Resolution/Quantization
Trade-offs in Image/Signal Representation

Alfred M. Bruckstein

Jury Lecture
Fig. 1. Original image of the authors (a) and digitized images (b) $M = 256, N = 256, b = 1$; (c) $M = 170, N = 85, b = 4$; (d) $M = 128, N = 64, b = 7$. The images require (b) 65 536, (c) 57 800, and (d) 57 344 bits. The resulting normalized values of the criterion are (b) 6.56, (c) 1, and (d) 1.59. The image (c) is closest to the optimal.
Optimal Digitization of 2-D Images
L. Nielsen, J. Aström, and E. I. Jury

Abstract—The problem of representing an image by \( M \times N \) samples with \( b \) bits/sample, subject to the constraint of a fixed total number of bits, is discussed. Reasonable assumptions are made in such a way that the optimization problem has a closed form solution. The solution is tested experimentally and agrees well with human perception of visual quality. The analytical solution brings out the dependence of the optimal digitization on image characteristics very clearly. The results explain and agree with results of other subjective tests.

I. INTRODUCTION

Vigorous research has been devoted to image processing and related fields during the last two decades. Several books have been written on various aspects of the theory and applications \([11]-[13]\). The problem of optimal digitization of 2-D images has been sporadically mentioned in several texts, but it has not been addressed in full detail \([22]-[24]\). Experimental investigations of the effect of coarse/fine print for bilevel images was initiated by Abdou and Wong \([5]\). No theory was given in this work. Steiglitz \([6]\) has presented a detailed theory of transmission of an analog signal over a fixed-bit-rate channel in the 1-D case. This work has motivated the extension to the 2-D case in this paper, which gives a definition of optimal digitization (quantization and sampling) of images for the first time. The definition seems to be meaningful in practical applications. It differs from Steiglitz work in the respect that our problem formulation admits an analytical solution.

II. PROBLEM FORMULATION

Let the original image \( f(x, y) \) be a function defined on \( \Omega = [0, L_x] \times [0, L_y] \subset R^2 \) with values in \( V = [\min f, \max f] \subset R^+ \). The image is sampled to give a sampled image \( \hat{f}(x_i, y_j) \) defined on an \( M \times N \) rectangular grid \( G \) with values in \( V \subset R^+ \). Ideal sampling is assumed, i.e., \( \hat{f}(x_i, y_j) = f(x_i, y_j) \) and \( x_i, y_j \in G \). In addition to sampling, the values of \( \hat{f}(x_i, y_j) \) are also quantized so that \( V \subset R^+ \) is represented by \( b \) bits in \( 2^b \) quantization levels. The quantization of \( \hat{f} \) is denoted by \( Q \hat{f} \), which is the digitized image defined on an \( M \times N \) grid with discrete values. We want to reconstruct the new function \( f \) defined on \( \Omega \) with values in \( V \). From the function \( Q \hat{f} \) the function \( f \) can be obtained using many different interpolation schemes \([11]\), \([5]\), \([6]\).

The optimal digitization problem can be formulated as follows. Assume that the image is represented by a fixed number of bits.

\[ M \cdot N \cdot b = C. \]  
(1)

Determine \( M, N, b \) such that the following error is minimum.

\[ E = \int \Omega [f(x, y) - \hat{f}(x, y)]^2 dx dy / \int \Omega dx dy. \]  
(2)

A number of restrictions are introduced to make the problem tractable analytically. The function \( f \) is characterized by its value range \( R = \max f - \min f \), and the mean fluctuation rates \( \sigma_x \) and \( \sigma_y \) defined by

\[ \sigma_x^2 = \frac{\int \Omega (f(x, y) - \bar{f})^2 dx dy}{\int \Omega dx dy} \]  
(3)

The quantization error \( n \) is defined as \( n = \hat{f} - \bar{f} \). Zero-order hold interpolation is used. This means that \( f(x, y) = Q \hat{f}(x_i, y_j) \) for \( x, y \) around \( x_i, y_j \).

III. SOLUTION

The criterion (2) is expanded by dividing the image in \( M \times N \) cells with sides \( \delta_x = L_x/M \) and \( \delta_y = L_y/N \). The cell midpoint \( x_i, y_j \) belongs to the grid \( G \). The contributions from all cells are then summed up. Insertion of the digitization and the interpolation schemes in the criterion (2) gives the mean square error

\[ E = \frac{1}{L_x L_y} \sum_{i, j} \left\{ \int \square [f(x - x_i, y - y_j) - f(x_i, y_j)]^2 dx dy + \int \square n(x_i, y_j)^2 dx dy \right\} \]  
(4)

where \( \square \) denotes integration over one cell. Steiglitz calls the first term in (4) the reconstruction error. This error depends only on the grid resolutions \( \delta_x \) and \( \delta_y \). A Taylor series of \( f \) gives the following expression for the reconstruction error in one cell.

\[ \frac{1}{12} \cdot (\delta_x^2 \delta_y^2 \sigma_x^2 + \delta_x \delta_y^2 \sigma_y^2). \]  
(5)

The second term in (4) is the quantization error, which depends only on the number of quantization levels \( 2^b \). Assuming equivalent quantization with the grain \( \delta = R \cdot 2^{-b} \), the quantization error is approximated by \( \delta^2 / 12 \) and

\[ \int \square n(x_i, y_j)^2 dx dy = \delta_x \delta_y \frac{R^2}{2^b}. \]  
(6)

The following formula is then obtained for the total mean square error (4).

\[ E = \frac{1}{12} \cdot \left( \frac{L_x^2 \sigma_x^2}{M^2} + \frac{L_y^2 \sigma_y^2}{N^2} + \frac{R^2}{2^b} \right). \]  
(7)

The optimal digitization problem is to minimize (7) subject to the constraint (1). The variables \( L_x, \sigma_x, L_y, \sigma_y \), and \( R \) are known constants which depend on the image. The problem is solved simply by completing the squares in (7) and inserting the constraint (1). The solution is

\[ b = \frac{1}{2 \ln 2} \ln \left[ \frac{R^2}{L_x \sigma_x L_y \sigma_y} \right]. \]  
(8)
\[ M = \sqrt{ \frac{L_x g_x}{L_y g_y} } \cdot \sqrt{\frac{C}{b}} \]  
(9) 

\[ N = \sqrt{ \frac{L_y g_y}{L_x g_x} } \cdot \sqrt{\frac{C}{b}} \]  
(10) 

It is interesting to see how the relation between the value range \( R \) and the fluctuation rates \( g_x \) and \( g_y \) influence the solution. More fluctuations leads to fewer bits (lower \( b \)) and more image resolution. Fewer fluctuations requires more bits and fewer samples. This agrees with earlier subjective tests [3], [4].

IV. The Experiment

The criterion (2) was chosen largely for mathematical convenience. A number of experiments have been carried out to test if the optimum digitization obtained corresponds to the subjective notation of a good digitization [7]. Fig. 1 illustrates a scan of the original image and a number of digitized versions. A scan of the original image gives the characteristics

\[ R = 4.06 \times 10^2 \ L_x g_x = 9.68 \times 10^2 \ L_y g_y. \]  
(11)

ACKNOWLEDGMENT

We are grateful to Dr. T. Pavlidis who suggested the problem to the third author. Support from the Swedish Board of Technical Development and the National Science Foundation is gratefully acknowledged.

REFERENCES


V. Conclusions

A theoretical formulation of the optimal digitization problem is given. The solution is obtained from an optimization of a criterion due to the constraint of fixed number of bits. The solution is tested experimentally and agrees well with human visual quality. An advantage is that the solution is given in closed form [see (8)-(10)]. This makes it easy to use as a rule of thumb. It also clearly points out the dependence on image characteristics. This dependence explains and agrees with results of other subjective tests. The criterion (2) is one of the simplest which admits an analytic solution. It would be interesting to look at other alternatives and make more extensive experimentation.
In 1984:

L. Nielsen, K.J. Aström & E.I. Jury published a paper titled:

OPTIMAL DIGITIZATION OF 2-D IMAGES

(IEEE T-ASSP-32/6, 1984)

The list of authors is very interesting:

- K.J. Aström - a world famous specialist in ADAPTIVE CONTROL THEORY
  Lund, Sweden

- L. Nielsen - professor of vehicular systems
  (student - first paper in 1984)
  Linköping U., Sweden
  (visual servoing for vehicles!)

- E.I. Jury - world famous in Control Theory
  Stability Criteria, etc.
  U. of Miami, Coral Gables
  Florida, U.S.A.
The paper has an ACKNOWLEDGMENT:

WE ARE GRATEFUL TO

Dr Theo PAULIDIS

WHO SUGGESTED THE PROBLEM TO
THE THIRD AUTHOR (E.I. Jury!)

Theo Paulidis, a leader in Image Processing, who once was at that time involved with
SYMBOL TECHNOLOGIES
and was interested in a wealth of "practical problems"
The Nielsen-Astrom Jury paper was subsequently quoted about 5/6 times, of these quotations by AMB and coauthors about 3/4 times.

- AMB: "On Optimal Image Digitization"
  IEEE T ASSP, 1987 (v6/35/4)

- N. Kiryati, AMB & A. Jones
  Bit Allocation in Piecewise Planar Representation of Images
  J. Vis Comm & Image Representation 6/1
  1995

- N. Kiryati, AMB
  Gray levels Can Improve the Performance of Binary Image Digitizers
  CVGIP: Graph. 53, 1991

- AMB, M Elad & R Kimmel
  Down Scaling for Better Transform Compression
  IEEE Trans Image. 12/9, 2002

* While writing these papers AMB corresponded with T. Panidis.
The Problem:

Given a set of signals \( f_w(x) \) over \([0,1]\) with values over \([-1,1]\), represent/describe them as best you can with less than \( B \) bits.

Solution

- Sample then quantize

(How to sample? How to quantize?)

If \( N \) samples will be quantized to \( b \) bits each we’ll have to have

\[ N \cdot b \leq B \]

How good is a representation

is measured by a distance between

\[ f_w(x) \] and \( f_{S(N)+G(b)}(x) \) estimated from samples quantized.
The prototype problem we want to solve is:

$$\min \text{ Distance} \left[ \int_{\mathbf{S}(N)+Q(b)}^{S(N)+Q(b)} f_u(x), \int_{\mathbf{S}(N)+Q(b)}^{S(N)+Q(b)} f_u(x) \right]$$

subject to $N, b \leq B$

Distance $\left[ \int_{\mathbf{S}(N)+Q(b)}^{S(N)+Q(b)} f_u(x), \int_{\mathbf{S}(N)+Q(b)}^{S(N)+Q(b)} f_u(x) \right]$ will be an expression involving $N, b$ and the properties of the family of functions $f(u)$.

Let us take the 1-D example of Nielsen Astrom A Jury, and look at it.
Given a function \( f(x) \) over \([0, 1]\)

- Sample by considering \( N \) equal sized intervals of length \( \delta = \frac{1}{N} \), \( R_i = \left[ \frac{i}{N}, \frac{i+1}{N} \right] \) \( i = 0, 1, \ldots, N-1 \) and describe \( f(x) \) over \( R_i \) by a single number \( f(c_i) \).

- Quantize the numbers \( f(c_i) \) by selecting \( 2^b \) values in the range \([-1, 1]\) and mapping \( f(c_i) \) to fit the closest of the \( 2^b \) values.

Then

\[
\sum_{i=0}^{N-1} \left( f(x) \right)_i = f(c_i) \quad \text{for} \quad x \in R_i
\]

is a piecewise constant representation of \( f(x) \) by \( N \cdot b \) bits.
Let us look at the "mean square" distance between $f(x)$ and its representation

$$D(f(x), f(x)) = \int_0^1 [f(x) - f(x)]^2 dx$$

$$= \sum_{i=0}^{N-1} \int_{R_i} [f(x) - f(i) + f(i) - f(i)]^2 dx =$$

$$= \sum_{i=0}^{N-1} \int_{R_i} [f(x) - f(i)]^2 dx +$$

$$\sum_{i=0}^{N-1} 2 (f(i) - f(i)) [f(x) - f(i)] dx +$$

where we see that selecting $f(i) dx = \int_{R_i} f(x) dx$ we get 0 cross terms.
Therefore we get

\[ D(f(x), f'(x)) = \sum_{i=0}^{N-1} \left[ f(x_i) - \frac{1}{6} \sum_{R_i} f(x) dx \right]^2 + \sum_{i=0}^{N-1} \delta(f(x_i) - f'(x_i))^2 \]

"Sampling Reconstruction Error"

Average Quantization Error

\[ \frac{1}{N} \sum_{i=0}^{N-1} (f(x_i) - f'(x_i))^2 \approx \frac{1}{2} \delta + \frac{1}{2} \]

by Taylor Expansion here

\[ R_i = \chi_N \approx \frac{1}{N^2} \left( \frac{1}{N} \sum_{i=0}^{N-1} f(x_i) \right) \]
Hence we obtained that

\[ D(f(x), f'(x)) = \frac{\sigma_f}{N^2} + \frac{K^2}{2b} \]

if we quantize optimally and select best representations over \( R_i \) to be quantized.

Now we can do:

\[ \min \left\{ \frac{\sigma_f}{N^2} + \frac{K^2}{2b} \right\} \]

s.t. \( N \cdot b = B \)

CUTE ISN'T IT!

So do:

\[
\nabla_{N,b} \left( \frac{\sigma_f}{N^2} + \frac{K^2}{2b} \right) + \lambda (B - N \cdot b) = 0
\]

\[
\frac{\partial}{\partial N} = 0, \quad \frac{\partial}{\partial b} = 0 \quad \text{and get solutions for } N, b.\]
$$D(B, b) = \frac{\sigma^2}{(B^2)^2} + \frac{k^2}{2^{2k}} = \frac{\sigma^2}{B^2} \cdot b^2 + \frac{k^2}{2^{2k}}$$

equation grows as $b$ increases.

equation decreases as $b$ decreases.

---

etcetera

eetcetera

eetcetera
This was just an example, somewhat similar to what NAJ did. There are many variations possible.

Obviously: "Sampling" can be done in many ways. Define a family of functions \( \{ \phi_i \} \) and project \( f \) on these via

\[
\langle f(x), \phi_i \rangle = f_i
\]

Then \( f_i \) are "generalized" samples of \( f \). (\( \phi_i \) can be Wavelets or even dictionaries!)

If \( \{ \phi_i \} \) are O.N., then we have
\[ f(x) = \sum_{i=1}^{N} f_{i} e^{i(x)} \] and the error is\[
D \left[ f(x), f_{\text{approx}} \right] = \int_{0}^{1} \left[ f(x) - f_{\text{approx}} \right]^{2} dx = \\
= \int_{0}^{1} \left[ f(x) - \sum_{i=1}^{N} f_{i} e^{i(x)} \right]^{2} dx = \\
= \int_{0}^{1} \left( f(x) - \sum_{i=1}^{N} f_{i} e^{i(x)} \right)^{2} dx + \sum_{i=1}^{N} \left( f_{i} - f_{\text{approx}_{i}} \right)^{2} dx = \\
= \int_{0}^{1} \left( f(x) - \sum_{i=1}^{N} f_{i} e^{i(x)} \right)^{2} dx + 2 \sum_{i=1}^{N} \left( f_{i} - f_{\text{approx}_{i}} \right) \left( \sum_{j=1}^{N} f_{j} \right) dx + \sum_{i=1}^{N} \left( f_{i} - f_{\text{approx}_{i}} \right)^{2} dx = \\
= \int_{0}^{1} \left( f(x) - \sum_{i=1}^{N} f_{i} e^{i(x)} \right)^{2} dx + \sum_{i=1}^{N} \left( f_{i} - f_{\text{approx}_{i}} \right)^{2} dx.
\]
• Our work on JPEG!
  • Our work on piecewise linear representation.
  • Our work on encoding B/W images.

All these can be regarded as processes of sampling and quantization.

Question:
Where did Nyquist disappear?
in all this!