Correspondence

On Optimal Image Digitization

A. M. BRUCKSTEIN

Abstract—**Nielsen *et al.* recently addressed the problem of determining the optimal discretization grid and quantization depth when a given bivariate function $f(x, y)$ has to be described with a predetermined number of bits. This was done under the assumption that the function value range and mean "fluctuation rates" in the x and y direction are given, and that ideal point sampling with zero-order-hold interpolation is used in reconstructing the image. This correspondence outlines an alternative approach, based on the assumption that $f(x, y)$ is the sample function of a 2-D stationary stochastic process with a known covariance function. We use standard integral sampling and obtain closed form solutions under the assumption that $f(x, y)$ is (the sample of) a homogeneous and separable Markov process.**

I. Introduction: The Image Digitization Problem

In order to enable the processing of images by a computer, we have to represent them as a matrix of quantized values. The process of transforming a bivariate function $f(x, y) \geq 0$ defined on, say $\Omega\,[0, 1] \times [0, 1]$, into a matrix of pixels with gray levels represented by a certain number of bits is the digitization process. This is usually done by some kind of *spatial sampling* of $f(x, y)$ which provides an $M \times N$ matrix of positive real values and their subsequent *quantization* to 2^b levels, a process that enables an approximate description of the continuous 2-D function with $M \times N \times b = B$ bits. Suppose we are given a certain sampling and quantization scheme with spatial resolution parameters M, N and quantization depth b and an interpolation process which defines the approximation $\hat{f}(x, y)$ of the original, continuous function that is recovered from a digitized image representation. Then the *optimal digitization problem* is to determine M, N, b so that $\hat{f}(x, y)$ is closest (in a well-defined sense) to $f(x, y)$, subject to the constraint $M \times N \times b = B$. In order to solve the above problem, we therefore need to define a distance measure between two continuous functions $f(x, y)$ and $\hat{f}(x, y)$. An obvious choice is, of course, the mean-square error:

$$\epsilon^2 \triangleq \iint_{\Omega} \left(f(x, y) - \hat{f}(x, y)\right)^2 dx\, dy. \tag{1}$$

Although this problem seems to be a very fundamental one in image processing, it was addressed in full detail only in a very recent paper by Nielsen *et al.* [1]. The reason for this might be that the image processing hardware usually dictates the range of values of M, N, and b, and therefore, the bit allocation implied by the digitization process is often unimplementable on existing devices. Furthermore, the digitization process as defined above is not expected to have good image coding or compression performance (in terms of the image quality achievable for a certain number of bits per pixel) when compared to other image coding methods. We note, however, that the digitization process is not supposed to replace an image coding method, the aim of optimal digitization being to generate the best image matrix that will subsequently become the input to various processing and coding algorithms. Also, the limits imposed on the image representation by the display hardware should not limit the range of possibilities in terms of how the image is represented by a set of numbers in the computer.

In [1], the optimal digitization problem was solved under the assumption of ideal point sampling, i.e., the $M \times N$ matrix of samples was $\{f(x_i, y_j)\}$ where $x_i = 2i + 1/2M$ and $y_i = 2j + 1/2N$ (the function values at midpoints of sampling cells) and given mean fluctuation rates of $f(x, y)$ in the x and y directions, and the value range of $f(x, y)$. Under some further reasonable assumptions, explicit formulas for M, N, and b were obtained showing how the resource allocation between spatial resolution (M, N) and quantization depth (b) is influenced by the mean directional fluctuation rates of $f(x, y)$.

In this correspondence, we address the optimal image digitization problem in conjunction with a different sampling scheme and under the assumption that $f(x, y)$ is the realization of a wide-sense stationary 2-D process with a given covariance function $R(\tau_x, \tau_y)$.

II. Optimal Digitization of a 2-D Stationary Process

We shall consider that $\{f_\omega(x, y)\}$ is a zero-mean, 2-D wide-sense stationary process with given covariance $R(\tau_x, \tau_y)$. (The zero-mean assumption is being made without loss of generality; simply assume that the ensemble mean of the positive valued images was already subtracted from the realizations of the process.) The sampling scheme proposed is the standard integral sampling. It divides the region $\Omega = [0, 1] \times [0, 1]$ into $M \times N$ equal area cells and the (real) value assigned to each cell (or pixel) is the mean value of $f_\omega(x, y)$ over the cell. Indeed, if over a region Δ we wish to replace the values of $f_\omega(x, y)$ by a number V_Δ, then the spatial average of the squared error induced (by this zero-order-hold interpolation process) is minimized for

$$V_\Delta = \frac{1}{\|\Delta\|} \iint_{\Delta} f_\omega(x, y)\, dx\, dy \tag{2}$$

where $\|\Delta\|$ is the area of the region Δ. Therefore, the image sampling procedure replaces $f_\omega(x, y)$ by the $M \times N$ matrix $\{f_\omega(i, j)\}$ of spatial averages of $f_\omega(x, y)$ over the equal area, rectangular sampling cells. With no further quantization on the entries of this matrix, the resulting zero-order-hold interpolation estimate of $f_\omega(x, y)$ will then be the best one possible, in the sense of minimizing the mean-square error criterion (1)—given the division of Ω into $M \times N$ equal pixels.

Let us now consider the following problem. Suppose that we replace, over a region Δ, the function $f_\omega(x, y)$ by V_Δ defined by (2): what will then be the ensemble average of the induced mean-square error? A standard calculation yields the answer to this question. Define

$$\epsilon_\Delta^2 \triangleq \frac{1}{\|\Delta\|} \iint_{\Delta} \left(f_\omega(x, y) - V_\Delta\right)^2 dx\, dy \tag{3}$$

and then, using the definition of V_Δ to simplify (3), we obtain that

$$E\epsilon_\Delta^2 = E\left[\frac{1}{\|\Delta\|} \iint_{\Delta} f_\omega^2(x, y)\, dx\, dy - V_\Delta^2\right]. \tag{4}$$

Using the definition of $R(\tau_x, \tau_y)$

$$R(\tau_x, \tau_y) = E f_\omega(x, y) f_\omega(x + \tau_x, y + \tau_y) \tag{5}$$

Manuscript received April 22, 1986; revised October 8, 1986.

The author is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

IEEE Log Number 8612866.

in the above expression for the expected mean-square error, we obtain that

$$E\epsilon_\Delta^2 = R(0,0) - EV_\Delta^2. \tag{6}$$

We also have, as immediate consequences of its definition (2), the following statistics of the random variable V_Δ:

$$EV_\Delta = Ef_\omega(x,y) = 0$$

$$EV_\Delta^2 = \frac{1}{\|\Delta\|^2}\iint_\Delta \iint_\Delta R(x-\xi, y-\eta)\,dx\,d\xi\,dy\,d\eta. \tag{7}$$

Therefore, (6) finally yields

$$E\epsilon_\Delta^2 = R(0,0) - \frac{1}{\|\Delta\|^2}\iint_\Delta \iint_\Delta R(x-\xi, y-\eta)\,dx\,d\xi\,dy\,d\eta. \tag{8}$$

If the region Δ is a rectangle, i.e., $x \in [-A_x/2, A_x/2]$ and $y \in [-A_y/2, A_y/2]$, we then have further

$$EV_\Delta^2 = \frac{1}{A_xA_y}\int_{-A_y}^{A_y}\int_{-A_x}^{A_x}\left(1-\frac{|\tau_x|}{A_x}\right)\left(1-\frac{|\tau_y|}{A_y}\right) \cdot R(\tau_x,\tau_y)\,d\tau_x\,d\tau_y \tag{9}$$

by repeatedly invoking the result that [2, p. 325]

$$\frac{1}{(2T)^2}\int_{-T}^{T}\int_{-T}^{T}R(x-y)\,dx\,dy = \frac{1}{2T}\int_{-2T}^{2T}R(\tau)\left(1-\frac{|\tau|}{2T}\right)d\tau. \tag{10}$$

Now (8) becomes

$$E\epsilon_\Delta^2 = R(0,0) - \frac{1}{A_xA_y}\int_{-A_x}^{A_x}\int_{-A_y}^{A_y}\left(1-\frac{|\tau_x|}{A_x}\right)\left(1-\frac{|\tau_y|}{A_y}\right) \cdot R(\tau_x,\tau_y)\,d\tau_x\,d\tau_y. \tag{11}$$

Assume now that the value V_Δ is quantized to L discrete levels in order to be represented by $\log_2 L$ bits. The spatial mean-square error induced by a zero-order-hold interpolation, when the quantized representation value V_Δ^Q will be used as $\tilde{f}_\omega(x,y)$ over Δ, is given by

$$(\epsilon_\Delta^Q)^2 = \frac{1}{\|\Delta\|}\iint_\Delta \left[(f_\omega(x,y)-V_\Delta)+(V_\Delta - V_\Delta^Q)\right]^2 dx\,dy. \tag{12}$$

This can be written as

$$(\epsilon_\Delta^Q)^2 = \frac{1}{\|\Delta\|}\iint_\Delta (f_\omega(x,y)-V_\Delta)^2\,dx\,dy + 2(V_\Delta - V_\Delta^Q)\frac{1}{\|\Delta\|}\iint_\Delta (f_\omega(x,y)-V_\Delta)\,dx\,dy + (V_\Delta - V_\Delta^Q)^2 \tag{13}$$

and the definition of V_Δ ensures that the cross term above vanishes, yielding

$$E(\epsilon_\Delta^Q)^2 = E\epsilon_\Delta^2 + E(V_\Delta - V_\Delta^Q)^2. \tag{14}$$

The mean-square error induced by quantizing a random variable with variance σ^2 is roughly proportional to the ratio between σ^2 and the square of the number of quantization levels; see, e.g., [3]. Thus, we can write that

$$E(V_\Delta - V_\Delta^Q)^2 = K_Q\frac{EV_\Delta^2}{L^2} \quad \text{with} \quad 2.72 > K_Q > 1 \tag{15}$$

and this yields the following expression for the error due to the zero-order-hold interpolation with a quantized representation (on a rectangle):

$$E(\epsilon_\Delta^Q)^2 = R(0,0) + \left\{\frac{1}{A_xA_y}\int_{-A_x}^{A_x}\int_{-A_y}^{A_y}\left(1-\frac{|\tau_x|}{A_x}\right)\left(1-\frac{|\tau_y|}{A_y}\right) \cdot R(\tau_x,\tau_y)\,d\tau_x\,d\tau_y\right\}\left(\frac{K_Q}{L^2}-1\right). \tag{16}$$

The above expression depends on A_x, A_y, and L and these variables are constrained in the digitization process. Indeed, suppose that we are given the sample function of a stationary process, and we wish to digitize it by using a total number of B bits that should be allocated between spatial resolution and quantization depth. Then, dividing the area Ω into $M \times N$ similar rectangular cells (pixels) $\{\Delta(i,j)\}$ over which the sample function is replaced by quantized $V_{\Delta(i,j)}(=f_\omega(i,j))$'s induces a total average squared error of

$$(\epsilon_{tot}^Q)^2 = \iint_\Omega \left(f_\omega(x,y)-\tilde{f}_\omega(x,y)\right)^2 = \frac{1}{MN}\sum_i\sum_j (\epsilon_{\Delta(i,j)}^Q)^2. \tag{17}$$

The mean-squared error over the individual pixels has the same expectation for each pixel; therefore, the expected mean-square error in the digitization process is given exactly by (16) above. This expression depends on the spatial resolution variables A_x and A_y and on the quantization depth L, and these variables are constrained in the digitization process since

$$A_x = \frac{1}{M},\ A_y = \frac{1}{N} \quad \text{and} \quad \log_2 L = b \tag{18}$$

and the product $M \times N \times b$ has, by assumption, to equal a given constant B, the total number of bits allocated to represent the image. Therefore, the optimal digitization parameters are obtained by solving the following minimization problem:

$$\text{minimize } E(\epsilon_\Delta^Q)^2 \quad \text{subject to} \quad \frac{1}{A_xA_y}\log_2 L = B. \tag{19}$$

This minimization problem can be solved numerically for general $R(\tau_x,\tau_y)$; however, in an important special case, explicit solutions are obtained. This occurs when $f_\omega(x,y)$ is a separable Markov process.

III. An (Almost) Explicit Solution for Markov Processes

Suppose that the process $f_\omega(x,y)$ is a separable, first-order, 2-D Markov process, i.e., that it has a covariance of the form $R(\tau_x,\tau_y) = R_x(\tau_x)R_y(\tau_y)$ with (see [4])

$$R_x(\tau_x) = e^{-\alpha_x|\tau_x|} \quad \text{and} \quad R_y(\tau_y) = e^{-\alpha_y|\tau_y|}. \tag{20}$$

Then $E(\epsilon_\Delta^Q)^2$ becomes

$$1 + \left\{\frac{1}{A_x}\int_{-A_x}^{A_x}\left(1-\frac{|\tau_x|}{A_x}\right)R_x(\tau_x)\,d\tau_x\,\frac{1}{A_y} \cdot \int_{-A_y}^{A_y}\left(1-\frac{|\tau_y|}{A_y}\right)R_y(\tau_y)\,d\tau_y\right\}\left(\frac{K_Q}{L^2}-1\right). \tag{21}$$

For first-order Markovian statistics, we have from

$$\frac{1}{T}\int_{-T}^{T}\left(1-\frac{|\tau|}{T}\right)e^{-\alpha|\tau|}\,d\tau = \frac{2}{T^2\alpha^2}(T\alpha - 1 + e^{-\alpha T}) \tag{22}$$

that

$$E(\epsilon_\Delta^Q)^2 = 1 - \frac{2}{A_x^2\alpha_x^2}(A_x\alpha_x - 1 + e^{-\alpha_xA_x})\frac{2}{A_y^2\alpha_y^2} \cdot (A_y\alpha_y - 1 + e^{-\alpha_yA_y})\left(1-\frac{K_Q}{L^2}\right). \tag{23}$$

If we assume that A_x and A_y are small enough so that also $\alpha_x A_x$, $\alpha_y A_y$ are small, we have that

$$E(\epsilon_\Delta^Q)^2 \cong 1 - \left(1 - \frac{1}{3}\alpha_x A_x\right)\left(1 - \frac{1}{3}\alpha_y A_y\right)\left(1 - \frac{K_Q}{L^2}\right) \quad (24)$$

and we wish to minimize (24) subject to $(1/A_xA_y)\log_2 L = B$. A standard Lagrange multiplier method provides that the product $(1 - \frac{1}{3}\alpha_x A_x)(1 - \frac{1}{3}\alpha_y \Delta_y)(1 - (K_Q/L^2))$ is maximized (i.e., $E(\epsilon_\Delta^Q)^2$ is minimized) for

$$\alpha_x A_x = \alpha_y A_y = P(\text{constant}). \quad (25)$$

The optimal L is then found by observing that the constraint

$$\frac{1}{A_xA_y}\log_2 L = B \quad \text{implies} \quad \log_2 L = \frac{BP^2}{\alpha_x\alpha_y}, \quad (26)$$

and therefore, P_{opt} maximizes

$$(1 - \tfrac{1}{3}P)^2\left(1 - K_Q e^{-2(\ln 2)BP^2/\alpha_x\alpha_y}\right) \triangleq \psi(P). \quad (27)$$

From $(\partial/\partial P)\,\psi(P) = 0$, we conclude that P_{opt} satisfies the equation

$$\frac{1}{6}\frac{\alpha_x\alpha_y}{\ln 2 B}\left(\frac{1}{K_Q}e^{2(\ln 2)BP^2/\alpha_x\alpha_y} - 1\right) = P\left(1 - \frac{1}{3}P\right). \quad (28)$$

The right-hand side (RHS) of this equation is a quadratic function positive in the interval $P \in [0, 3]$, and achieves its maximum at $P = \frac{3}{2}$, the maximal value being $\frac{3}{4}$. The left-hand side (LHS) is an exponential function increasing from a slightly negative, since $(K_Q > 1)$, value at $P = 0$ to $+\infty$. Therefore, (28) has a unique solution that may be determined numerically. In order to obtain an upper bound for P_{opt}, let us ask what is $\tilde{P}$ for which the LHS reaches $\frac{3}{4}$. Surely, then, $P_{\text{opt}} < \tilde{P}$. Now the equation

$$\frac{1}{3}\frac{\alpha_x\alpha_y}{\ln 2 B}\left(\frac{1}{K_Q}e^{2(\ln 2)(B\tilde{P}^2/\alpha_x\alpha_y)} - 1\right) = \frac{3}{2} \quad (29)$$

yields

$$\tilde{P}(B) = \left\{\frac{\alpha_x\alpha_y}{2(\ln 2)B}\ln\left[\left(\frac{9}{2}\frac{(\ln 2)B}{\alpha_x\alpha_y} + 1\right)K_Q\right]\right\}^{1/2}. \quad (30)$$

Note that $\tilde{P}(B)$ is very small indeed for large B's since

$$\frac{\ln B}{B} \to 0 \quad \text{as} \quad B \to \infty. \quad (31)$$

Therefore, we have

$$P_{\text{opt}} < \left\{\frac{\alpha_x\alpha_y}{2(\ln 2)B}\ln\left[\left(\frac{9}{2}\frac{(\ln 2)B}{\alpha_x\alpha_y} + 1\right)K_Q\right]\right\}^{1/2} \quad (32)$$

and the initial assumption that $A_x\alpha_x$ and $A_y\alpha_y$ are small is satisfied. We also note that (for large values of B) the LHS of (28) is a very steeply increasing function of P; therefore, the difference between P_{opt} and $\tilde{P}(B)$ is small and decreases with increasing B. $\tilde{P}(B)$ thus becomes, for large B, a very good approximation for P_{opt}. We have then

$$A_x = P_{\text{opt}}/\alpha_x \cong \left\{\frac{\alpha_y}{\alpha_x}\frac{1}{2\ln 2}\frac{1}{B}\ln\left[\left(\frac{9}{2}\frac{(\ln 2)B}{\alpha_x\alpha_y} + 1\right)K_Q\right]\right\}^{1/2}$$

$$A_y = P_{\text{opt}}/\alpha_y \cong \left\{\frac{\alpha_x}{\alpha_y}\frac{1}{2\ln 2}\frac{1}{B}\ln\left[\left(\frac{9}{2}\frac{(\ln 2)B}{\alpha_x\alpha_y} + 1\right)K_Q\right]\right\}^{1/2} \quad (33)$$

which shows that the ratio between the x and y resolutions is directly controlled by the spatial correlation distances $1/\alpha_x$ and $1/\alpha_y$, not an unexpected result. It is also interesting to note the similarity between our results (33) and those of Nielsen *et al.* [1, eq. (9), (10)] with correlation distances taking the place of mean directional fluctuation rates.

IV. Concluding Remarks

We presented an approach to optimal image digitization, based on standard integral sampling and assuming that the image is the realization of a 2-D homogeneous process with known autocorrelation. We note that since standard integral sampling is but one method of representing an image via an orthogonal basis of functions, one could consider other optimal digitization problems as well. We could assume, for example, that the image is represented by using the first (in terms of energy or variance ordering) $M \times N$ eigenfunctions of the covariance (Karhunen–Loève expansion), and that the corresponding coefficients are digitized to a certain number of bits. This leads to a different, very interesting optimization problem allocating B bits among MN uncorrelated random variables (the coefficients of the KL expansion) according to their variances so as to minimize the mean-square reconstruction error. If the product MN is given *a priori*, the problem leads to a classical bit-allocation process (see [3]–[5]); however, we do not have such prior constraints here. We have to determine both the product MN and the corresponding allocation of resources (bits or integer quantization levels) to minimize the expected mean-square error in reconstructing the image.

Other image representation, sampling, and reconstruction methods could be considered as well. In [1], Nielsen *et al.* considered point sampling with zero-order-hold reconstruction, and found the parameters of the optimal digitization under the assumption that the value range and mean directional fluctuations of the 2-D function are given. We could also consider point sampling in conjunction with the ideal low-pass filter reconstruction, with parameters adjusted to correspond to the sampling grid. If the process is band-limited and if we increase the number of bits to be allocated to infinity, it would be nice to show that the spatial sampling grid approaches the Nyquist grid, more and more bits being allocated, not to spatial resolution, but rather to increase the quantization depth.

The digitization process described in this correspondence assumed that a total number of B bits has to be allocated between the spatial resolution and the quantization depth so as to minimize the total mean-square reconstruction error. However, in some applications, we might be ready to allocate a variable number of bits per image in order to bound the error to some predetermined value ϵ. In fact, we can formulate a general optimal digitization problem as follows. We define two cost functionals $\mathbb{C}_e(\cdot)$ and $\mathbb{C}_b(\cdot)$, both monotonically increasing in their arguments so that the total cost induced by a certain digitization and reconstruction process is

$$\mathbb{C}_{\text{TOTAL}} = \mathbb{C}_e\left(E(\epsilon_\Delta^Q)^2\right) + \mathbb{C}_b\left(\frac{1}{A_xA_y}\log_2 L\right). \quad (34)$$

$\mathbb{C}_{\text{TOTAL}}$, of course, depends on A_x, A_y, and L, and in it, $\mathbb{C}_e$ is the penalty function weighting the expected mean-square error due to a digitization $A_x = 1/M$, $A_y = 1/N$, $\log_2 L = b$, and $\mathbb{C}_b$ is the cost of bit usage. The optimal digitization can be defined as the one that minimizes $\mathbb{C}_{\text{TOTAL}}$. It is clear that by using various forms for $\mathbb{C}_e$ and $\mathbb{C}_b$, we can put both the previous bit allocation problem and the predetermined error bound problem into this framework (see, e.g., [5]).

References

[1] L. Nielsen, K. J. Astrom, and E. I. Jury, "Optimal digitization of 2-D images," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1247–1249, Dec. 1984.

[2] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill, 1965.

[3] J. J. Y. Huang and P. M. Scultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Commun. Syst.*, vol. CS-11, pp. 289–296, Sept. 1963.

[4] A. Rosenfeld and A. Kak, *Digital Picture Processing, Vol. 1*. New York: Academic, 1982.

[5] A. M. Bruckstein, "On soft bit allocation," Technion—Israel Inst. Technol., Haifa, EE Rep. 575, Jan. 1986 (also to appear in *IEEE Trans. Acoust., Speech, Signal Processing*).