Best linear unbiased estimators for properties of straight lines

Dorst, L.; Smeulders, A.W.M.

Published in:
IEEE Transactions on Pattern Analysis and Machine Intelligence

DOI:
10.1109/TPAMI.1986.4767781

[9] T. Kishimoto and Y. Sato, "Simultaneous transmission of voices and handwriting signals: Sketchphone," *IEEE Trans. Commun.*, vol. COM-29, Dec. 1981.
[10] K. Kondo *et al.*, "Chain encoding of telewriting signals using linear interpolation" (in Japanese), IECE Japan Tech. Rep. IE 83-107, 1983.
[11] T. Kaneko and M. Okuhara, "An efficient data compression for chain-coded line drawings" (in Japanese), IECE Japan, Tech. Rep. IE83-82, Nov. 1983.
[12] K. Shinohara and T. Minami, "Data compression for chain-coded line drawings by level code" (in Japanese), in *Proc. 1984 IIEE Japan Nat. Conf. Rec.*, no. 12-24, June 1984.
[13] T. Minami and K. Shinohara, "Multiple grid chain coding," presented at the 1984 Picture Coding Symp., no. 6-1, Cession-Sevigne, France, July 1984.
[14] J. Koplowitz, "On the performance of chain codes for quantization of line drawings," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-3, Mar. 1981.
[15] L. A. Santalo, "Geometry and geometric probability," in *Encyclopedia of Mathematics and Its Applications*, vol. 1. Reading, MA: Addison-Wesley, 1976, pp. 27-34.
[16] Y. Sato *et al.*, "Telewriting system" (in Japanese), *Electric. Comm. Lab. Tech. J.*, vol. 32, Mar. 1983.

Best Linear Unbiased Estimators for Properties of Digitized Straight Lines

LEO DORST AND ARNOLD W. M. SMEULDERS

***Abstract*—This paper considers the problem of measuring properties of digitized straight lines from the viewpoint of measurement methodology. The measurement and estimation process is described in detail, revealing the importance of a step called "characterization" which was not recognized explicitly before. Using this new concept, BLUE (Best Linear Unbiased) estimators are found. These are calculated for various properties of digitized straight lines, and are briefly compared to previous work.**

***Index Terms*—BLUE estimators, chain-code string, digitization error, digitized straight lines, length measurement, measurement accuracy, quantization.**

I. Introduction

This paper aims at finding optimal estimators for properties of digitized straight lines in an image. To find these, we first need a precise understanding of the measurement process since this will reveal how improvement over existing methods can be achieved. This knowledge can then be used to arrive at "optimal" estimators, in some specified sense.

The measurement situation is the following. Before digitization, there is a line in the continuous world with specific properties (such as slope, length, intercept, etc.) If one wishes to measure a property, a *digitization* of the continuous line is performed, leading to a chain-code string. This digitization reduces the information in an essential way since it maps a set of continuous lines into a set of discrete strings. Therefore, exact measurement is impossible; the best one can do is *estimate* the continuous property from the string.

We will discriminate two steps in this estimation procedure (Section II): a *characterization* in which the information present in the actual chain-code string is reduced to some characterizing parameters, and a *calculation* in which these parameters are used in a formula for an estimator of the property. The results by other investigators on measuring line length [1]-[4] may also be expressed in these two terms (Section III).

The importance of this unraveling of the estimation procedure (recognized here for the first time) follows from the fact that given the digitization, one can optimize estimators in two independent ways: by improving the characterization step, and by improving the calculation, i.e., the formula used as an estimator. Both are considered in Section IV. It is shown that to every characterization there corresponds a "BLUE" estimator, which is the optimal estimator for that particular characterization (optimal in the sense of minimal MSE, linearity, and being unbiased). The search for optimal estimators therefore becomes a search for an optimal characterization. It is shown that a so-called "faithful characterization," in which no information is lost, results in optimal BLUE estimators.

In Section V, this basic result is applied to the measurement of the length per chain-code of a digital straight line.

II. A Formal Description of the Measurement Process

In this section, the digitization and measurement process is described in detail. This is necessary to arrive at a precise formulation of the optimal measurement methods.

All continuous straight lines form a set $\mathcal{L}$. An element l of this set can be characterized by intercept e and slope α in a Cartesian grid (l: $y = \alpha x + e$). A property f can be associated with each l. For instance, $f(e, \alpha) = \alpha$ represents the slope of l.

Properties of the continuous line l are to be measured after *digitization* symbolized by the operator D. Digitization results in a string c where $c = Dl$. Digitization of all lines in $\mathcal{L}$ results in the set of all straight chain-code strings $\mathcal{C}$.

Given a string c, there is an *equivalence class* of continuous lines, all having the same string c as its digitization. This equivalence class is called the *domain* $\mathfrak{D}_D(c)$ of c corresponding to the digitization D.

Formal definition:

$$\mathfrak{D}_D(c) = \{l \in \mathcal{L} \mid Dl = c\}. \tag{1}$$

Thus, the set of domains indicates the finest distinction of the original continuous lines that is still possible after digitization.

The digitized line c can be represented in various ways, for instance, as a string of (x, y) coordinates, a series of run codes, or the Freeman directional code [1]. Furthermore, one can digitize in different ways, such as grid intersection quantization (GIQ) or object boundary quantization (OBQ) [2]. The difference between these methods is not essential to the basic idea of the present paper. Specific results will be given for OBQ, 8-connected Freeman chain codes. For this case, the domains have been given in a previous paper [5]. In Fig. 1, they are depicted for all lines having a chain code consisting of 6 codes 0 or 1. The representation is in a part of the parameter space (e, α) where each point (e, α) represents a continuous line $l: y = \alpha x + e$.

The estimation of the original continuous property f is to be based on the discrete string resulting from the digitization. This is done by *characterization* and *calculation*. The characterization K reduces the chain-code string c to a tuple t of parameters (where $t = Kc$) characterizing the string. For instance, one might characterize a string by the number of odd and even Freeman codes in the string. Characterizing all strings of $\mathcal{C}$ leads to the set of all tuples $\mathfrak{I}$. A tuple can be used in an *estimator* g which attributes a value $g(t)$ to the tuple t. This *value* $g(t) = g(Kc) = g(KDl)$ is used as an *estimate* of the original continuous property $f(l)$, based on the digitization Dl. The recognition of the characterization K is essential in optimizing the estimator g of the property f.

In the same way as domains are the equivalence classes into which the set of lines is divided by digitization, we have equiva-

Manuscript received August 6, 1984; revised July 15, 1985.
L. Dorst is with the Department of Applied Physics, Delft University of Technology, 2628 CJ Delft, The Netherlands.
A.W.M. Smeulders is with the Department of Pathology and Medical Informatics, Free University, Amsterdam, The Netherlands.
IEEE Log Number 8407172.

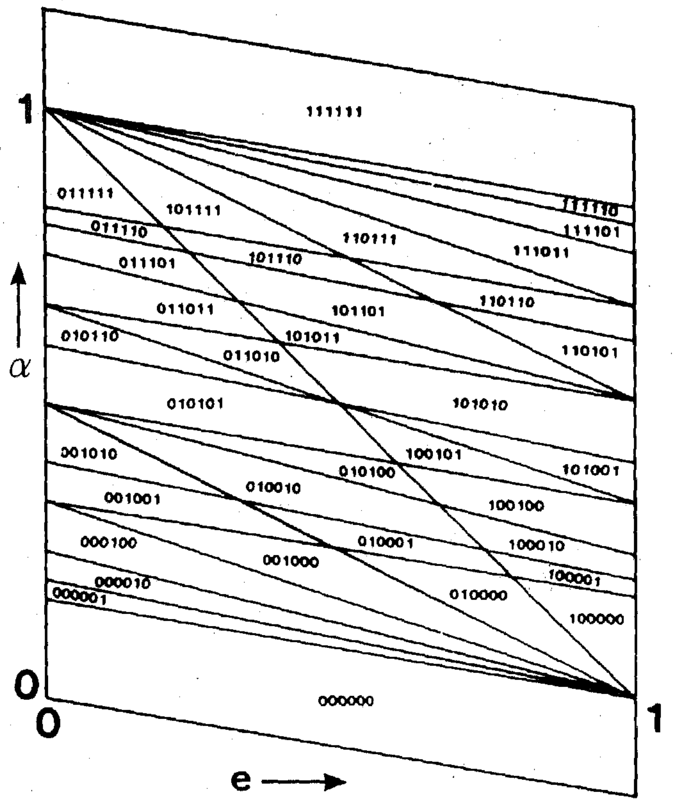

Fig. 1. The domains of the OBQ digitization, represented in the (e, α) plane for all strings consisting of 6 elements 0 or 1 (from [5]).

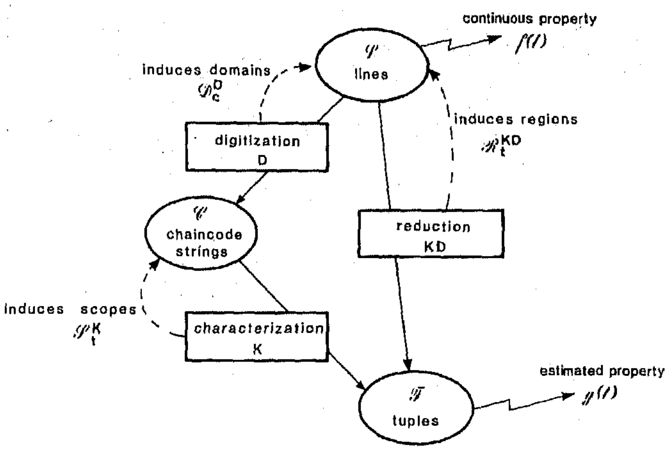

Fig. 2. A schematic representation of the measurement process for properties of digitized straight lines.

lence classes into which the set of strings is divided by characterization. Therefore, we introduce the *scope* $\mathcal{S}_K(t)$ of a tuple t as the equivalence class of all strings having the same tuple t under the characterization K:

$$\mathcal{S}_K(t) = \{c \in \mathcal{C} \mid Kc = t\}. \tag{2}$$

Taking digitization and characterization together, the original set of lines $\mathcal{L}$ is mapped into the set $\mathcal{T}$ of all tuples. The equivalence classes of this mapping will be called *regions*. Thus, the region $\mathcal{R}_{K,D}(t)$ of a tuple t is the set of all lines having the same tuple t after digitization D and characterization K:

$$\mathcal{R}_{K,D}(t) = \{l \in \mathcal{L} \mid KDl = t\}. \tag{3}$$

We will often omit the subscripts D and K and write $\mathcal{D}(c)$, $\mathcal{S}(t)$, and $\mathcal{R}(t)$ if it is clear which digitization and characterization are meant.

Fig. 2 summarizes the terms introduced. A relation between the different equivalence classes which follows immediately from the definitions is

$$\mathcal{R}_{K,D}(t) = \bigcup_{c \in \mathcal{S}_K(t)} \mathcal{D}_D(c). \tag{4}$$

From this description, it is seen that at two points a (potential) loss of information occurs. Digitization unavoidably implies loss of information since it maps a continuous set $\mathcal{L}$ into a discrete set $\mathcal{C}$. Characterization, however, maps one discrete set ($\mathcal{C}$) into another ($\mathcal{T}$). Here loss of information is avoidable if the characterization is chosen properly. Let us use the symbol F to signify a *faithful characterization*, i.e., a characterization that is a bijective mapping between a string c and its corresponding tuple t. In that case, the scope $\mathcal{S}_F(t)$ consists of a single string $c = F^{-1}t$ and the region $\mathcal{R}_{F,D}(t)$ is equivalent to the domain $\mathcal{D}_D(F^{-1}t)$. Thus, a faithful characterization causes no loss of information.

III. Description of Previous Work

Extensive work has been done for almost 15 years in the measurement of the length of a digitized straight line segment [1]–[4]. This work can be described within the framework of terms associated with the measurement process introduced in the previous section.

A very simple and straightforward way to associate a length to a given chain-code string might be the total number n of chain codes in the string. In this case, the string is characterized by the "tuple" (n); the regions in $\mathcal{L}$ corresponding to this (n) characterization are indicated in (e, α) space in Fig. 3(a). As a length estimator, we have

$$l_S(n) = n. \tag{5}$$

This estimator is too simple, and not often used in practice; it is introduced here as an illustration.

Freeman [1] based a length estimate on the number of even and odd codes in a string n_e and n_o. In terms of the present paper, he used an (n_e, n_o) characterization. This results in a finer division of $\mathcal{L}$ into regions than the (n) characterization [Fig. 3(b)]. But still, many strings are lumped together. The length estimator given in [1]

$$l_F(n_e, n_o) = n_e + \sqrt{2}n_o \tag{6}$$

counts even codes as having length 1 and odd codes as $\sqrt{2}$. Freeman thus gave an exact measure of the length of the *digital* arc. But as Groen and Verbeek [2] and Proffitt and Rosen [3] almost simultaneously realized, this is not necessarily a good measure of the length of the arc *before* digitization. The estimator (6) has a limit root mean-square error of 6.6 percent for long strings (see Table I).

Groen and Verbeek [2] used the same (n_e, n_o) characterization, but calculated the probability of occurrence of even and odd codes based on the distribution of continuous lines. This led to different coefficients for the length estimator, namely,

$$l_G(n_e, n_o) = 1.059n_e + 1.183n_o. \tag{7}$$

This estimator is somewhat less accurate than l_F (Table I).

Proffitt and Rosen [3] independently calculated a length estimator, again based on the (n_e, n_o) characterization, which also took into account the relative probabilities of even and odd codes. In contrast to [2] where the probabilities were calculated for strings with $n = 1$, they performed their calculations for the case $n \to \infty$, obtaining

$$l_P(n_e, n_o) = 0.984n_e + 1.340n_o. \tag{8}$$

(Actually, this estimator can be found in [4] since there the basic idea which was applied to 4-connected strings in [3] was applied to 8-connected strings.) This estimator has a limit root MSE of 2.7 percent, which is considerably better than l_F (see Table I).

In the same paper [3], an essential new step was taken, which we would now describe as choosing a new characterization. A new

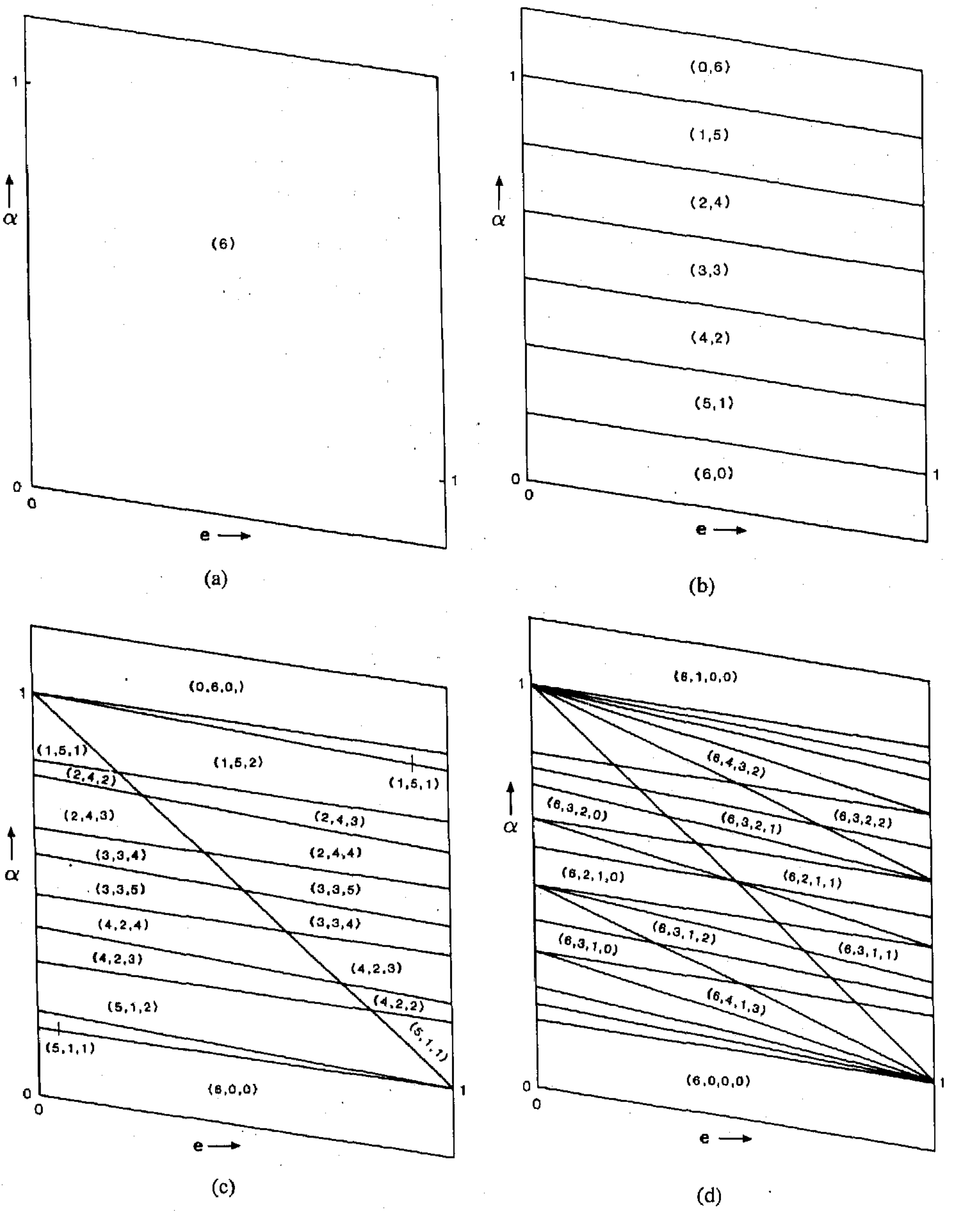

Fig. 3. The regions corresponding to various characterizations, represented in the (e, α) plane). Compare Fig. 1. (a) (n) characterization. (b) (n_e, n_o) characterization. (c) (n_e, n_o, n_c) characterization (note that some disconnected polygons belong to one region). (d) (n, q, p, s) characterization (for clarity, not all labels are indicated).

TABLE I
A COMPARISON OF LENGTH ESTIMATORS

Estimator	BLUEness	1	2	5	10	20	50	100
$l_F(n_e, n_o)$	Biased	0.223	0.117	0.076	0.069	0.067	0.067	0.066
$l_G(n_e, n_o)$	BLUE for $n = 1$, otherwise biased	0.217[a]	0.141	0.103	0.091	0.085	0.082	0.080
$l_P(n_e, n_o)$	Unbiased for $n \to \infty$, not BLUE	0.232	0.114	0.053	0.037	0.031	0.028	0.027
$l_C(n_e, n_o, n_c)$	Unbiased for $n \to \infty$, not BLUE	0.228	0.103	0.040	0.021	0.013	0.0088	0.0081
$l_V(n_e, n_o, n_c)$	BLUE for (n_e,n_o,n_c) char., optimal BLUE for $n \leq 4$	0.217[a]	0.104[a]	0.033	0.014	0.0064	0.0024	0.0017
$l_O(n, q, p, s)$	Optimal BLUE, is faithful characterization	0.197	0.094	0.029	0.011	0.0038	0.0010	0.0004

[a] The difference of the entries marked with footnote [a] with the optimal BLUE estimator l_O can be explained by a different treatment of strings consisting solely of odd or even codes in [2] and [4].

parameter was introduced, the so-called *corner count* n_c, being the number of code transitions (01 and 10 sequences) in the chain-code string. This extends the characterization to an (n_e, n_o, n_c) characterization, of which the corresponding regions in $\mathcal{L}$ are indicated in Fig. 3(c). The length estimate proposed in [3] is

$$l_C(n_e, n_o, n_c) = 0.980n_e + 1.406n_o - 0.091n_c. \tag{9}$$

obtained by least squares approximation of a linear formula to infinitely long strings. (Again, for 8-connectivity, this estimator is found in [4] rather than [3].) This estimator tends to a limit accuracy of approximately 0.8 percent (Table I). The increase in accuracy in going from l_P to l_C shows the effect of the extra parameter n_c, and hence the importance of the choice of the characterization. With l_C, the length estimators are better "tuned" to the strings for which they are used, and this is the main reason for their increased accuracy.

The next step to accurate length measurement was taken in a paper by Vossepoel and Smeulders [4]. They applied the same (n_e, n_o, n_c) characterization, but did not use a linear formula in the parameters of the tuple. Instead, they arrived at estimators for the length corresponding to a tuple (n_e, n_o, n_c) by averaging the lengths per chain code of all lines (e, α) in the region $\mathcal{R}(n_e, n_o, n_c)$ corresponding to this tuple. Note that this optimizes estimators per tuple (n_e, n_o, n_c) rather than for all lines of a specific number of chain codes, as in [3]. The integration can be performed, but leads to complicated formulas

$$l_V(n_e, n_o, n_c) = g(n_e, n_o, n_c) \tag{10}$$

where the function g can be found in [4]. This estimator is much more accurate than the previous methods (Table I). A comparison to l_C, which is also based on the (n_e, n_o, n_c) characterization, shows the importance of optimizing the "calculation" step in the estimation procedure.

In a previous paper [5], we defined four parameters (n, q, p, s) which can be extracted from a straight string (i.e., a string that could be the digitization of a straight line), and showed that there is a unique correspondence between the string and this quadruple. In terms of the present paper, this (n, q, p, s) characterization is a faithful characterization. As is illustrated in Fig. 3(d), this faithful characterization leads to the finest tesselation of the (e, α) plane still possible after digitization, and therefore a length estimator can be tuned to each individual chain-code string. It is obvious from the previous that this (n, q, p, s) characterization can potentially lead to the most accurate estimators possible. In the next section, we will show that is indeed the case.

IV. BLUE Estimators

In the previous section, it was seen that estimators based on increasingly finer characterizations can be better tuned to individual strings, and can thus in principle be a more accurate estimate for the original continuous property. It was also seen that given the characterization, still many formulas are possible for the estimator.

In this section, we will show that for each characterization, there exists an "optimal" estimator in the sense that an unbiased estimate of the length is provided, with minimal mean-square error (MSE). This kind of estimator is well known from parameter estimation theory, and is called BLUE: Best (minimal MSE), Linear (being an average over the "observations"), Unbiased Estimator. In the case of measuring line properties, the continuous property $f(l)$ of the line l is estimated by a function $g(t)$ based on the tuple $t = Kc$ of the string $c = Dl$ corresponding to the line l. The requirements for $g(t)$ to be a BLUE estimator $f(l)$ are as follows.

1) The estimator should be linear in $f(l)$. This implies that the estimator $g(t)$ for a tuple t should have the form

$$g(t) = E_{\mathcal{A}(t)}\{f(l)\} \tag{11}$$

where $E_{\mathfrak{X}}\{x\}$ denotes the expectation of x over a set $\mathfrak{X}$ and $\mathcal{A}(t)$ is some set dependent on the tuple t.

2) The estimator $g(t)$ should be an unbiased estimate of $f(l)$ over the set of all straight line segments $\mathcal{L}$:

$$E_{\mathcal{L}}\{f(l) - g(t)\} = 0. \tag{12}$$

3) Of all estimators satisfying (11) and (12), the BLUE estimator $g_K(t)$ for a given characterization K should have minimal MSE over $\mathcal{L}$:

$$E_{\mathcal{L}}\{[f(l) - g_K(t)]^2\} \text{ minimal.} \tag{13}$$

Note that in all three requirements, $t = Kc$ denotes the tuple corresponding to the string $c = Dl$, so $t = KDl$.

It turns out that the estimator obtained by attributing to a tuple t the expectation of $f(l)$ over the region $\mathcal{R}(t)$ is BLUE.

Theorem 1: The estimator

$$g_K(KDl) = E_{\mathcal{R}_{K,D}(KDl)}\{f(l)\} \tag{14}$$

is BLUE.

Proof:

1) $g_K(t)$ is a linear estimator, as follows immediately from (11).

2) Consider a single region $\mathcal{R}_{K,D}(KDl)$. Omitting the subscripts, we have

$$E_{\mathcal{R}(KDl)}\{f(l) - g_K(KDl)\} = E_{\mathcal{R}(KDl)}\{f(l)\} - g_K(KDl) = 0$$

Thus, $g_K(t)$ is unbiased for a region, and hence also for a collection of regions, such as $\mathcal{L}$. So, (12) is satisfied. Note that $\mathcal{R}(KDl)$ is the smallest set of lines still distinguisable using the tuple t; averaging over a smaller set than $\mathcal{R}(KDl)$ would result in a biased estimator.

3) Comparing the general estimator $g(KDl)$ in (11) to $g_K(KDl)$ in (14), regarding the MSE over a region $\mathcal{R}_{K,D}(KDl)$ [abbreviated as $\mathcal{R}(KDl)$], we have

$$\begin{aligned} &E_{\mathcal{R}(KDl)}\{[f(l) - g(KDl)]^2\} \\ &\quad= E_{\mathcal{R}(KDl)}\{[f(l) - g_K(KDl)]^2\} \\ &\qquad+ E_{\mathcal{R}(KDl)}\{[g_K(KDl) - g(KDl)]^2\} \\ &\quad\geq E_{\mathcal{R}(KDl)}\{[f(l) - g_K(KDl)]^2\}. \end{aligned}$$

Hence, g_K has a smaller MSE than any linear unbiased estimator based on averaging over more than one region. Hence, it is the Best Linear, Unbiased Estimator. Q.E.D.

Thus, for any given characterization K, the BLUE estimator for a given t is obtained by averaging over the region $\mathcal{R}_{K,D}(t)$. An example is the estimator $l_V(n_e, n_o, n_c)$ given in formula (10), which is the BLUE estimator for the (n_e, n_o, n_c) characterization.

If the characterization is faithful, the regions reduce to domains. These are the smallest possible sets of lines distinguishable after digitization, and the BLUE estimators corresponding to this faithful characterization are therefore the most accurate estimators possible, given the digitization D. This is expressed in the following theorem.

Theorem 2: Of all BLUE estimators

$$g_K(KDl) = E_{\mathcal{R}_{K,D}(KDl)}\{f(l)\}$$

the estimator g_F, corresponding to a faithful characterization F, has minimal MSE.

Proof: Consider the MSE over a domain $\mathfrak{D}_D(Dl) = \mathcal{R}_{F,D}(FDl)$ [abbreviated to $\mathfrak{D}(Dl)$]:

$$\begin{aligned} &E_{\mathfrak{D}(Dl)}\{[f(l) - g_K(KDl)]^2\} \\ &\quad= E_{\mathfrak{D}(Dl)}\{[f(l) - g_F(FDl)]^2\} \\ &\qquad+ E_{\mathfrak{D}(Dl)}\{[g_F(FDl) - g_K(KDl)]^2\} \\ &\quad\geq E_{\mathfrak{D}(Dl)}\{[f(l) - g_F(FDl)]^2\}. \end{aligned}$$

Hence, the MSE of g_F is smaller than that of an arbitrary g_K, unless $K = F$. As in the proof of Theorem 1, no decrease of the set over which is averaged is possible beyond $\mathcal{R}_{F,D}(FDl) = \mathfrak{D}_D(Dl)$, without resulting in an unbiased estimator. Therefore, g_F is the optimal BLUE estimator. Q.E.D.

In principle, this solves the optimal estimation problem, not only for straight lines, but for any two-step digitization and characterization measurement process. To derive the result of Theorem 2, only the terminology was borrowed from the straight line case. In all cases, the estimator

$$g_F(FDl) = E_{\mathcal{R}_{F,D}(FDl)}\{f(l)\} \tag{15}$$

is "optimal BLUE."

Note that since $\mathcal{R}_{F,D}(FDl) = \mathfrak{D}_D(Dl)$, the estimator (15) can be written as

$$g_F(FDl) = E_{\mathfrak{D}_D(Dl)}\{f(l)\} \tag{16}$$

which is independent of the specific faithful characterization used (as it should be).

To evaluate (15) in a particular situation, one needs

- a faithful characterization F
- an expression of the domains $\mathfrak{D}_D(Dl)$ in terms of this faithful characterization, as regions $\mathcal{R}_{F,D}(FDl)$
- calculation of the expectation of any desired function f over this domain.

For straight lines, this will be done in the next section.

V. Optimal BLUE Estimators for Properties of Straight Lines

To evaluate (15) for properties of straight lines, we need both a faithful characterization and an expression for the domains. Both have already been given in [5]. Here we repeat these results in the terminology of the present paper.

Main Theorem (from [5])

A straight string c (with ith element c_i) can be faithfully characterized by the quadruple (n, q, p, s) where

n is the number of elements of c,

q is the shortest period present in c or any of its extensions:

$$q = \min_k \{k \in \{1, 2, \cdots, n\} | k = n \vee \forall i \in \{1, 2, \cdots, n-k\} : c_{i+k} = c_i\},$$

p is the number of odd codes in a period q:

$$p = \sum_{i=1}^{q} c_i \text{ (for a string consisting of codes 0 and/or 1)},$$

s is a phase shift, the position at which a template pattern can be found in the string c:

$$s: s \in \{0, 1, 2, \cdots, q-1\} \wedge \forall i \in \{1, 2, \cdots, q\} : c_i = \left\lfloor \frac{p}{q}(i - s) \right\rfloor - \left\lfloor \frac{p}{q}(i - s - 1) \right\rfloor,$$

$\lfloor x \rfloor$ is the floor function, indicating the largest integer not larger than x, $\lceil x \rceil$ will denote the ceiling function, indicating the smallest integer not smaller than x.

Domain Theorem (from [5])

The domain of a string c, expressed as a region of the faithful (n, q, p, s) characterization, is the set of all lines $y = \alpha x + e$ satisfying the following two conditions.

1) $p_-/q_- < \alpha < p_+/q_+$.
2) For $p_-/q_- < \alpha \le p/q$,

$$\left\lceil L(s)\frac{p}{q} \right\rceil - \alpha L(s) \le e < 1 + \left\lfloor F(s+l)\frac{p}{q} \right\rfloor - \alpha L(s+l).$$

For $p/q \le \alpha < p_+/q_+$,

$$\left\lceil F(s)\frac{p}{q} \right\rceil - \alpha F(s) \le e < 1 + \left\lfloor L(s+l)\frac{p}{q} \right\rfloor - \alpha L(s+l).$$

The two functions $L(\cdot)$ and $F(\cdot)$ are defined as

$$L(x) = x + \left\lfloor \frac{n-x}{q} \right\rfloor q$$

and

$$F(x) = x - \left\lfloor \frac{x}{q} \right\rfloor q$$

and the integers l, q_+, p_+, q_-, and p_- are defined by

$$l: 0 \le l < q \wedge lp = q - 1 \pmod q$$

$$p_+ = 1 + \left\lfloor L(s+l)\frac{p}{q} \right\rfloor - \left\lceil F(s)\frac{p}{q} \right\rceil$$

$$q_+ = L(s+l) - F(s)$$

$$p_- = -1 - \left\lfloor F(s+l)\frac{p}{q} \right\rfloor + \left\lceil L(s)\frac{p}{q} \right\rceil$$

$$q_- = L(s) - F(s+l).$$

The proofs are given in [5].

The general shape of the domain of the string with tuple (n, q, p, s) is quadrangular (see also [6]). The domain is widest at $\alpha = p/q$, indicating that this slope is the most probable slope in agreement with the string with tuple (n, q, p, s). The domain tapers linearly and reaches a width 0 at $\alpha = p_-/q_-$ and $\alpha = p_+/q_+$. It can be shown that p_-/q_-, p/q, and p_+/q_+ are three consecutive terms in the Farey series of order n [6], [7], implying that $q_- \le n$ and $q_+ \le n$.

With these results, formula (15) becomes

$$\begin{aligned} g_F(n, q, p, s) = & \int_{p_-/q_-}^{p/q} \int_{\lceil L(s)p/q \rceil - \alpha L(s)}^{1 + \lfloor F(s+l)p/q \rfloor - \alpha F(s+l)} \cdot f(e, \alpha)\, P(e, \alpha)\, de\, d\alpha \\ & + \int_{p/q}^{p_+/q_+} \int_{\lceil F(s)p/q \rceil - \alpha F(s)}^{1 + \lfloor L(s+l)p/q \rfloor - \alpha L(s+l)} \cdot f(e, \alpha)\, P(e, \alpha)\, de\, d\alpha, \end{aligned} \tag{17}$$

where $P(e, a)$ is the probability density describing the distribution of the lines.

This formula is the main result of this paper, applied to straight lines, in its most general form. It provides the BLUE estimator for an arbitrary property $f(e, \alpha)$ of a continuous straight line segment, given a particular chain-code string c, faithfully characterized by the tuple (n, q, p, s).

Evaluation of the BLUE Estimators

Estimator (17) will now be calculated for the property "line length per code element," which is $f(e, \alpha) = \sqrt{1 + \alpha^2}$. This requires an assumption to be made about $P(e, \alpha)$. Generally, we can assume a uniform distribution of the lines in distance to an origin, and orientation [7]. This implies

$$P(e, \alpha) \sim (1 + \alpha^2)^{-3/2}. \tag{18}$$

Following [4], we use a moment-generating function f_i:

$$f_i(n, q, p, s) = \iint_{\mathcal{R}(n,q,p,s)} \frac{f^i(e, \alpha)}{(1 + \alpha^2)^{3/2}}\, de\, d\alpha \tag{19}$$

which allows the estimator (17) to be written in the form

$$g_F(n, q, p, s) = \frac{f_1(n, q, p, s)}{f_0(n, q, p, s)} \tag{20}$$

and its variance as

$$\operatorname{var}\{g_F(n, q, p, s)\} = \frac{f_2(n, q, p, s)}{f_0(n, q, p, s)} - \left\{\frac{f_1(n, q, p, s)}{f_0(n, q, p, s)}\right\}^2. \tag{21}$$

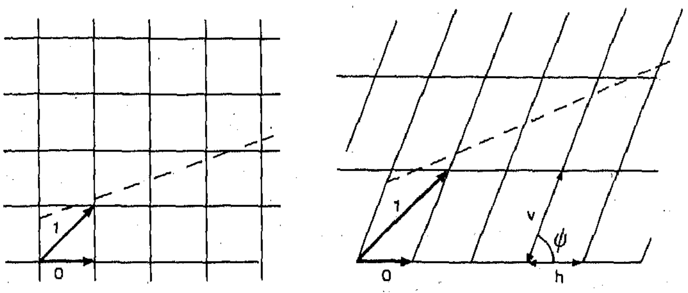

Fig. 4. Generalization to skew grids by a linear transformation.

(As an aside at this point, it should be noted that the estimators of formulas (19)–(21) can easily be generalized to 4- and 6-connected grids and other regular grids, using the concept of a "column" introduced in [4]. Calculating the appropriate linear transformations to transform a skew grid [Fig. 4(a)] to a square grid [Fig. 4(b)] and then applying (19), we find

$$f_i(n, q, p, s) = Kv\left(\frac{h}{v}\sin\psi\right)^2 \cdot \iint_{\mathcal{R}(n,q,p,s)} \cdot \frac{f^i(e_s, \alpha_s)}{\left\{\left(\alpha_s + \frac{h}{v}\cos\psi\right)^2 + \left(\frac{h}{v}\sin\psi\right)^2\right\}^{3/2}} \, de_s \, d\alpha_s. \tag{19'}$$

Here K is a normalization constant which cancels out in the computation of g_F and var (g_F) with formulas (20) and (21). We will not use (19′); it is mentioned here for completeness.)

Since the property line length per code only depends on α, we will from now on only treat properties independent of e: $f(e, \alpha) = f(\alpha)$. In that case, the integral over e in (19) can be performed, yielding

$$f_i(n, q, p, s) = \int_{p/q}^{p_+/q_+} (p_+ - \alpha q_+)\,(1 + \alpha^2)^{-3/2} f^i(\alpha)\, d\alpha + \int_{p_-/q_-}^{p/q} (\alpha q_- - p_-)\,(1 + \alpha^2)^{-3/2} f^i(\alpha)\, d\alpha.$$

This can be rewritten as

$$f_i(n, q, p, s) = [F_i(\alpha | -p_+, -q_+)]_{p/q}^{p_+/q_+} + [F_i(\alpha | p_-, q_-)]_{p_-/q_-}^{p/q} \tag{23}$$

where the functions F_i are defined by

$$\frac{\partial}{\partial \alpha} F_i(\alpha | P, Q) = (\alpha Q - P)\,(1 + \alpha^2)^{-3/2} f^i(\alpha). \tag{24}$$

For the line length per code, we have $f(\alpha) = \sqrt{1 + \alpha^2}$, and hence from (24),

$$F_0 = -\frac{\alpha P + Q}{\sqrt{1 + \alpha^2}} \tag{26}$$

$$F_1 = \frac{Q}{2}\ln(1 + \alpha^2) - P\arctan(\alpha) \tag{27}$$

$$F_2 = Q\sqrt{1 + \alpha^2} - P\ln(\alpha + \sqrt{1 + \alpha^2}). \tag{28}$$

With (20) and (23), this is the BLUE estimator for the line length corresponding to a chain-code string (n, q, p, s).

Two other properties for which we can now easily find the BLUE estimators are the angle $f(\alpha) = \arctan(\alpha)$ and the slope $f(\alpha)$. F_0 is as specified in (26), but F_1 and F_2 now become for the angle $\arctan(\alpha)$:

$$F_1 = \frac{(Q\alpha - P) - (Q + P\alpha)\arctan(\alpha)}{\sqrt{1 + \alpha^2}} \tag{29}$$

$$F_2 = -\frac{(Q + P\alpha)\arctan^2(\alpha) + 2(P - Q\alpha)\arctan(\alpha) - 2(Q + P\alpha)}{\sqrt{1 + \alpha^2}}, \tag{30}$$

and, for the slope α,

$$F_1 = \frac{(P - Q\alpha)}{\sqrt{1 + \alpha^2}} + Q\ln(\alpha + \sqrt{1 + \alpha^2}) \tag{31}$$

$$F_2 = \frac{Q\alpha^2 + P\alpha + 2Q}{\sqrt{1 + \alpha^2}} - P\ln(\alpha + \sqrt{1 + \alpha^2}), \tag{32}$$

respectively.

Approximate Behavior

The formulas (20)–(24) do not reveal the behavior of the estimator with varying $f(\alpha)$ nor do they reveal the dependence of the estimators on (n, q, p, s). Extensive calculations yield Taylor approximations to order $O(n^{-2})$ clarifying these issues:

$$g_F(n, q, p, s) = f\left(\frac{p}{q}\right) + \frac{1}{3q}\left(\frac{1}{q_+} - \frac{1}{q_-}\right) f'\left(\frac{p}{q}\right) + O(n^{-4}) \tag{33}$$

$$\text{var}\,\{g_F(n, q, p, s)\} = \frac{1}{18q^2}\left(\frac{1}{q_+^2} + \frac{1}{q_+ q_-} + \frac{1}{q_-^2}\right) \cdot \left\{f'\left(\frac{p}{q}\right)\right\}^2 + O(n^{-6}). \tag{34}$$

Note that q_+ and q_- are implicitly dependent on (n, q, p, s). It is seen that the first term, dominating g_F, is $f(p/q)$, which is the value of the property f at the "middle" of the domain. As stated before, this is just the most probable slope in agreement with the string with tuple (n, q, p, s). The second term compensates for the asymmetry of the domain relative to $\alpha = p/q$. This term can be shown to be of the order $O(n^{-2})$.

VI. Conclusion

Consideration of the estimation process involved in measuring properties of line segments leads to the discrimination of three steps.

1) *Digitization:* The complete description of this step for straight lines and the loss of information it unavoidably implies were studied in [5]. An optimal estimation procedure aims at using all information remaining after digitization.

2) *Characterization:* This is an essential step in the whole process, which can potentially destroy information. This step accounts, to a large extent, for the differences in accuracy of estimators previously given. An optimal estimator must be based on a faithful characterization, which is loss free.

3) *Calculation:* Based on a specific characterization, many estimators can be given. BLUE estimators are optimal in the sense of being linear and unbiased with minimal MSE. It was shown how they are related to a specific characterization.

Optimal BLUE estimators result from a calculation of BLUE estimators based on a faithful characterization preserving all information left after digitization. This is the general recipe. For some properties of straight lines, this procedure was performed.

For the property "line length per chain code," the optimal BLUE estimators are briefly compared to those of previous authors in Table I with respect to BLUEness and values of the root mean-square error (We hope to give a more detailed comparison, including algorithmic issues, in a future paper. A preview is given in [8].) Table I shows that increasingly accurate characterizations (with more elements in the characterizing tuple) generally result in more accurate estimators. This is partly explained by Fig. 3 where it is seen that the regions corresponding to these characterizations are increasingly smaller, allowing better tuning of the estimator. Thus, the estimators l_F, l_G, and l_P, based on the (n_e, n_o) characterization, all have a high root MSE (≥ 2.6 percent). Use of the (n_e, n_o, n_c) characterization, but still using a linear formula, leads to l_C, which has a limit MSE of 0.6 percent. The BLUE estimator for this characterization is l_V, which is much more accurate, especially for longer strings. The optical BLUE estimator l_O of formulas (23)–(28) is even more accurate, and the proofs of this paper show that beyond l_O, no improvement in accuracy is possible.

Acknowledgment

B. Duin and A. van den Bos are gratefully remembered for offering their willing ears and outspoken throughts, and J. de Bruin, J. Goudriaan, and C. Durville because they stayed friendly and helpful, even after the xth draft of this paper.

References

[1] H. Freeman, "Boundary encoding and processing," in *Picture Processing and Psychopictorics*, B. S. Lipkin and A. Rosenfeld, Eds. New York: Academic, 1970, pp. 241–266.

[2] F. C. A. Groen and P. W. Verbeek, "Freeman-code probabilities of object boundary quantized contours," *Comput. Graphics Image Processing*, vol. 7, pp. 391–402, 1978.

[3] D. Profitt and D. Rosen, "Metrication errors and coding efficiency of chain-encoding schemes for the representation of lines of finite length," *Comput. Graphics Imaging Processing*, vol. 10, pp. 318–322, 1979.

[4] A. M. Vossepoel and A. W. M. Smeulders, "Vector code probability and metrication error in the representation of straight lines of finite length," *Comput. Graphics Image Processing*, vol. 20, pp. 347–364, 1982.

[5] L. Dorst and A. W. M. Smeulders, "Discrete representation of straight lines," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 450–463, 1984.

[6] M. D. McIlroy, "A note on discrete representation of lines," *AT&T Tech. J.*, vol. 64, no. 2, 1985.

[7] L. Dorst and R. P. W. Duin, "Spirograph theory," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, no. 5, pp. 632–639, 1984.

[8] L. Dorst and A. W. M. Smeulders, *Length Estimators Compared, Pattern Recognition in Practice II.* Amsterdam, The Netherlands: North-Holland, 1985, pp. 73–80. Also available in: "Image analysis," in *Proc. 4th Scandinavian Conf. Image Anal.*, vol. 2, Trondheim, Norway, 1985, pp. 743–751.

One-Dimensional Scan Selection for Two-Dimensional Signal Restoration

JORGE L. ARAVENA AND WILLIAM A. PORTER

***Abstract*—The problem of *m*-D filtering using sequential scanning is considered. It is shown that the optimal causal filter and the performance measure depend on the scan selected. Examples show that this effect can be significant. Possible techniques to select a suitable scan are analyzed.**

***Index Terms*—Image enhancing, *m*-D filtering, partially ordered resolution spaces, scan selection, selective memory.**

I. Introduction

A simple method of applying 1-D techniques in the design of discrete m-D filters is raster scanning whereby the m-D domain set is linearly ordered. Raster scanning has been used both in the frequency domain [7] and in the time domain. In particular, it has enabled the design of recursive, Kalman type filters [4], [6], [11], [12] for m-D signal extraction.

Manuscript received February 19, 1985; revised August 15, 1985. Recommended for acceptance by Y. T. Chien. This work was supported in part by the National Science Foundation under Grant ECS-83/16731.

The authors are with the Department of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA 70803.

IEEE Log Number 8405792.

Most authors have assumed that the scanning is *a priori* fixed. However, in studying related problems [1] we have shown that the choice of scan can affect the performance of the subsequent filter. Scan selection can also be important in the sequential processing of bidimensional sensor arrays and other multiplexed data aquisition systems.

In this correspondence we consider scan selection as a special case of selective memory (formally defined later). We determine an optimal mean square filter with respect to a general selective memory constraint. Using simple computational examples we show that the error may vary significantly with the specified scan.

The total number of possible scans in an $N \times N$ raster is $(N^2)!$. In only the most trivial cases is an exhaustive search for the best scan practical. We present a search algorithm, which in simulation tests, produces a scan procedure close to the optimal. A search technique derived from dynamic programming is also considered.

II. Mathematical Preliminaries

We use the basic ideas of causality theory [5], [10] to develop our concept of selective memory. For sake of brevity we consider 2-D signals only, and make use of the power of functional analysis concepts [8], [9]. The index or time set will be taken as the set $\mu = \{a = (i, j) : 0 \leq i, j \leq N - 1\}$. We shall model the image space by $X = l_2^2(\mu)$ and use the orthoprojector family $\{\Delta(a); a \in \mu\}$ defined by

$$[\Delta(a)x](b) = \begin{cases} 0 & b \neq a \\ x(a) & b = a. \end{cases}$$

The vector $\Delta(a)x$ is often called the value of the image x at point $a \in \mu$.

For a given subset $\sigma \subset \mu$ define the related projector

$$P^\sigma = \sum_{a \in \sigma} \Delta(a);$$

and consider the equality

$$\Delta(a)Lx = \Delta(a)LP^\sigma x, \qquad \text{all } x.$$

Since $P^\sigma x$ is always zero for all points outside the subset σ, the value of the output at $a \in \mu$ does not depend on the input outside the subset σ. We say that the projector P^σ identifies the memory of the system L at $a \in \mu$.

For each $a \in \mu$ let there be specified a subset $\sigma(a) \subset \mu$ and a projector $P^a = P^{\sigma(a)}$. A given processor L has the specified selective memory if and only if L satisfies

$$\Delta(a)L = \Delta(a)LP^a, \qquad a \in \mu. \tag{1}$$

The projectors $\{P^a\}$ need not have any relationship among themselves nor with the projectors $\{\Delta(a)\}$. However, for every scanning of the elements in the index we can choose the P^a so that (1) implies causality.

We shall use the name "l-causal" to denote a map with selective memory specified by a linear order "l." If the input is scanned one pixel at a time according to the order l then l-causal maps are realizable by on line sequential processing.

III. The Restricted Optimal Filter

We will determine the filter which is mean square optimal with respect to the constraint of a specified selective memory. The standard such problem is shown in Fig. 1. The operator L describes a known blurring or degradation effect on the image x. The signal η is an additive noise and D is the reconstructive or enhancing filter.

The image processor D has a selective memory specified by the projector family $\{P^a\}$. The restricted optimal filtering problem is that of determining a filter D to minimize the index $J(D) = E[\|x - \hat{x}\|^2]$, over the class of systems with the same selective memory, hereafter called realizable systems. Here $E[\cdot]$ denotes statistical expectation.