Information-Theoretic Models in Image Processing

Gadiel Seroussi
Information Theory Research
Hewlett-Packard Laboratories
Palo Alto, California
Crossed Views

(representation of reality)

information theory

computer vision

neurosciences

(not to scale)
What is “reality” in information theory?

- Reality is represented by the (information-theoretic) models we use to solve data processing problems.
- Models are generally described in the language of probability.
- We can often prove very strong theorems for abstract data models:
  - tight asymptotic bounds, convergence rates, complexity statements.
- But, also often, reality is messier than our models!
  - We compromise, introducing domain knowledge that does not always lend itself to “clean” formalization.
  - In the case of images, the knowledge comes from physics, signal processing, and other approaches to image processing.
- In the end, practical success depends on the right choice of a model class (essentially an art), and optimizing algorithms for the model class.
Context models

- Data: a sequence $x_1^n = x_1 \ldots x_n$ of symbols over an alphabet $A$ of size $\alpha = |A|$ (e.g. $\alpha = 256$ for 8-bit pixels)

- A causal context model is a set of conditional distributions $P(x_t | C(x_{1:t-1}))$ where $C$ maps the past sequence $x_{1:t-1}$ to a conditioning state $C(x_{1:t-1})$ for $x_t$ (from a finite set of $|C|$ states)

- The model assigns a probability $P(x_1^n) = p_0 \prod_t P(x_t | C(x_{1:t-1}))$ to $x_1^n$

- Total number of free statistical parameters in the context model: $K = |C| \cdot (\alpha - 1)$ (assuming a full parametrization)

- For image compression, we generally make $C$ a function of a (causal) template of pixels surrounding $x_t$
Universal modeling and coding

- Context models can be seen as variants of Markov models.
- Under various settings, and with appropriate technical assumptions, it can be shown that $x_1^n$ can be losslessly encoded using a minimum possible

$$L = -\log_2 P(x^n) + \frac{K}{2} \log_2 n + o(K \log n) \text{ bits}$$

- This can be implemented efficiently, and holds in a doubly universal setting where $K$ is not fixed in advance. (Rissanen, Ryabko, many others; paper by Martín, Seroussi, & Weinberger, IEEE IT July 2004 gives history, biblio, and most recent results)

Minimum Description Length (MDL) principle of statistical inference:

- choose the model class that provides the shortest code length for the model and the data in terms of the model

A constructive, algorithmic way out of Kolmogorov complexity’s limitations
Lossless image compression

- Universal modeling and coding generally do not work when applied directly to continuous tone images.
  - Conditioning directly on template values results in an astronomical number of possible states, even for moderate template sizes.
  - With full parametrization, each state adds \((\alpha-1)\) free parameters.
  - \(\Rightarrow\) Model cost \((K)\) out of control (\(=\) “sparse statistics” = “curse of dimensionality”)

- Solutions
  - Context quantization: merge statistics of “similar” contexts
  - Use prediction, collect statistics on prediction errors
  - Use low-dimensional parametric families of conditional distributions (e.g. two-sided geometric or discrete Laplacian)

Use prior knowledge on characteristics of image data to keep model cost in check while still capturing useful data redundancies
How does prediction help?

- Prediction has been used since the early days of image coding as a means to “decorrelate” data, in conjunction with memoryless symbol-by-symbol coding (e.g. Huffman).
- In the context of universal coding, it was less clear why the step was needed, as the universal modeler would eventually learn the predictive pattern.
- The explanation: context-dependent prediction allows for merging contexts that have similar distributions up to alphabet shift.

Prior knowledge on images tells us that such contexts exist and will merge (we don’t need to wait for the universal modeler to learn that).

Prediction allows for significant model cost reductions, and, thus faster convergence to the limiting optimal code length [formalized in Weinberger & Seroussi IEEE IT 1997]
The JPEG-LS standard

- The **LOCO-I** algorithm, a low complexity embodiment of these principles, was adopted as the core of the JPEG-LS lossless image compression standard [Weinberger, Seroussi, Sapiro *IEEE IP 2000*].

![Graph showing relative running time vs. average bits/symbol](image)

Mars Rover image encoded with LOCO-I
A hierarchy of learning

- decide to use prediction
- fixed prediction
- adaptive prediction
- model prediction errors
- design time
- each stage builds on "knowledge" acquired in the previous stages, and adjusts parameters left free by the previous stages
- run time
- wire brain for language
- learn about grammar
- adjust grammar
- speak Spanish
- genetic time
- brain time
Universal image denoising

- **Discrete Universal DEnoiser (DUDE)**
  - Giovanni Motta and Ignacio Ramírez – adaptations and implementations for continuous tone images
Discrete Denoising

- $X_i, Z_i, \hat{X}_i$ take values from finite alphabets.

- **Goal:** Choose $\hat{X}_1, \ldots, \hat{X}_n$ on the basis of $Z_1, \ldots, Z_n$ to minimize a fidelity criterion *(which we might not be able to measure!)*.

(For simplicity, we will use 1D indexing notation $X_i$ for images — assume an appropriate scanning order, or interpret an index $i$ as a coordinate pair; $n \rightarrow \infty$ then means *all* dimensions go to $\infty$ simultaneously.)
Universal Denoising Setting: Assumptions

- **$n$-block denoiser**: $\hat{X}^n : A^n \rightarrow A^n$.

- **Normalized cumulative loss** of the denoiser $\hat{X}^n$ when the observed sequence is $z^n \in A^n$ and the channel input block is $x^n \in A^n$:

  $$L_{\hat{X}^n}(x^n, z^n) = \frac{1}{n} \sum_{i=1}^{n} \Lambda(x_i, \hat{X}^n(z^n)[i])$$

- Define a sequence of distinct finite **context templates** $S_k \subset \mathbb{Z}^2 \setminus \{(0,0)\}$, with $K = |S_k|$ non-decreasing with $k$ (e.g., $L_p$-balls of radius $k$, with the center $(0,0)$ removed).

Example:

- $K = 12$
- $\{(0, \pm 1), (\pm 1, 0), (\pm 1, \pm 1), (0, \pm 2), (\pm 2, 0)\}$
The DUDE Algorithm for 2D data: General Idea

- Use the context templates \( S_k \subset \mathbb{Z}^2 \setminus \{(0, 0)\} \)

- Fix \( k \). For each symbol \( Z_i \) to be denoised, do:
  - Determine the \( S_k \)-context centered at \( Z_i \), \( S_k(Z^n_1, i) \).
  - Count all occurrences of letters with the same \( S_k \) context pattern. This gives a conditional empirical distribution of the noisy symbol given the context pattern.
  - Use channel transition probability to estimate the conditional empirical distribution of the noiseless symbol \( X_i \) given the noisy context \( S_k(Z^n_1, i) \).
  - Make decision using
    - the loss function,
    - the channel transition probability,
    - the conditional empirical distribution
    - the observed symbol to be denoised.
Optimality Result: Semi-Stochastic Setting

Minimum $S_k$-sliding-window loss of $(x^n, z^n)$:

$$D_k(x^n, z^n) = \min_{f : A^{S_k} \times A \to A} \left[ \frac{1}{n} \sum_{i=1}^{n} \Lambda(x_i, f(S_k(z^n_i, i), z_i) \right]$$

We define a **Discrete Universal DEnoiser** (DUDE) $\hat{X}^n_{univ}$ that satisfies:

**Theorem.** For any input image $X$, a.s.,

$$\lim_{n \to \infty} \left[ L_{\hat{X}^n_{univ}} (x^n, Z^n) - D_{kn} (x^n, Z^n) \right] = 0$$

No asymptotic penalty for universality in the individual sequence setting.
**Theorem.** For all stationary (shift invariant) $X$, 

$$\lim_{n \to \infty} EL_{\hat{X}_{\text{univ}}^n}(X^n, Z^n) = \inf_{n \geq 1} \min_{\hat{X}^n \in D_n} EL_{\hat{X}^n}(X^n, Z^n),$$

where $D_n$ is the class of all (including non-universal) $n$-block denoisers.

*No asymptotic penalty for universality in the probabilistic setting.*
The Discrete Universal Denoiser

Fix $k$, $S_k$

Pass 1 For every index $i$ in the image:
compute the count vectors $m(z^n_1, S_k(z^n_1, i)) =$
histogram of symbols for each context pattern assumed by $S_k$ in $z^n_1$.

Pass 2 Correct according to:

$$
\hat{X}^n_{S_k}(z^n)[i] = \arg \min_{\hat{x} \in \mathcal{A}} m^T(z^n_1, S_k(z^n_1, i)) \Pi^{-1} (\lambda \hat{x} \odot \pi z_i).
$$

$$
\Pi = \{\Pi(i, j)\}_{i, j \in \mathcal{A}} = [\pi_1 | \cdots | \pi_M]
$$

$$
\Lambda = \{\Lambda(i, j)\}_{i, j \in \mathcal{A}} = [\lambda_1 | \cdots | \lambda_M].
$$

$$(v \odot w)_i = v_i w_i$$

Two passes, each in linear time.
A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist and Hartley on this subject. In the present paper we will extend the theory to include a
A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist and Hartley on this subject. In the present paper we will extend the theory to include a random bit error rate: 5.0%
A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist¹ and Hartley² on this subject. In the present paper we will extend the theory to include a

denoised bit error rate: 0.4%
Bi-level image denoising: original

image: 896x1160 half-tone (600x350 segment shown)
Noisy image

random bit error rate: 2.0%
denoised bit error rate: 0.7%
Continuous Tone Image Denoising: Issues

• Large alphabet (e.g., \( \alpha = 256 \) for 8 bpp)

  - huge number of possible context patterns (e.g. \( 256^8 \approx 1.8 \cdot 10^{19} \) for 3×3)
  - large number of statistical parameters per context
  - the problem is not that we don’t have enough resources (e.g. memory), but that we don’t have enough image
Continuous Tone Image Denoising: Issues

- Large alphabet (e.g., $\alpha = 256$ for 8 bpp)
  - huge number of possible context patterns (e.g. $256^8 \approx 1.8 \cdot 10^{19}$ for $3 \times 3$)
  - large number of statistical parameters per context
  - the problem is not that we don’t have enough resources (e.g. memory), but that we don’t have enough image

model cost (or sparse statistics) problem:
learning is expensive!
Continuous Tone Image Denoising: Issues

- Large alphabet (e.g., $\alpha = 256$ for 8 bpp)
  - huge number of possible context patterns (e.g. $256^8 \approx 1.8 \cdot 10^{19}$ for $3 \times 3$)
  - large number of statistical parameters per context
  - the problem is not that we don’t have enough resources (e.g. memory), but that we don’t have enough image

  **model cost** (or **sparse statistics**) problem:
  * learning is expensive!

- Raw DUDE does not exploit **prior knowledge** of image properties (e.g. smoothness)

- Inversion of channel matrix $\Pi$ can be numerically problematic
Continuous Tone Image Denoising: Approaches

- Key component of the DUDE: model the conditional distribution

\[ P(Z_i | \text{context of } Z_i) \] noisy conditioned on noisy

from which we then estimate

\[ P(X_i | \text{context of } Z_i) \] clean conditioned on noisy
Continuous Tone Image Denoising: Approaches

- Key component of the DUDE: model the conditional distribution
  \[ P(Z_i | \text{context of } Z_i) \]  noisy conditioned on noisy
  from which we then estimate
  \[ P(X_i | \text{context of } Z_i) \]  clean conditioned on noisy

- Modeling task similar to that of a *lossless image compressor*
  (but with non-causal contexts)

- Problems of *model cost* and use of *prior knowledge* have been
  successfully addressed in image compression (e.g., LOCO-I/JPEG-LS;
  model cost considerations going back to Rissanen, Langdon—early’80s)
Continuous Tone Image Denoising: Approaches

- Key component of the DUDE: model the conditional distribution

\[ P(Z_i \mid \text{context of } Z_i) \text{ noisy conditioned on noisy} \]
\[ P(X_i \mid \text{context of } Z_i) \text{ clean conditioned on noisy} \]

- Modeling task similar to that of a lossless image compressor (but with non-causal contexts)

- Problems of model cost and use of prior knowledge have been successfully addressed in image compression (e.g., LOCO-I/JPEG-LS; model cost considerations going back to Rissanen, Langdon—early’80s)

- Some building blocks:
  - context quantization (reduce the number of conditioning states)
  - prediction (incorporate prior knowledge)
  - parametric distributions (e.g. Laplacian, Gaussian, “binning”)

*with suitable adaptations, similar blocks can also be used in the DUDE*
What happened to universality?

- Lessons learned from data compression:

  * asymptotic results are great, but data doesn’t always cooperate
  * be universal, but only when you must
  * if prior knowledge is available and can be used effectively, use it! (don’t pay the price of learning what you already know)
  * use of prior knowledge translates to reduced model costs and faster convergence (formalized in [Weinberger-Seroussi’97])
What happened to universality?

- Lessons learned from data compression:
  
  *asymptotic results are great, but data doesn’t always cooperate*
  
  - be universal, but *only when you must*
  - if prior knowledge is available and can be used effectively, *use it!* (don’t pay the price of learning what you already know)
  - use of prior knowledge translates to reduced model costs and faster convergence (formalized in [Weinberger-Seroussi’97])

- *The best algorithms use a judicious mix of prior knowledge and universality appropriate for the application at hand*
“Salt and pepper” noise $\delta = 30\%$ (PSNR=10.6dB)
Denoised by the DUDE (PSNR=38.5dB)
Denoised by selective median filter (PSNR=28.8dB)
More Salt and Pepper

Salt & Pepper

$\delta = 50\%, \quad \text{PSNR} = 7.3 \, \text{dB}$
More Salt and Pepper

Salt & Pepper
\[ \delta = 50\%, \quad \text{PSNR} = 7.3 \text{ dB} \]

denoised
\[ \text{PSNR} = 33.6 \text{ dB} \]
More Salt and Pepper

Salt & Pepper
\[ \delta = 50\%, \quad \text{PSNR} = 7.3 \, \text{dB} \]

denoised
\[ \text{PSNR} = 33.6 \, \text{dB} \]

original
Salt and Pepper: vs. prior state of the art

δ = 30%  PSNR = 10.7 dB
Salt and Pepper: vs. prior state of the art

\[ \delta = 30\% \quad \text{PSNR} = 10.7 \, \text{dB} \]

\[ \text{denoised} \quad \text{PSNR} = 37.8 \, \text{dB} \]
Salt and Pepper: vs. prior state of the art

\[ \delta = 30\% \quad \text{PSNR} = 10.7 \text{ dB} \]

\[ \text{denoised} \quad \text{PSNR} = 37.8 \text{ dB} \]

- Best previous result in the literature Pok, Liu & Nair (IEEE IP Jan’03):
  \[ \text{PSNR} = 34.3 \text{ dB} \quad @ \quad \delta = 30\% \]
Gaussian noise

\[ \sigma = 20, \quad \text{PSNR} = 20.2 \text{ dB} \]
Gaussian noise

\[ \sigma = 20, \quad \text{PSNR} = 20.2 \text{ dB} \]

denoised
\[ \text{PSNR} = 24.9 \text{ dB} \]
Gaussian noise

\[ \sigma = 20, \quad \text{PSNR} = 20.2 \, \text{dB} \]

denoised
\[ \text{PSNR} = 24.9 \, \text{dB} \]

original
Denoising “real” data

portion of scanned book page
Denoising “real” data

denoised with DUDE tuned to Gaussian $\sigma = 20$
Denoising “real” data

denoised with median
How do our models see?

- *Universal simulation* in a model class: pick a random sequence $y_1^n$ that is indistinguishable from $x_1^n$ by any model in the class
  - Merhav & Weinberger 2003 (fixed structure models), Seroussi 2004 (doubly universal)

- The simulated sequence shows *only that part of* $x_1^n$ that is captured by the model
Gray scale image simulation

original

(joint work with G. Brown and G. Sapiro)

simulated:
steerable wavelets +
universal types