of the model class capture the statistical properties of the data. The theory and practice of universal lossless data compression address a fundamental trade-off between these two components of the code length: a richer model class with a higher accuracy of models therein results in a shorter modeling component of the code length at the expense of a greater model cost component.

In denoising, a similar trade-off exists between the richness of a model class for the underlying clean and noisy signals and the ability to learn accurate models. The results of [1] show a strong dependence of the convergence of the DUDE performance to optimal performance on the size of the context model. This convergence is determined largely by the degree to which the law of large numbers has taken hold on random subsequences of samples z_i occurring in a given context and having a given underlying clean sample value. Convergence requires that these subsequences be relatively long, implying numerous occurrences of each noisy context and underlying clean sample value. A large neighborhood and a large sample alphabet clearly lead toward shorter subsequences thereby increasing the "denoising model cost."

The results of [2] showed that the original DUDE scheme of [1], with few modifications, was effective in denoising bi-tonal images. However, it is a practical fact that even for the smallest useful neighborhoods one might contemplate, the size of a typical continuous tone image (now or in the foreseeable future) is not sufficient for a context model of the sort described in Section 2 to approach an asymptotic regime where the optimality results of [1] are meaningful. In most cases, there will be very few repetitions of observed contexts in an image, and each context will "capture" only a small number of samples.

To address this problem, we exploit prior knowledge on the structure of image data to let contexts share their information, and allow many different contexts to "contribute" to the conditional distribution used at each image location. In lossless image compression, this has often been achieved by *prediction*, i.e., the assumption that groups of conditional distributions depend on the conditioning context only through a context dependent offset (the predicted value), and *context clustering*, i.e., partitioning the space of actual contexts into a much smaller number of *conditioning classes*, which are generally disjoint. The two techniques are used, for example, in [5] (see [7] for a theoretical analysis of the role of prediction in compression). However, there is no compelling reason for the disjointness of the classes, or for a given context not to contribute its information to more than one conditioning class. An extreme case of this paradigm is presented in [6], where every context contributes, in an appropriately weighted form, to the denoising of every location of an image. Thus, we address the DUDE model cost problem for continuous tone images by augmenting the baseline DUDE algorithm with two additional tunable components: a prediction component and a context clustering component. These components serve to "blend" information from different contexts to determine the required conditional distributions of the data.

Let $\hat{z}(\cdot)$ denote a prediction function mapping contexts $z(\mathcal{S}+i)$ to values in $\mathcal{A}$. Let $C(\cdot)$ denote a clustering function mapping contexts into integer valued cluster indices. The augmented DUDE framework then replaces the conditional distribution computed in Step 1 of the procedure of Fig. 1 with $\mathcal{A}(\hat{p}(z_i - \hat{z}(z(\mathcal{S}+i))|C(z(\mathcal{S}+i))))$, where $\hat{p}(e|C(z(\mathcal{S}+i)))$ denotes the empirical distribution of the prediction error $z_j - \hat{z}(z(\mathcal{S}+j))$ along the subsequence of indices j for which $C(z(\mathcal{S}+j)) = C(z(\mathcal{S}+i))$, and

$\mathcal{A}(\cdot)$ denotes a clamping operation that forces the support of the final distribution to be contained in $\mathcal{A}$. The underlying assumption behind this modification is that the conditional distribution of z_i given $z(\mathcal{S} + i)$ depends on $z(\mathcal{S} + i)$ only through $C(z(\mathcal{S} + i))$ and the "DC shift" $\hat{z}(z(\mathcal{S} + i))$. Thus, a distribution is determined per cluster and predicted value.

Note that although our approach to reducing denoising model cost is motivated by methods from lossless image compression, we cannot directly apply the predictors and clustering techniques underlying these methods since they are restricted to be causal in nature and, more significantly, have been designed to model clean images, while here we allow non–causal contexts and seek to model noisy images. Therefore, like the channel parameter $\mathbf{\Pi}$, the prediction and clustering components also need to be matched to the noise corrupting process at hand.

As a common first step in context clustering and prediction, the image can be initially denoised by a "rough" denoiser (e.g., a median denoiser for salt and pepper noise, or a Wiener filter for Gaussian noise, or simply the identity function). Contexts are then built of the roughly denoised symbols, but the conditional statistics collected still correspond to the original noisy symbols. These statistics are used to denoise the image according to the above modification of the scheme of Fig. 1. The procedure can then be repeated, using the DUDE-denoised image as the (roughly denoised) starting point of the next iteration (but always collecting statistics on original noisy symbols). The stopping point of the iteration can be determined using various heuristics, depending on the noise channel and the mode of operation of the denoiser (e.g., by visual inspection in an interactive environment). Notice that for nontrivial context sizes, each iteration increases the total number of original samples affecting the denoising of each location.

The specific clustering function $C(\cdot)$ we have considered is a composition of the iterative denoising procedure just described with a possible "DC removal" implemented as $z(\mathcal{S} + i) - \hat{z}(\mathcal{S} + i)$, followed by a spatial transformation to exploit assumed spatial (e.g. rotational) symmetries in the conditional distribution, followed finally by a standard (vector or scalar) quantization procedure. In general, we allow both the clustering and prediction mappings to themselves depend on the entire noisy image. The quantization procedure playing a role in $C(\cdot)$ for instance may be a vector quantizer designed using the Generalized Loyd algorithm [9] operating on the entire set of contexts.

Matrix inversion. In a first sub-step of Step 2 of Fig. 1, the inverse of the channel transition matrix $\mathbf{\Pi}$ is used to determine a posterior distribution of the clean symbol x_i given the noisy context $z(\mathcal{S} + i)$. Although $\mathbf{\Pi}$ is formally non-singular for most channels of interest in image denoising, it is very badly conditioned in some important cases, and, most notably, in the Gaussian case. In practice, for these channels, a numerical procedure can be used to solve for the desired conditional distribution $\hat{P}_x$ of x_i, by minimizing a function of the form $||\hat{P}_x \mathbf{\Pi} - \hat{P}_z||$, subject to numerical

%S&P	Lena* ([8])	Lena* (DUDE)	Lena (DUDE)
30	35.6	36.8	38.3
50	32.3	32.9	34.1
70	29.3	29.6	30.7

Table 1. PSNRs (dB) for S&P denoising on Lena* and Lena by DUDE and [8].

tolerances and stability, and the constraint that $\hat{P}_x$ be a valid distribution (written as a vector). Here, $\hat{P}_z$ is a probability vector derived from the empirical counts of noisy samples.

4. RESULTS

Denoisers based on the augmented DUDE framework and incorporating the tools described in Section 3 were implemented and tested on a variety of channels and continuous tone images. We present a sample of the results, comparing, whenever possible, to other schemes in the literature. All images tested were gray-scale, with intensity values between $a_{\min} = 0$ (black) and $a_{\max} = 255$ (white).

Salt and pepper noise. The *salt and pepper* (S&P) channel corrupts each image sample with probability δ, independently of other samples. When a sample is corrupted, it is switched to maximum or minimum intensity with equal probability. The S&P channel resembles an erasure channel, in that any value strictly between $a_{\min}$ and $a_{\max}$ is known to be uncorrupted. Samples valued $a_{\min}$ or $a_{\max}$, on the other hand, may be clean or noisy. The S&P channel is not additive, and its transition matrix is well conditioned and easily inverted (except when δ approaches one).

The DUDE implementation for S&P uses a context model, nicknamed Napkin, based on a fixed neighborhood comprised of the L_1 ball of radius 2 about the origin. Contexts are quantized using quantized gradients, activity levels, and texture information (as in JPEG-LS and CALIC; see, e.g., [3]), and are used for both prediction and (after further context quantization) statistical modeling. The denoiser was iterated as described in Section 3.

Table 1 compares the augmented DUDE framework with an algorithm recently proposed in [8]. The results reported in [8] appear to be for a non–standard grayscale version of the Lena image (denoted Lena* hereafter) which we obtained from the first author's website and to which we applied the augmented DUDE framework to generate the PSNR figures in column 3 of Table 1. For reference, in column 4 we give the PSNRs of the DUDE framework on a more widely used grayscale version of Lena (http://www.dsp.ece.rice.edu/~wakin/images/). We see from Table 1 that the DUDE framework attains favorable performance relative to the algorithm of [8], which is based on a total variation–like minimization approach, and is shown in [8] to, in turn, outperform a variety of previously proposed S&P denoisers.

Gaussian noise. The tuning of the prediction and clustering components of the augmented DUDE framework to the Gaussian noise setting is still a work in progress. Prelim-
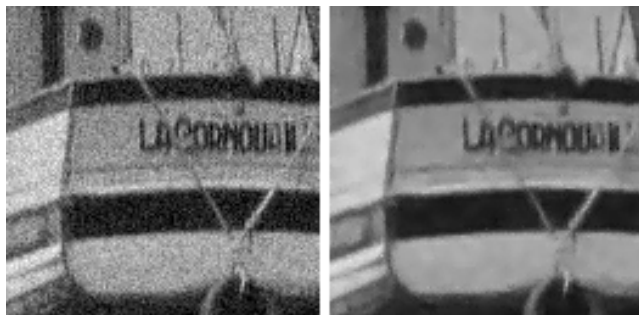
Fig. 2. Denoising on the Gaussian channel ($\sigma = 15$); left: noisy image (PSNR=24.7dB), right: DUDE-denoised (PSNR=31.7dB).

inary results are promising. We experimented with a prediction unit based on the Napkin model used above for S&P noise, and a context quantizer based on a Generalized Lloyd optimized vector quantizer. Fig. 2 shows a portion of the standard 720×576 'boats' image, corrupted by Gaussian noise with $\sigma = 15$, and its denoised version. In terms of PSNR, the augmented DUDE framework outperforms the early generations of wavelet based denoisers, but falls short of the state of the art wavelet based denoiser in [10].

"Real life" denoising. Although well characterized channels are useful in the design and analysis of denoising schemes, real-life noisy images seldom abide by the abstract models. In a practical setting, the channel structure and parameters can be regarded as "knobs" of a denoising system, which can be tuned to achieve the best performance. The latter is often characterized by visual inspection. In such a setting, flexibility and robustness of a scheme against a mismatch in channel parameters or other assumptions is a very desirable property. Preliminary experiments indicate that the augmented DUDE framework is indeed quite robust. The upper left of Fig. 3 shows a segment of a scan of an antique book page, showing "bleeding" from the reverse side of the page. The resulting image was denoised with an augmented DUDE denoiser tuned for Gaussian noise with $\sigma = 20$ (lower right) and two other denoisers designed for Gaussian noise: the adaptive Wiener filter implemented as the *wiener2* function in Matlab Ver. 6.1 (lower left) and the BLS–GSM wavelet based denoiser of [10] (upper right), as embodied by the Matlab implementation made available by the authors. The parameters of the Wiener filter and BLS–GSM denoiser were hand optimized to yield the best subjective results. For the images shown, *wiener2* was run with a 15x15 neighborhood and noise power 1000, while the BLS–GSM denoiser was run with a noise standard deviation of 100 and all other parameters set to those yielding the best results in [10] (as specified in the Matlab implementation's *denoi_demo.m* file).

5. REFERENCES

[1] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. J. Weinberger, "Universal discrete denoising: Known channel," *IEEE Trans. Inform. Theory*, Jan. 2005.
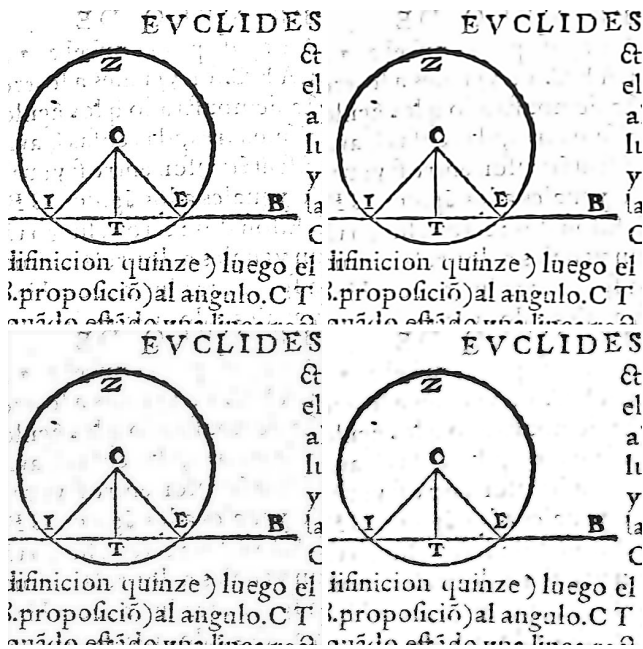
Fig. 3. Denoising an old book page. *Clockwise* starting with top–left: scanned image; BLS–GSM denoiser of [10]; DUDE framework for Gaussian channel; adaptive Wiener filter. PSNRs relative to a hand–cleaned version are 23.2 dB (scanned), 22.6 dB (BLS–GSM), 25.9 (DUDE), and 25.2 dB (Wiener).

[2] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. J. Weinberger, "A discrete universal denoiser and its application to binary images," in *Proc. of ICIP'03*, Barcelona, Sept. 2003.

[3] B. Carpentieri, M. J. Weinberger, and G. Seroussi, "Lossless compression of continuous-tone images," *Proceedings of the IEEE*, vol. 88, no. 11, pp. 1797–1809, Nov. 2000.

[4] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July 1984.

[5] M. J. Weinberger, G. Seroussi, and G. Sapiro, "The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS," *IEEE Trans. Image Proc.*, vol. 9, no. 8, pp. 1309–1324, Aug. 2000.

[6] A. Buades, B. Coll, and J. M. Morel, "A review of image denoising algorithms, with a new one," to appear in SIAM J. of Multiscale Modelling and Simulation (MMS).

[7] M. J. Weinberger and G. Seroussi, "Sequential prediction and ranking in universal context modeling and data compression," *IEEE Trans. Inform. Theory*, vol. 43, no. 5, pp. 1697–1706, Sept. 1997.

[8] R. H. Chan, C. Ho, and M. Nikolova, "Salt–and–pepper noise removal by median–type noise detectors and detail–preserving regularization," To appear in *IEEE Transactions on Image Processing*, preprint and images available at http://www.math.cuhk.edu.hk/~rchan/paper/impulse/.

[9] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 45, pp. 437–444, April 1997.

[10] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *IEEE Transactions on Image Processing*, vol. 12, no. 11, pp. 1338–1351, November 2003. Software implementation available at http://decsai.ugr.es/~javier/denoise/index.html.