N 69 36261

NASA CR 105647

Entropy of Printed Matter at the Threshold of Legibility for Efficient Coding in Digital Image-Processing

by

Ronald Barthold Arps

June 1969

Scientific Report No. 31

Prepared under

National Aeronautics and Space Administration

Grant NGL-05-020-014

RADIOSCIENCE LABORATORY

STANFORD ELECTRONICS LABORATORIES

STANFORD UNIVERSITY · STANFORD, CALIFORNIA

ENTROPY OF PRINTED MATTER AT THE THRESHOLD OF LEGIBILITY FOR EFFICIENT CODING IN DIGITAL IMAGE-PROCESSING

By

Ronald Barthold Arps

June 1969

Scientific Report No. 31

Prepared under

National Aeronautics and Space Administration
Grant NGL-05-020-014

Radioscience Laboratory
Stanford Electronics Laboratories
Stanford University Stanford, California

ABSTRACT

Printed matter (text or typewriting) as a two-dimensional, two-valued, stochastic process, is studied to determine how its statistical properties vary with resolution. Spatial resolution is digital, with the document image dissected into small rectangular scan "elements." Quantizing is two-level with a scan element defined as either black or white, using a 50 percent decision level after integration over the element area.

The tradeoff between document quality and document entropy is explored by taking each factor separately and modeling its dependence on spatial frequency. Since these factors really depend on the ratio between resolution and character size, all results are reported in terms of normalized spatial frequency. Stroke width is used to express character size in this normalization.

Legibility of characters, out-of-context, is used as the document quality measure for printed matter. A piece-wise linear model for the dependence of legibility on resolution is postulated, based on concepts from the sampling theorem. The recovery of character-strokes when under-sampled is assumed proportional to the intelligible recovery of characters (legibility). An empirical model is also fitted to the data, as it varies over combinations of two-dimensional resolution.

A succession of increasingly complex source alphabets are examined for encoding the document image. One- and two-dimensional alphabets are tried out before exploring resolution variation for the better ones. Horizontal and vertical resolution are kept equal for these experiments. An empirical model is derived to express the behavior of document entropy with resolution for the various alphabets. Resolution efficiency is defined in order to compare the compression achieved at a given resolution with compression at the Nyquist interval for character-strokes. This measure is proposed in order to evaluate the high compression values obtained in using sampling rates much higher than the Nyquist rate.

CONTENTS

CONTENTS (Cont)

TABLES

ILLUSTRATIONS

ILLUSTRATIONS (Cont)

ACKNOWLEDGMENT

I would like to express my indebtedness to Professor Allen M. Peterson for his unfailing support and encouragement, to Professor James C. Bliss for his guidance and suggestions, and to Professor Joseph W. Goodman for his help in reviewing this manuscript.

Grateful acknowledgment is made to the International Business Machines Corporation for material and financial support. In addition, I wish to express my appreciation to my colleagues at IBM: Carl E. Schlaepfer, under whose guidance the experimental system was constructed; and to Robert L. Erdman and Allan S. Neal, whose human factors analysis made possible the legibility data.

This dissertation was published under the support of NASA Contract No. NSG 377.

Chapter I

INTRODUCTION

1.1. Background

In the two decades since Shannon's fundamental contribution [Shannon - 1948], considerable effort has been expended in applying information theory to a variety of problem areas. The reduction of redundancy in digitized images is but one such application. More recently, industry has been grappling with the practical implementation of these ideas giving rise to new questions and problem areas.

For example, office copiers have developed a vast market in the business world; the use of machines for handling business documents has become commonplace. A natural extension is the storage and forwarding of document images. To be sure, facsimile has been around for a long time. But only recently has a large market and new technology combined to make sophisticated image processing systems feasible. Applying information theory to the processing of digital images provides the capability to encode for compression or noise immunity. Regeneration of the image is possible in the face of many processing steps that might otherwise cause image degradation.

The design of such systems requires more than the theoretical concepts of entropy, redundancy reduction, and compression codes. Statistics characterizing typical documents are required to design matching compression codes. In this case, "representative" documents must be defined. For such representative documents the statistics are dependent on the resolution of the digitizing process. Resolution also impacts the "quality" of a document image. This dependence on resolution, and the resulting tradeoffs for document coding and document quality, are the major topics of this study.

Typically, empirical results in the literature are hard to extend or generalize quantitatively. One must subjectively judge quality using the arbitrary image of a face, crowd, etc., for comparison before and after some processing step. Compressions using a code are reported for a specific resolution, without feel for its dependency on the resolution chosen. Extrapolation of these results to other documents or resolutions can be made only qualitatively.

What is presented here is a quantitative exploration of document quality and compression. Concentration on print is motivated by the preponderance of it in the images of business documents. Also, the emphasis on printed matter and the nature of processing in contemporary copying machines, has caused me to restrict my attention to two-level image processing (as contrasted with "grey-scale").

1.2. Summary

Two measures for the information inherent in printed documents --legibility and entropy--have been explored here, in an attempt to characterize their behavior with resolution. Though philosophically similar their characteristics seemed not to overlap, with mechanical measurements of entropy tending only to approximate crudely the information discerned by human observers.

In Chapter II, the legibility of alphanumeric characters is used as a measure for document quality. The physical recovery of character structure is proposed as a major factor determining the behavior of legibility with decreasing resolution. A piece-wise linear model is presented for the recovery of characters, based on undersampling of the strokes that constitute their structure (Fig. 2.1-4). Legibility measurements for the same documents are fitted with an analytic relationship, expressed in terms of spatial frequency over both dimensions (Eq. 2.2-18). This empirical model has been normalized with respect to character size and appears to have the same shape for differing type fonts.

Entropy is used in Chapter III to measure the performance of successively complex coding schemes. Resolution is varied here over the same spatial frequencies used in Chapter II. Both one- and two-dimensional codes are explored, in an attempt to understand the mechanisms underlying compression for the images of printed matter.

From the insight gained, a model for document entropy is constructed based on page size, character size and density, and resolution

(Eq. 3.4-20). For a particular type font this model fitted to data appears to be independent of the actual distribution of characters in the sample.

A definition is proposed for resolution efficiency, relative to the Nyquist rate for the character structure (Eq. 3.4-23). This is called ϵ^*-efficiency and embodies all the resolution variation. It has been used to define simple expressions for ϵ^*-symbol entropy, ϵ^*-symbol compression, and ϵ^*-page entropy (Eq. 3.4-25, 26, and 27). Also, the ability of a scan pattern to handle varying character sizes is specified by ϵ^*-efficiency.

Possible extensions to this work are noted in Chapter IV.

Chapter II

THE THRESHOLD OF LEGIBILITY WITH SPATIAL FREQUENCY

A useful measure for image quality in printed matter is how well it can be read. Business documents add a constraint to this, in that context may not always be helpful in deciphering an unclear image. Financial statements or part number listings, for example, must preserve their character integrity throughout any image processing. Therefore, out-of-context character "legibility" is used here for print quality specification.

Historically, legibility has been used as a measure of quality to investigate the effects of type font, stroke width, and other parameters of printed characters. More recently legibility has been used to measure quality in various kinds of visual displays. Many of the basic investigations of print are summarized in Legibility of Print [Tinker - 1963]. The extent of current literature on display quality is indicated in recent bibliographies [Cornog and Rose - 1967], and [Shurtleff - 1967].

My interest here lies in the effects of resolution on the digitized images of printed matter. Attention has been restricted to rectangular resolution elements variable in both dimensions. Parameters dependent on resolution are specified over a "resolution plane" where vertical and horizontal resolution are the axes.

2.1. Theoretical Model Based on the Sampling Theorem

An example of throughput for the experimental system (Appendix A.1) can be seen in Fig. 2.1-1, where a "before" and "after" microphotograph of the letter K is shown. The original type font, which is called Mid-Century, was selected for its clean strokes which are not adorned with serifs. The particular character illustrated came from the word THINK in document A1 (Appendix A.2). The digitizing process is readily evident in the reproduced character, which has been built up from small black or white rectangles.

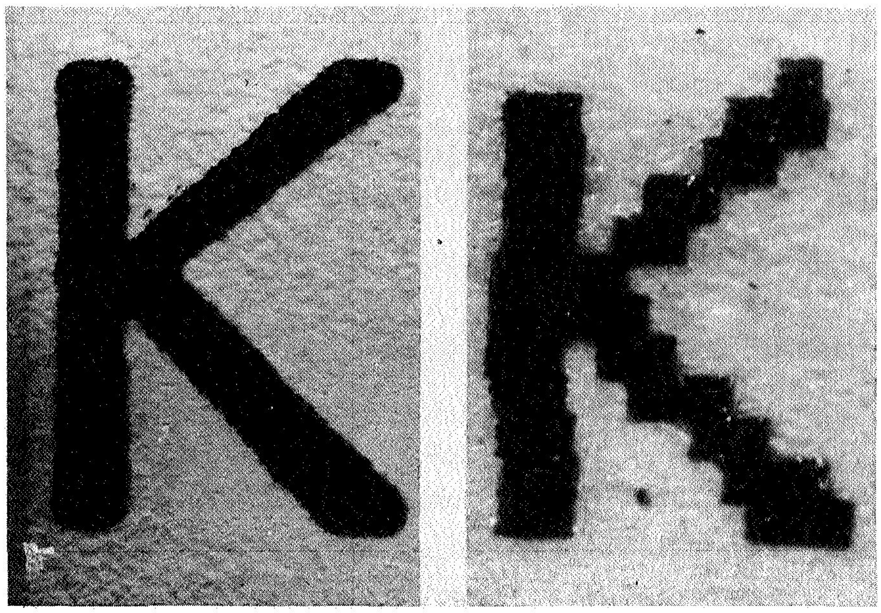


Fig. 2.1-1. CHARACTER BEFORE AND AFTER DIGITAL PROCESSING (20X MAGNIFICATION).

The scanning process is illustrated in Fig. 2.1-2, where an idealized K is displayed superimposed with a rectangular scan element. As the area-integrating element moves in the direction of scan, it is periodically sampled and classified as either black or white. The following dimensions are also shown in Fig. 2.1-2:

w_s	= stroke width	mils
l_s	= stroke length (for a straight stroke)	mils
d_s	= scan element dimension in the scanning direction	mils
d_p	= scan element dimension perpendicular to the scanning direction	mils

Angles are all defined relative to the abscissa in a rectangular x,y coordinate system:

θ_s	= stroke angle	degrees
θ_e	= element scan direction	degrees

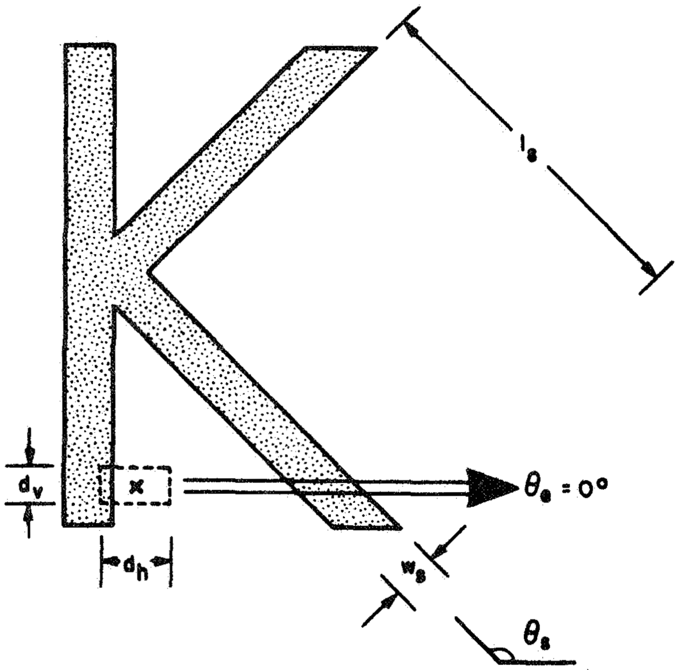


Fig. 2.1-2. SCANNING AN IDEALIZED CHARACTER.

The following functions [Goodman - 1968] will also be useful in describing the scanning process.

Comb function:

$$\text{comb}\ (x) = \sum_{-\infty}^{\infty} \delta(x-n),$$ where $\delta(x)$ is the Dirac delta function

Rectangle function:

$$\text{rect}\ (x) = \begin{cases} 1, & |x| < \frac{1}{2} \\ 0, & |x| > \frac{1}{2} \end{cases}$$

Sinc function:

$$\text{sinc}\ (x) = \frac{\sin\ (\pi x)}{\pi x}$$

2.1.1. Transfer function of the digital image-processing system

For a general model of document digitizing, define:

$g_d(x,y)$ = document reflectivity function

$a(x,y)$ = aperture function of the scan element

Scanning results in a two dimensional-convolution between the two:

$$g_e(x,y) = \text{scan element output}$$

$$= g_d(x,y) * a(x,y) \qquad (2.1\text{-}1)$$

For digital image processing, the scanner output must be sampled:

$$g_s(x,y) = \text{sampler output}$$

$$= g_e(x,y) \ \mathrm{comb}\left(\frac{x}{X}\right) \ \mathrm{comb}\left(\frac{y}{Y}\right) \qquad (2.1\text{-}2)$$

where the x and y sampling intervals are X and Y respectively. Finally, analog-to-digital (A/D) conversion introduces a nonlinear operation that I symbolize with a subscripted C-bracket:

$$g_c(x,y) = \text{A/D converter output}$$

$$= C\{g_s(x,y)\}_{\ell} \qquad (2.1\text{-}3)$$

where, ℓ, is the number of grey levels used in the conversion.

Signal restoration can be represented in the spatial frequency domain by multiplication with a filter characteristic:

$$H(f_X, f_Y) = \text{filter spatial frequency response}$$

or one can equivalently use convolution in the space domain:

$$h(x,y) = \text{filter spatial response}$$

The restored document image then becomes:

$$\hat{g}_d(x,y) = \text{D/A converter output}$$

$$= g_c(x,y) * h(x,y) \qquad (2.1\text{-}4)$$

Combining Eq. 2.1-1,2,3, and 4 produces the glorious result:

$$\hat{g}_d(x,y) = C\left\{[g_d(x,y)*a(x,y)]\left[\mathrm{comb}\left(\frac{x}{X}\right)\mathrm{comb}\left(\frac{y}{Y}\right)\right]\right\}_{\ell} * h(x,y) \qquad (2.1\text{-}5)$$

Case A. rect filtering for analog sampling

Classically, in the Whittaker-Shannon sampling theorem [Whittaker - 1915], [Shannon - 1949] one has:

$$G_d(f_X, f_Y) = 0, \quad \text{for} \quad f_X \geq \frac{1}{2X}, \; f_Y \geq \frac{1}{2Y} \qquad \text{(band-limited)}$$

$$a(x,y) = \delta(x,y)$$

$$H(f_X, f_Y) = \mathrm{rect}(Xf_X)\ \mathrm{rect}\ (Yf_Y)$$

$$h(x,y) = \frac{1}{XY}\ \mathrm{sinc}\left(\frac{x}{X}\right)\ \mathrm{sinc}\left(\frac{y}{Y}\right)$$

$$\ell \to \infty \qquad \text{(analog)}$$

where the two-dimensional Dirac delta function is defined as [Goodman - 1968]:

$$\delta(x,y) = \lim_{N\to\infty} N^2 \exp[-N^2\pi(x^2+y^2)]$$

and the use of generalized Fourier transforms is assumed throughout. (For a detailed discussion of delta functions, refer also to [Bracewell - 1965].)

Using the sifting property of delta functions in the convolution of Eq. 2.1-1 yields:

$$g_e(x,y) = g_d(x,y)*\delta(x,y) = g_d(x,y) \qquad (2.1\text{-}1a)$$

The A/D conversion of Eq. 2.1-3 does not apply:

$$g_c(x,y) = g_s(x,y) \qquad (2.1\text{-}3a)$$

so that combination of Eq. 2.1-1a,2, and 3a yields:

$$g_c(x,y) = g_d(x,y)\ \mathrm{comb}\left(\frac{x}{X}\right)\ \mathrm{comb}\left(\frac{y}{Y}\right) \qquad (2.1\text{-}6)$$

In the spatial frequency domain, the following identity holds for Eq. 2.1-4 with the given filter:

$$\hat{G}_d(f_X,f_Y) = G_c(f_X,f_Y)\ \mathrm{rect}(Xf_X)\ \mathrm{rect}(Yf_Y) = G_d(f_X,f_Y) \qquad (2.1\text{-}4a)$$

assuming that $G_d(f_X,f_Y)$ is band-limited:

$$G_d(f_X,f_Y) = 0\ , \quad \text{for} \quad f_X \geq \frac{1}{2X},\ f_Y \geq \frac{1}{2Y}$$

Since the expanded form of the transform in Eq. 2.1-6 is:

$$G_c(f_X,f_Y) = \sum_{n=-\infty}^{\infty}\ \sum_{m=-\infty}^{\infty} G_d\left(f_X - \frac{n}{X},\ f_Y - \frac{m}{Y}\right) \qquad (2.1\text{-}6a)$$

the band-limiting assumed for $G_d(f_X,f_Y)$ just matches the assumed rect filter, and only the replication of $G_d(f_X,f_Y)$ passes through at the origin:

$$\hat{G}_d(f_X,f_Y) = G_d(f_X,f_Y) \qquad (2.1\text{-}5a)$$

Case B. sinc filtering for digital sampling

In the system used here (Appendix A.1), samples are reconstructed using a two-level output printer with stair-step interpolation from sample to sample. In effect, a two-dimensional rect function:

$$h(x,y) = \mathrm{rect}\left(\frac{x}{X}\right)\ \mathrm{rect}\left(\frac{y}{Y}\right)$$

is convolved with two-level samples in the space domain:

$$\ell = 2$$

$$g_c(x,y) = C\left\{g_s(x,y)\right\}_2 \qquad (2.1\text{-}3b)$$

$$\hat{g}_d(x,y) = g_c(x,y) * h(x,y) \qquad (2.1\text{-}4b)$$

This is equivalent to filtering in the spatial frequency domain with a two-dimensional sinc function:

$$H(x,y) = XY\ \mathrm{sinc}(Xf_X)\ \mathrm{sinc}(Yf_Y)$$

The scan aperture is an identical rectangle to the above:

$$a(x,y) = \mathrm{rect}\left(\frac{x}{X}\right)\ \mathrm{rect}\left(\frac{y}{Y}\right)$$

$$g_e(x,y) = g_d(x,y)*\mathrm{rect}\left(\frac{x}{X}\right)\ \mathrm{rect}\left(\frac{y}{Y}\right) \qquad (2.1\text{-}1b)$$

both with dimensions matching the sampling intervals:

$$g_s(x,y) = g_e(x,y)\ \mathrm{comb}\left(\frac{x}{X}\right)\ \mathrm{comb}\left(\frac{y}{Y}\right) \qquad (2.1\text{-}2b)$$

The resulting system description becomes:

$$\hat{g}_d(x,y) = C\left\{\left[g_d(x,y)*\mathrm{rect}\left(\frac{x}{X}\right)\mathrm{rect}\left(\frac{y}{Y}\right)\right]\left[\mathrm{comb}\left(\frac{x}{X}\right)\mathrm{comb}\left(\frac{y}{Y}\right)\right]\right\}_2 * \left[\mathrm{rect}\left(\frac{x}{X}\right)\mathrm{rect}\left(\frac{y}{Y}\right)\right] \qquad (2.1\text{-}5b)$$

In effect, the image is dissected by a rectangular X by Y grid (Eq. 2.1-1b and 2b). The elements of the grid are then deemed black or white, depending on the "color" predominating (Eq. 2.1-3b). Finally, they are reproduced as solid-color rectangles of the same dimension (Eq. 2.1-4b). [Deutsch - 1957] used this criterion for two-level quantization.

Notice that using space domain analysis aids in understanding the physical process. In the spatial frequency domain, it is especially awkward to visualize the effect of two-level quantization.

The process obviously does not achieve perfect restoration. However, area-integration does serve to eliminate "fly-specks"--document noise of less than 1/2 the sample element area. The two-level nature of the data is also preserved. This aids in regeneration of the document during multiple passes through such an image-processing system.

2.1.2. The probability of stroke recovery vs resolution

Linear systems techniques can be applied to simplify analysis of the generalized document. First, line drawings and characters can be decomposed into strokes (as defined previously in Fig. 2.1-2). The "impulse response" to an idealized stroke can be measured next for an image-processing system, and then superposition used to extend the results to specific cases. (This is not truly linear because amplitude must be forced to one of two levels after superposition.)

A simplifying approximation can be made by assuming that:

$$\ell_s \gg d_s, d_p \qquad \text{mils} \qquad (2.1\text{-}7)$$

That is, in practical applications the scan elements are smaller than the stroke length, and about the size of the stroke width.

A further simplification can be made by considering the critical angle between the stroke and direction of scan. When

$$|\theta_s - \theta_e| = 90° \qquad \text{mils} \qquad (2.1\text{-}8)$$

the stroke appears narrowest to the sweeping element. This is important in the case of undersampling, where:

$$w_s < d_s, d_p \qquad \text{mils}$$

Strokes are not reproduced if they are too narrow to fill 50 percent of the scan element area.

Finally, for convenience I have assumed a horizontal scan so that:

$$\theta_e = 0° \qquad (2.1\text{-}9)$$

$d_h = d_s$ = scan element dimension in the scanning direction mils

$d_s = d_p$ = scan element dimension perpendicular to the scanning direction mils

Figure 2.1-3a illustrates the assumptions of Eq. 2.1-7, 8, and 9. The document reflectivity is now represented by:

$$g_d(x,y) = \mathrm{rect}\left(\frac{x - Z}{w_s}\right) ,$$

where, Z, is a random variable with uniform distribution:

$$P[Z] = \frac{1}{X_d},\ |x| \leq \left(\frac{X_d}{2}\right)$$

$$= 0\ ,\ |x| > \left(\frac{X_d}{2}\right)$$

representing the relative phase difference between the center line of the stroke (X_d, y) and the origin (0,0). This phase difference determines whether a stroke will be sampled by the combs of Eq. 2.1-2b centered at the origin.

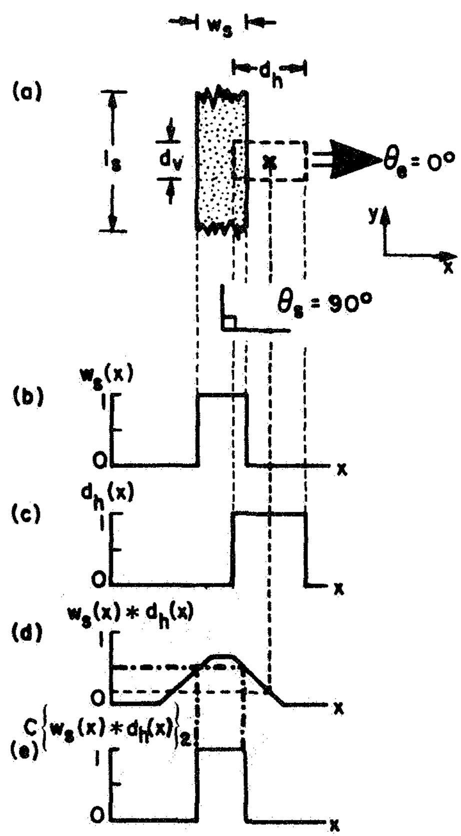


Fig. 2.1-3. SCANNING AN IDEALIZED STROKE.

For simplicity, the problem can now be viewed in one dimension:

$$g_d(x,y) = w_s(x) = \text{rect}\left(\frac{x - Z}{w_s}\right) \qquad (2.2\text{-}10)$$

$$a(x,y) = d_h(x) = \text{rect}\left(\frac{x}{X}\right) \qquad (2.1\text{-}12)$$

and Eq. 2.1-5b becomes:

$$\hat{w}_s(x) = C\left\{\left[\text{rect}\left(\frac{x - Z}{w_s}\right) * \text{rect}\left(\frac{x}{X}\right)\right] \text{comb}\left(\frac{x}{X}\right)\right\}_2 * \text{rect}\left(\frac{x}{X}\right) \qquad (2.1\text{-}12)$$

The convolution between $w_s(x)$ and $d_h(x)$ is illustrated in Fig. 2.1-3b, c, and d. In Fig. 2.1-3a, $\frac{1}{6}$th of the scan element overlaps the stroke. This value is plotted directly below, as a point of height $\frac{1}{6}$ on the curve of Fig. 2.1-3d.

The A/D conversion process is illustrated in Fig. 2.1-3d and e, with the values 0.5- and-above in Fig. 2.1-3d being assigned value "1" in Fig. 2.1-3e. Notice that the output of the A/D converter is a faithful version of $w_s(x)$. This is true so long as $w_s > (d_h)/2$. If $w_s \leq (d_h)/2$, the convolution will not exceed 0.5 and the stroke will drop out.

In Fig. 2.1-3 sampling has not yet occurred. In this case, the order of sampling and quantizing is not important. It has been shown second in order to clearly demonstrate the effects of the random phase, Z. Since Z is uniformly distributed, a stroke will be sampled with probability:

$$\begin{aligned} P_{rs}(x) = P[\text{recovery of stroke}] &= 0\,, \quad \text{for} \quad w_s \leq \frac{d_h}{2} \\ &= \frac{w_s}{d_h}, \quad \text{for} \quad \frac{d_n}{2} < w_s \leq d_h \\ &= 1\,, \quad \text{for} \quad d_h < w_s \qquad (2.1\text{-}13) \end{aligned}$$

The probability of recovery is zero for the case of drop-out, one for strokes wider than the sampling interval, and linear in-between.

Generalizing to two dimensions, a simple piece-wise linear model is made assuming that the smaller element dimension predominates in modeling the chance of recovery. Equation 2.1-13 now becomes:

$$
\begin{aligned}
P_{rs}(x,y) &= 0.0, \quad \text{for} \quad \frac{w_s}{d_h} \ \text{or} \ \frac{w_s}{d_v} \leq 0.5 \\
&= \frac{w_s}{d_h}, \quad \text{for} \quad 0.5 < \frac{w_s}{d_h} \leq \frac{w_s}{d_v} \leq 1.0 \\
&= \frac{w_s}{d_v}, \quad \text{for} \quad 0.5 < \frac{w_s}{d_v} \leq \frac{w_s}{d_n} \leq 1.0 \\
&= 1.0, \quad \text{for} \quad 1.0 < \frac{w_s}{d_h} \ \text{and} \ \frac{w_s}{d_v} \qquad (2.1\text{-}14)
\end{aligned}
$$

In order to express the piece-wise linear ranges simultaneously for both dimensions, the inequalities have been rearranged. This gives stroke width over element size, which is dimensionless and constitutes a normalization. In other words, if:

$$\frac{1}{d_h} = f_h = \text{horizontal spatial frequency} \quad \text{elem./inch}$$

$$\frac{1}{d_v} = f_v = \text{vertical spatial frequency} \quad \text{elem./inch}$$

then multiplying next by w_s (inch/stroke) yields normalized spatial frequencies:

$$f_h^* = w_s f_h \qquad \text{elem./stroke}$$

$$f_v^* = w_s f_v \qquad \text{elem./stroke}$$

Using these normalized coordinates, the piecewise-linear model of Eq. 2.1-14 has been plotted in Fig. 2.1-4. The dotted vertical

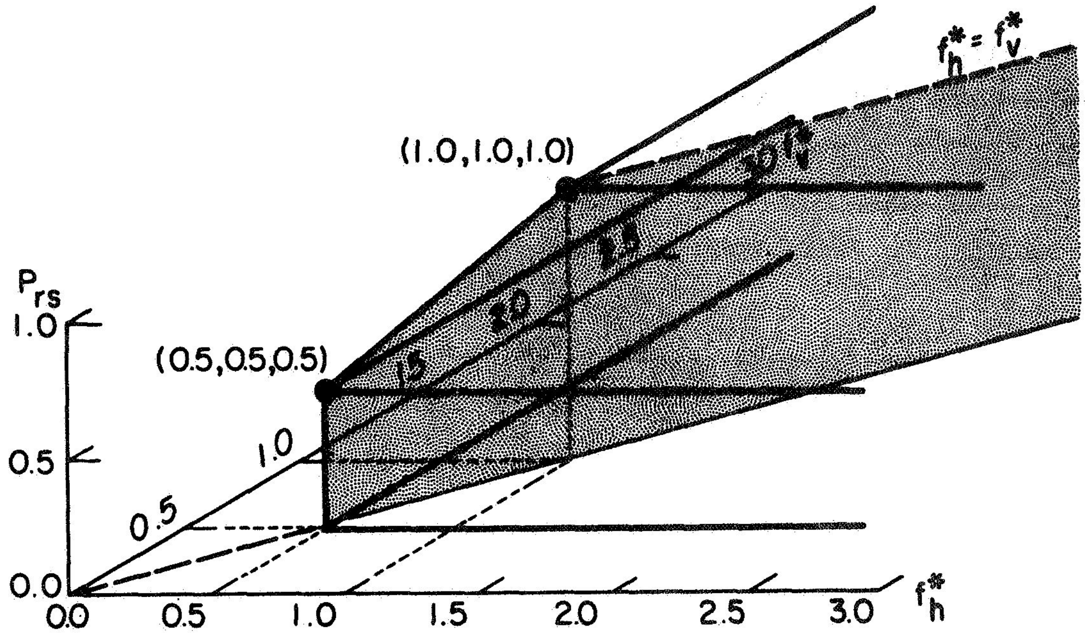


Fig. 2.1-4. THEORETICAL MODEL FOR THE THRESHOLD OF LEGIBILITY.

plane, where $f_h^* = f_v^*$, represents combinations of equal element dimension (i.e., square scan elements).

Figure 2.1-4 is proposed as a theoretical model for legibility using the following superposition argument:

i) Percent legibility of a character set is defined to be the average percent legibility of its individual characters.

ii) The percent legibility of individual characters is assumed to be proportional to the percent recovery of their strokes.

The assumption ii) is admittedly rough, but provides a first order model from which to make further refinements as insight is gained.

2.2. Empirical Model Based on Legibility Measurements

Document A.1 (see Appendix 2) with four sizes of Mid-Century type was passed through the experimental system at all 25 resolution combinations for both vertical and horizontal scanning. The legibility for these outputs was measured by associates at IBM, and the total experiment reported internally in a joint paper [Arps, et al - 1966]. Subsequently these results were presented externally at the 1968 IEEE Conference on Communications [Arps, et al - 1968]. Raw data from this joint effort are reported in Appendix A.3. The subsequent contributions in sections 2.2.1 - 2.2.3 are my own and heretofore not published.

2.2.1. Normalizing with respect to stroke width

To combine the data taken for different character sizes, normalization was desirable. A check of the data indicated that the results of scaling were fairly linear. For example, scan aperture dimension had to double when character height was doubled to achieve the same legibility (Fig. 2.2-1). The particular dimension to be used for normalization (character height, stroke width, etc.) appeared unimportant at first, since the different character sizes were photographic reductions of each other. Stroke width, w_s, was selected in order to relate results to the theory in sections 2.1.1 and 2.1.2.

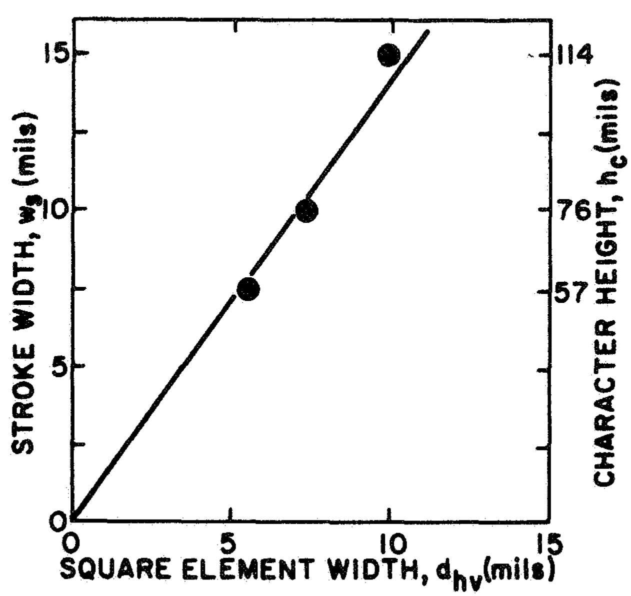


Fig. 2.2-1. SCALING CHECK FOR LINEARITY.

The resulting normalized resolution frequency combinations are illustrated in Fig. 2.2-2.

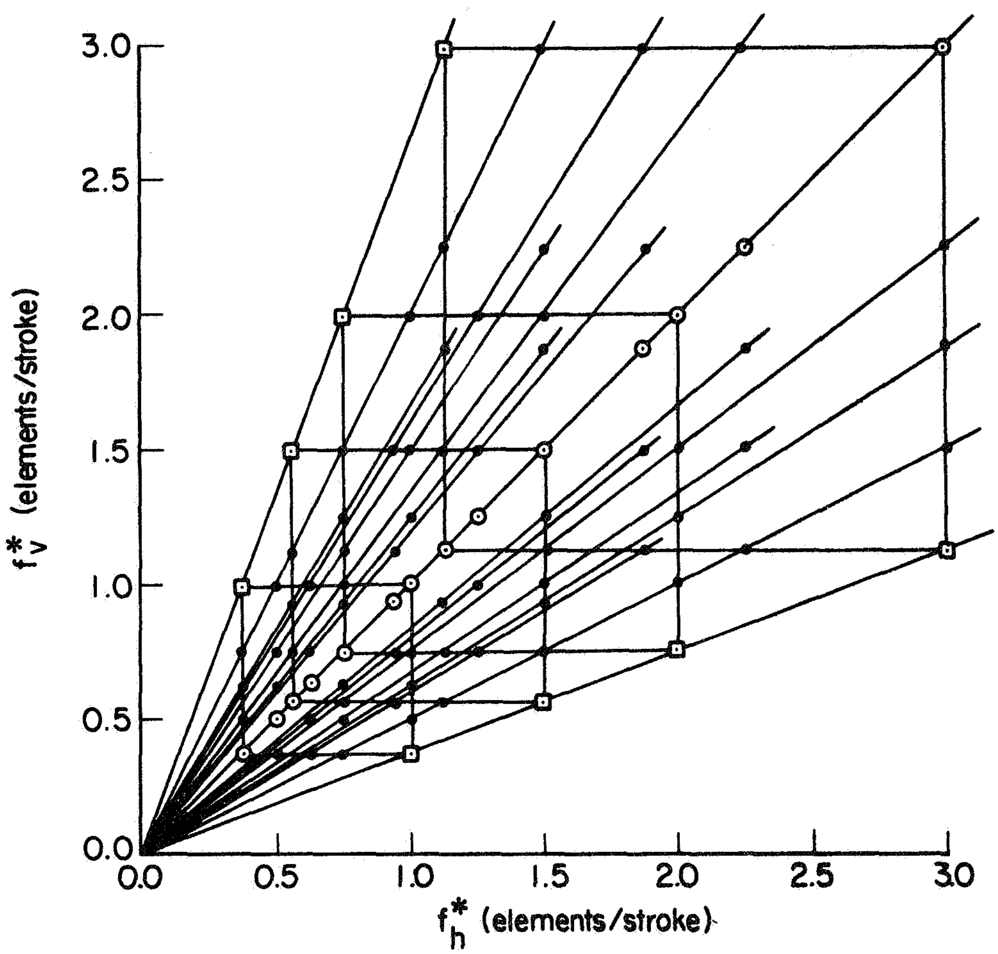


Fig. 2.2-2. NORMALIZED RESOLUTION COMBINATIONS, WITH $a_{hv} = 0.0$.

They represent all the combinations of:

f_h^* = normalized horizontal spatial frequency

$$= \frac{f_h}{f_s} = \frac{w_s}{d_h} \qquad \text{elem./stroke}$$

and:

f_v^* = normalized vertical spatial frequency

$$= \frac{f_v}{f_s} = \frac{w_s}{d_v} \qquad \text{elem./stroke}$$

for all combinations of d_h, d_v, and w_s with values:

$$d_h = 5.0,\ 6.7,\ 8.0,\ 10.0,\ 13.3 \qquad \text{mils/elem.}$$

$$d_v = 5.0,\ 6.7,\ 8.0,\ 10{,}0,\ 13.3 \qquad \text{mils/elem.}$$

$$w_s = 5.0,\ 7.5,\ 10.0,\ 15.0 \qquad \text{mils/stroke}$$

The 200 combinations are grouped into 4 large overlapping squares. These represent the same 25 resolution settings normalized varying amounts corresponding to the 4 character sizes on Document A1. Rays have been drawn from the origin to indicate data points with the same f_v^* to f_h^* ratio. The circled data points along the 45° ray, represent the resolution combinations for square scan elements. The square-marked data points have the largest f_v^* to f_h^* ratio (marked for use in later discussion along with the specification that $a_{hv} = 0$).

A rough feel for the data in Table A.3-1 was obtained by jotting values down at corresponding resolutions in Fig. 2.2-2. When contours of equal legibility were sketched in, they appeared to have a hyperbolic structure.

$$(f_h^* - a_v)(f_v^* - a_h) = K^*, \quad \text{for} \quad P_{rc}[K^*] = K_L \tag{2.2-1}$$

Its asymptotes would be the lines:

$$f_h^* = a_v$$

and:

$$f_v^* = a_h$$

and the origin for the hyperbolas would lie at (a_h, a_v). To preserve symmetry I added the constraint that:

$$a_h = a_v = a_{hv} \tag{2.2-2}$$

Another plot was made using only legibility values for square scan elements (Fig. 2.2-3). Values were taken along a 45° line in the resolution plane for both Table A.3-1 and the theoretical model

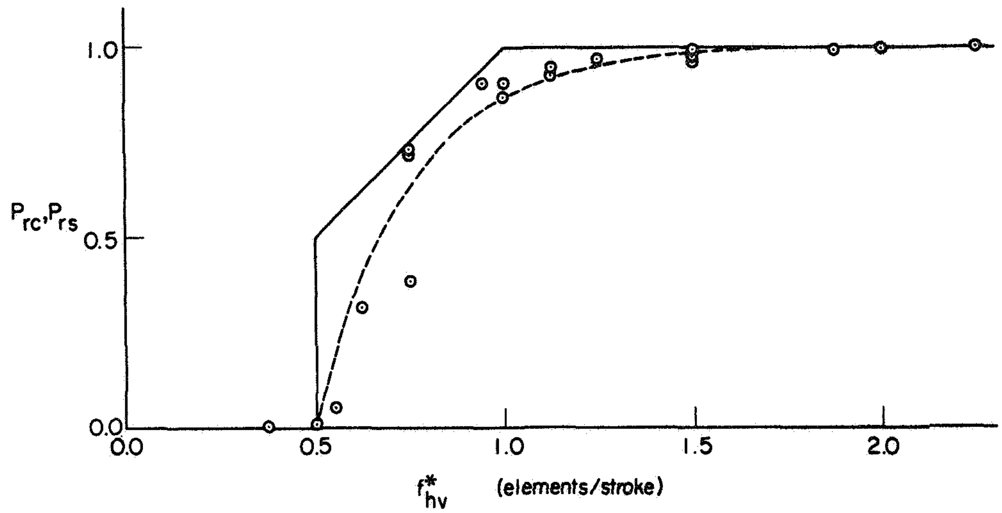


Fig. 2.2-3. LEGIBILITY DATA FOR SQUARE SCAN ELEMENTS (UPPERCASE, MID-CENTURY TYPE).

of Fig. 2.1-4. The theoretical model became a piece-wise linear curve appearing to bound the empirical data points. A sketch through the empirical points resembled a classical $1-e^{-x}$ curve. This resemblance was stronger when allowance was made for the greater standard error of legibilities around 50 percent.

The exponential structure suggested by Fig. 2.2-3 was further evaluated by plotting it again on semi-log paper (Fig. 2.2-4). The result was linear enough to justify a representation of the form:

$$\ln(1 - P_{rc}) = -\frac{1}{c}(f^*_{hv} - b), \quad \text{for} \quad f^*_{hv} \geq b \tag{2.2-3}$$

or alternatively:

$$P_{rc} = 1 - \exp\left[-\frac{1}{c}(f^*_{hv} - b)\right], \quad \text{for} \quad f^*_{hv} \geq b \tag{2.2-4}$$

where

b = the intercept at zero legibility

c = the exponential decay constant

This model for legibility is a physically reasonable statistic in that it is asymptotic (in contrast to the piece-wise linear model). As resolution frequency increases, the illegibility $(1 - P_{rc})$, decreases in direct proportion to the remaining illegibility.

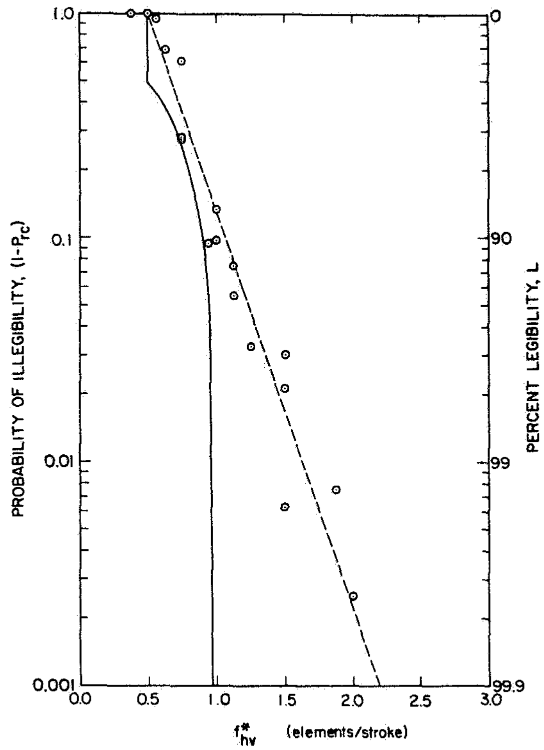


Fig. 2.2-4. REGRESSION LINE FOR SQUARE ELEMENTS.

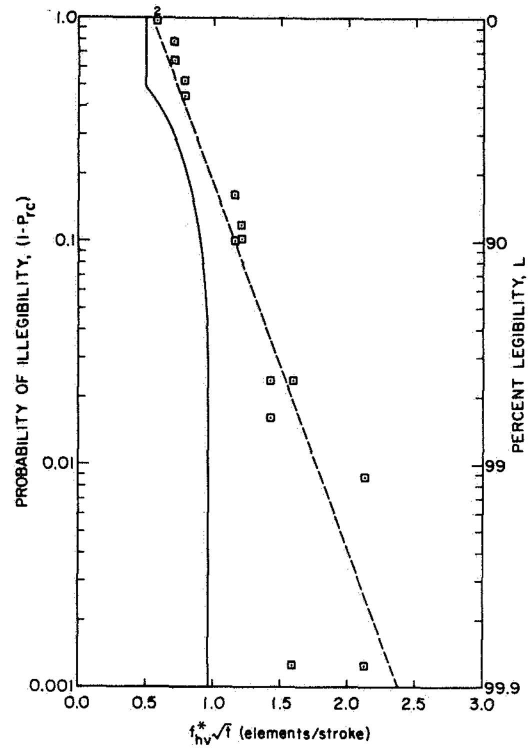


Fig. 2.2-5. REGRESSION LINE FOR RECTANGULAR ELEMENTS ALONG $t = 0.5, 2.0$

Notice that f^*_{hv} is a projection onto either axis of points along the 45° ray where:

$$f^*_{hv} = \sqrt{f^*_h \, f^*_v}\ , \quad \text{for} \quad f^*_h = f^*_v = f^*_{hv} \qquad \text{elem./stroke}$$

This is not the same as the dimension along the 45° ray which is:

$$1.414\ f^*_{hv} = \sqrt{(f^*_h)^2 + (f^*_v)^2}\ , \quad \text{for} \quad f^*_h = f^*_v = f^*_{hv} \qquad \text{elem./stroke}$$

The next step was to combine Eq. 2.2-1,-2, and-3 into a 3-dimensional analytic expression for legibility over all resolutions. The exponential form (Eq. 2.2-3) is like a boundary condition for the hyperbolas (Eq. 2.2-1) along the 45° ray. Solving Eq. 2.2-3 for f^*_{hv} yielded:

$$f^*_{hv} = b - c \ln(1 - P_{rc}), \qquad \text{for} \quad f^*_{hv} \geq b \tag{2.2-5}$$

Next the constraint of Eq. 2.2-2 was combined with Eq. 2.2-1 and solved for the case where $f^*_h = f^*_v$:

$$f^*_{hv} = \sqrt{K^*} + a_{hv} \tag{2.2-6}$$

Combining Eq. 2.2-5 and -6, and solving for K produced:

$$K = [b - c \ln(1 - P_{rc}) - a_{hv}]^2 , \qquad \text{for} \quad c \ln(1 - P_{rc}) \geq 0 \tag{2.2-7}$$

Having solved for K at the $f^*_h = f^*_v$ boundary condition, insertion of Eq. 2.2-7 into Eq. 2.2-1 yielded the complete form:

$$(f^*_h - a_{hv})(f^*_v - a_{hv}) = [b - c \ln(1 - P_{rc}) - a_{hv}]^2, \quad \text{for} \quad c \ln(1 - P_{rc}) \geq 0 \tag{2.2-8}$$

And solving for P_{rc} produced:

$$P_{rc} = 1 - \exp\left[-\frac{1}{c}\left(\sqrt{(f^*_h - a_{hv})(f^*_v - a_{hv})} + a_{hv} - b\right)\right],$$

$$\text{for} \sqrt{(f^*_h - a_{hv})(f^*_v - a_{hv})} + a_{hv} \geq b \tag{2.2-9}$$

The form of Eq. 2.2-4 has been preserved, given the relationship:

$$f^*_{hv} = \sqrt{(f^*_h - a_{hv})(f^*_v - a_{hv})} + a_{hv}, \quad \text{for} \quad f^*_h = f^*_v = f^*_{hv} \tag{2.2-10}$$

Notice that the same exponential character can be preserved away from the 45° ray. Along rays emanating from an origin, a_{hv}, and with a slope defined as:

$$t = \frac{(f^*_v - a_{hv})}{(f^*_h - a_{hv})} \tag{2.2-11}$$

For $a_{hv} = 0$, the form for Eq. 2.2-9 becomes:

$$P_{rc} = 1 - \exp\left[-\frac{1}{c}\left(f_h^* \sqrt{t} - b\right)\right], \quad \text{for} \quad f_h^* \sqrt{t} \geq b \qquad (2.2\text{-}12a)$$

Examples of the rays being described can be seen in Fig. 2.2-2 emanating from an origin (0,0). The term $\sqrt{t}$ represents the scaling required to express distance along the ray in terms of the horizontal axis. A similar expression can be obtained for f_v^*, reflecting the scale along such a ray as Eq. 2.2-12 in terms of the vertical axis.

This fact permitted initial checks of the proposed model. Data points along rays such as in Fig. 2.2-2 could be plotted on semi-log paper to see if the results looked straight enough. Figure 2.2-5 illustrates such a plot. The sixteen data points are from 30° and 60° rays used in combination on the assumption of symmetry ($t = 0.5$ and 2.0).

The development of the analytic model allows a fit to the data using only one overall semi-log plot (rather than one for each ray). Since only the dependent variable, P_{rc}, is statistical, the independent variables, f_h^* and f_v^*, that define a data point, can be readily transformed. When the f_h^* and f_v^* values for a data point are mapped into an equivalent f_{hv}^* value using the model:

$$f_{hv}^* = \sqrt{(f_h^* - a_{hv})(f_v^* - a_{hv})} + a_{hv} \qquad (2.2\text{-}10)$$

then all data points can be fitted simultaneously on a plot of $\ln(1 - P_{rc})$ vs f_{hv}^*. This will be illustrated in the next section for various values of a_{hv}.

2.2-2. Fitting the model using multiple regression analysis.

The exponential and hyperbolic structure of the legibility data suggested the functional relationship of Eq. 2.2-9. This function can be considered a regression equation with multiple independent variables f_h^* and f_v^*, and dependent variable P_{rc}. The other parameters such as a_{hv}, b, and c then become the coefficients to be fitted in the regression analysis. (Some good references for the statistical discussion in this section are [Ezekiel and Fox - 1963], and [Ostle - 1963]).

The first analysis was done on the exponential relationship of Eq. 2.2-4. Here, only the data for square elements was used. Of special interest was the parameter b, which has physical meaning. According to the system design (Appendix 1), total dropout should occur for strokes less than half the width of the sampling interval, when $dh = dv$:

$$P_{rc} = 0, \quad \text{for} \quad f^*_{hv} = \frac{\omega_s}{d_{hv}} \leq 0.5$$

For the analytic expression of Eq. 2.2-4:

$$P_{rc} = 0, \quad \text{for} \quad f^*_{hv} = b$$

The critical point is how well the physical constraint

$$b = 0.5 \qquad (2.2\text{-}13)$$

matches to the data taken.

For mathematical tractability a linear analysis of the regression equation was performed, using the transformed expression (Eq. 2.2-3) obtained by taking the natural logarithm of Eq. 2.2-4. This fits the linear regression form

$$Y = a_0 + a_1 X$$

with

$$X = f^*_{hv}$$

$$Y = \ln(1 - P_{rc})$$

$$a_0 = \frac{b}{c}$$

$$a_1 = -\frac{1}{c}$$

All the data points on Fig. 2.2-4 were used except the single value where $f^*_{hv} < 0.5$ (two other data points where $(1 - P_{rc}) = 0$, could also not be used nor plotted).

The results of the analysis were that:

$$a_0 = 1.837 \pm 0.294$$

$$a_1 = 3.929 \pm 0.248$$

$$\bar{S}_Y = 0.445$$

where the accompanying intervals are the standard error of the regression coefficients, and $\bar{S}_Y$ is the standard error of estimate (standard deviation about the regression line) defined as:

$$\bar{S}_Y = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{(N - 2)}$$

Transformed back into a probability by taking the antilog yields:

$$e^{\bar{S}_Y} = 1.560$$

which is an expression for multiplicative error which increases with the value of illegibility, $(1 - P_{rc})$.

To properly evaluate the regression results, residuals between real and estimated values must be computed in the original domain of $(1 - P_{rc})$:

$$\bar{S}_L = \frac{\sum_i (L_i - \hat{L}_i)^2}{(N - 2)}$$

where:

$$\hat{L}_i = 100\left(1 - e^{\hat{Y}_i}\right)$$

are the estimates of L_i taken from the estimate $\hat{Y}_i$ in the transformed analysis. This additive standard error of estimate for L_i turns out to be:

$$\bar{S}_L = 9.14$$

The adjusted coefficient of determination, R_Y^2 for the Y analysis was

$$R_Y^2 = 0.981$$

This would usually imply an estimate that 98.1 percent of the variation in legibility is accounted for by the regression model. Caution must be observed, however, in interpreting this parameter; as well as in interpreting the standard error of the regression coefficients. Their meaning is ordinarily based upon random sampling of the independent variable from an underlying normal distribution. In this experiment the independent variables are controlled, and in effect their distribution has been chosen in advance. Under these circumstances the meanings of the coefficient of determination and standard error of regression coefficients are restricted to the specific distribution of this experiment. Thus, since the variance of the set of 100 resolution combinations is arbitrary comparisons based on this variance are only valid for the exact same set of observations. It will turn out that even in this restrictive sense these parameters are useful. They can be used for comparison between various models fitted to both the uppercase and lowercase data, since the identical resolution values are used in all the analyses.

Returning to the evaluation of results from the regression analysis, solving for b gives:

$$b = -\frac{a_0}{a_1} = 0.468$$

This result is strongly encouraging, being close to the theoretical value of 0.5 that was anticipated. A value slightly lower than 0.5 could be caused by intersections of strokes in a character that might not drop out yet at $f^*_{hv} = 0.5$. Also, variations in a_0 and a_1 over one standard error would produce values of b ranging from 0.578 to 0.369, easily including the point 0.5. The value for the second parameter, the decay constant, c, was:

$$c = -\frac{1}{a_1} = 0.255 \pm 0.016$$

Having found b close to 0.5, I decided to fix it to its theoretical value and repeat the regression analysis without a constant term using:

$$X = f^*_{hv} - 0.5$$

$$Y = \ln(1 - P_{rc})$$

$$a_0 = 0$$

$$a_1 = -\frac{1}{c}$$

The result has only a slightly larger value for a_1

$$a_0 = 0.0 \pm 0.110$$

$$a_1 = -4.068 \pm 0.147$$

and the value for c turned out to be similar:

$$c = -\frac{1}{a_1} = 0.246 \pm 0.009$$

It is this fit for c_1 with $b = 0.5$, that is plotted as a dotted line on Fig. 2.2-3 and -4.

Similar measurements were made for lowercase Mid-Century type, and upper- and lowercase Dual Gothic type, (Documents B1-B5). These results are summarized in Table 2.2-1 below:

Table 2.2-1

REGRESSION ANALYSES FOR SQUARE ELEMENTS

Type Font	N	a_0	a_1	b	c	R_Y^2	$\bar{S}_L$
Uppercase, Mid-Century	17	1.837 ± 0.294	-3.929 ± 0.248	0.468	0.255	0.943	11.43
		0.0 ± 0.110	-4.068 ± 0.147	0.5	0.246	0.981	9.11
Lowercase, Mid-Century	17	2.299 ± 0.332	-3.564 ± 0.177	0.585	0.255	0.929	14.45
		0.0 ± 0.131	-3.564 ± 0.177	0.5	0.273	0.964	15.22
Uppercase Dual-Gothic	5	0.0 ± 0.272	-5.504 ± 0.455	0.5	0.183	0.981	---
Lowercase, Dual-Gothic	5	0.0 ± 0.344	-6.228 ± 0.562	0.5	0.161	0.976	---

Case A. $a_{hv} = 0.0$:

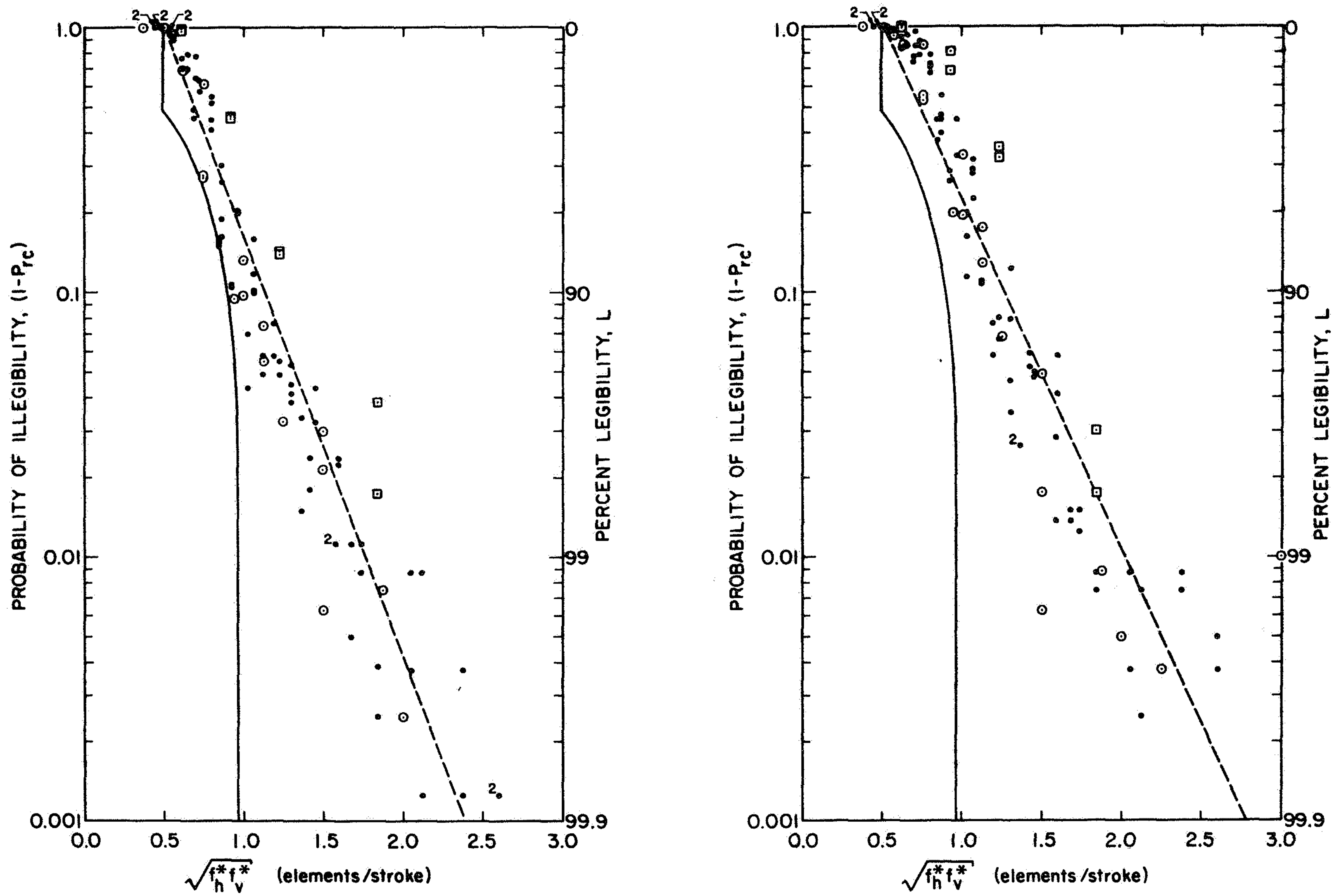
A simple regression was next performed to plot the line in Fig. 2.2-5. The question was whether the data looked like a $1-e^{-x}$ form along the rays of Fig. 2.2-2. Having been plotted in terms of equivalent square element dimensions, $f^*_{hv}\sqrt{t}$, a comparison could be made directly with the data in Fig. 2.2-4. The encouraging results included an intercept near 0.5. As a result, the assumption $a_{hv} = 0$ was first used to fit all the data using:

$$P_{rc} = 1 - \exp\left[-\frac{1}{c}\left(\sqrt{f^*_h f^*_v} - b\right)\right], \quad \text{for} \quad \sqrt{f^*_h f^*_v} \geq b \qquad (2.2\text{-}9a)$$

Figure 2.2-6a illustrates all the uppercase data for Mid-Century type, with the independent variables f^*_h and f^*_v transformed into equivalent square-element dimensions:

$$f^*_{hv} = \sqrt{f^*_h f^*_v} \qquad (2.2\text{-}10a)$$

and plotted as before, using Eq. 2.2-3. Figure 2.2-6b shows the same data for lowercase, Mid-Century type. For no constraint on a_0, the



a. Uppercase

b. Lowercase

Fig. 2.2-6. REGRESSION LINE FOR ALL ELEMENTS, WITH $a_{hv} = 0.0$.

intercepts of their regression lines were at $b = 0.445$ and 0.569 respectively (with $b = 0.5$ well within their variations as a_0 and a_1 range over one standard error). The dashed regression lines that are shown come from the successive analyses that were constrained to pass through $b = 0.5$.

The regression analyses for the data in Fig. 2.2-6a and b, are summarized in Table 2.2-2:

Table 2.2-2

REGRESSION ANALYSES FOR ALL ELEMENTS, WITH $a_{hv} = 0.0$

Type Font	N	a_0	a_1	b	c	R_Y^2	$\overline{S}_L$
Uppercase, Mid-Century	89	1.549 ± 0.159	-3.479 ± 0.120	0.445	0.288	0.906	11.85
		0.0 ± 0.061	-3.657 ± 0.070	0.5	0.274	0.969	9.18
Lowercase, Mid-Century	89	1.896 ± 0.162	-3.335 ± 0.122	0.569	0.300	0.895	12.87
		0.0 ± 0.063	-3.121 ± 0.071	0.5	0.320	0.956	15.28

Comparison with Table 2.2-1 revealed that the standard error of estimate for legibility, $\overline{S}_L$, had remained about the same for each analysis. Apparently, the addition of data for rectangular scan elements had not disturbed the model.

However, the model with $a_{hv} = 0$ had one drawback when the loci of equal legibility were considered. Equation 2.2-1 had become:

$$f_h^* f_v^* = K^*, \quad \text{for} \quad P_{rc}[K^*] = K_L \tag{2.2-1a}$$

These hyperbolas with origin (0,0) were also loci of equal scan element area:

$$f_h^* f_v^* = \frac{w_s^2}{A} \tag{2.2-14}$$

But according to previous observations [Arps, et al - 1968] along loci of constant legibility, square scan elements had the largest surface area. Stated in another way, given values along Eq. 2.2-14, legibility was maximum for square elements:

$$P_{rc}[K^*] = K_L, \quad \text{for} \quad f_h^* = f_v^*$$

$$< K_L, \quad \text{for} \quad f_h^* \neq f_v^* \qquad (2.2\text{-}15)$$

The inequality in Eq. 2.2-15 is also consistent with the theoretical model proposed in section 2.2.2. When non-square elements are modeled, their legibility is determined entirely by the smaller resolution frequency.

These contradictions could be resolved by reverting to the general hyperbola of Eq. 2.2-1 with the constraint:

$$a_{hv} > 0 \qquad (2.2\text{-}16)$$

Case B. $a_{hv} = 0.5$.

Further consideration of the theoretical model raised the possibility that the analytic model of Eq. 2.2-9 could be matched at more points than just $b = 0.5$ along the 45° ray. Using $a_{hv} = 0.5$, $b = 0.5$, it becomes:

$$P_{rc} = 1 - \exp\left[- \frac{1}{c}\sqrt{(f_h^* - 0.5)(f_v^* - 0.5)}\right] ,$$

$$\text{for } \sqrt{(f_h^* - 0.5)(f_v^* - 0.5)} > 0.5 \qquad (2.2\text{-}9b)$$

along the asymptotes $(f_h^*, 0.5)$ and $(0.5, f_v^*)$ the legibility goes to zero, as in the case for the theoretical model (see Fig. 2.1-4). As a result, the origin for the hyperbolas was moved to (0.5, 0.5), resulting in rays with the form:

$$P_{rc} = 1 - \exp\left[\frac{1}{c}(f_h^* - 0.5)\sqrt{t}\right] \qquad (2.2\text{-}12b)$$

with:

$$t = \frac{(f_v^* - 0.5)}{(f_h^* - 0.5)} \qquad (2.2\text{-}11b)$$

as illustrated in Fig. 2.2-7.

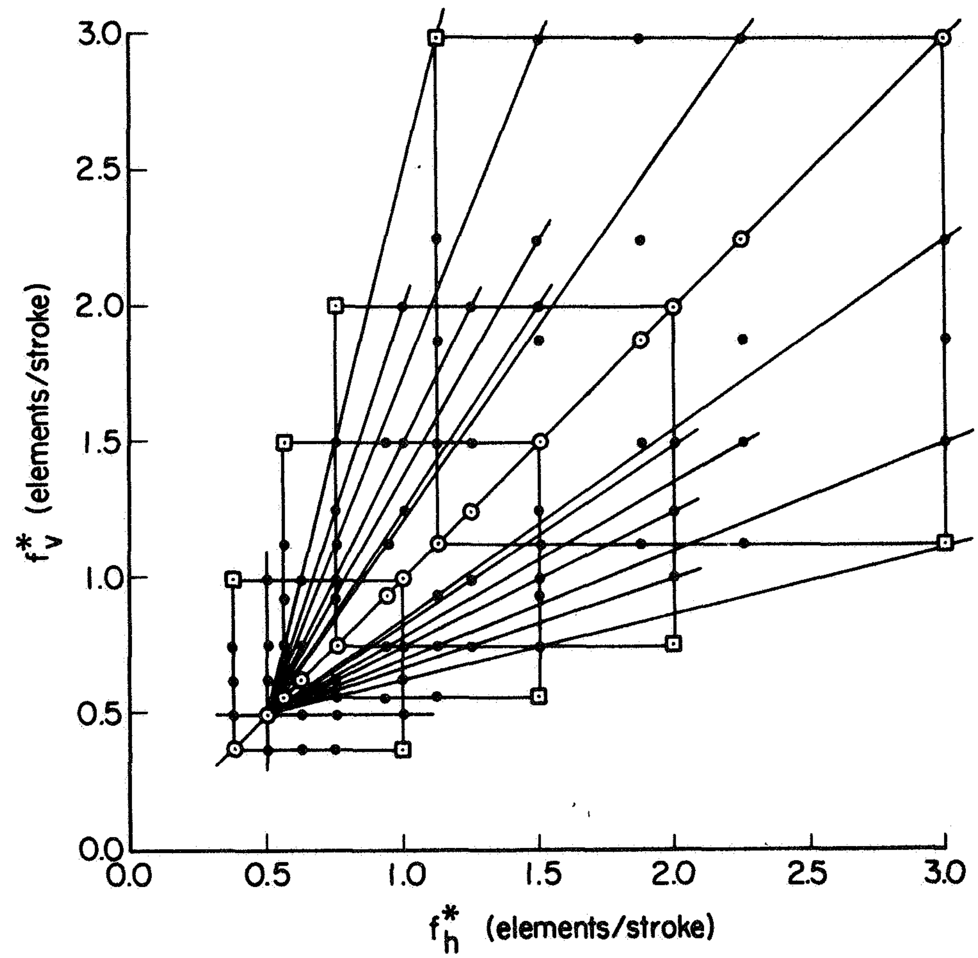


Fig. 2.2-7. NORMALIZED RESOLUTION COMBINATIONS, WITH $a_{hv} = 0.5$.

With hopes for a better model the data was plotted again as shown in Fig. 2.2-8a and b. This time, a simple regression to test for $b = 0.5$ was not practical. The regression lines were obtained using the form:

$$X = (f_h^* - 0.5)(f_v^* - 0.5)$$

$$Y = \ln(1 - P_{rc})$$

$$a_0 = 0$$

$$a_1 = -\frac{1}{c}$$

and the analyses are summarized in Table 2.2-3:

Table 2.2-3

REGRESSION ANALYSES FOR ALL ELEMENTS, WITH $a_{hv} = 0.5$

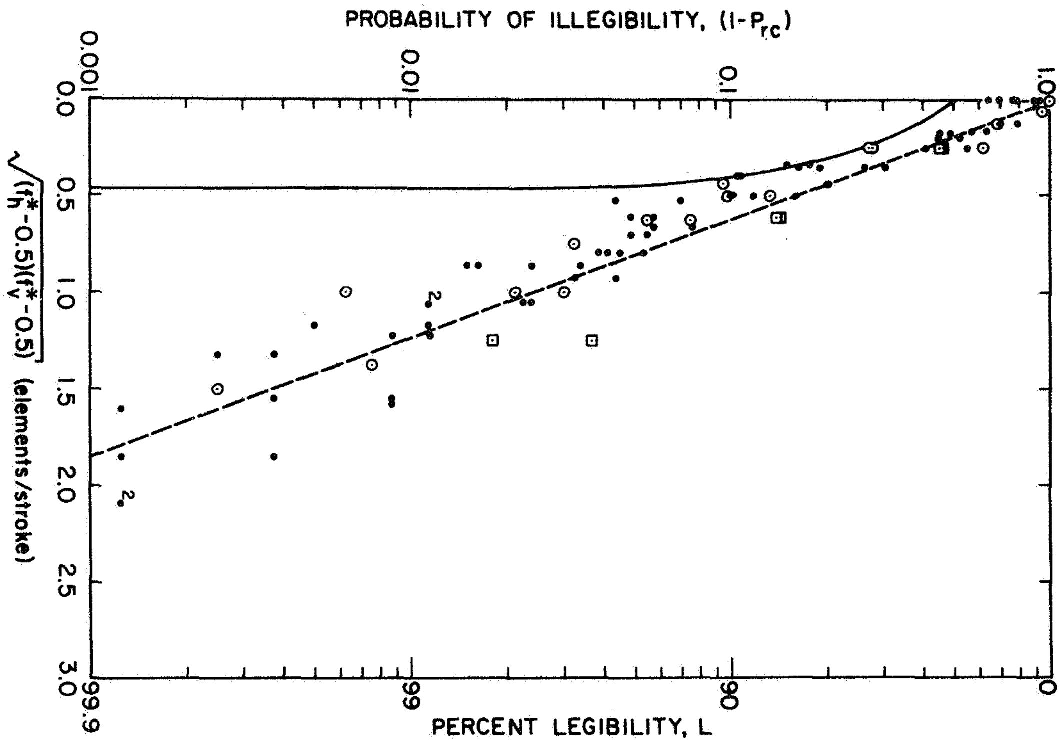
Type Font	N	a_0	a_1	b	c	R_Y^2	$\overline{S}_L$
Uppercase, Mid-Century	89	0.0 ± 0.058	-3.747 ± 0.068	0.5	0.267	0.973	8.74
Lowercase, Mid-Century	89	0.0 ± 0.056	-3.206 ± 0.065	0.5	0.312	0.965	12.37

The standard errors of estimate for legibility, $\overline{S}_L$, showed significant improvement over previous results. This confirmed the desirability of modeling the hyperbolic origin at (0.5,0.5). The coefficient of determination, R_Y^2, also improved steadily with each step; and the decay constants, c, are closer to the values obtained for square elements alone.

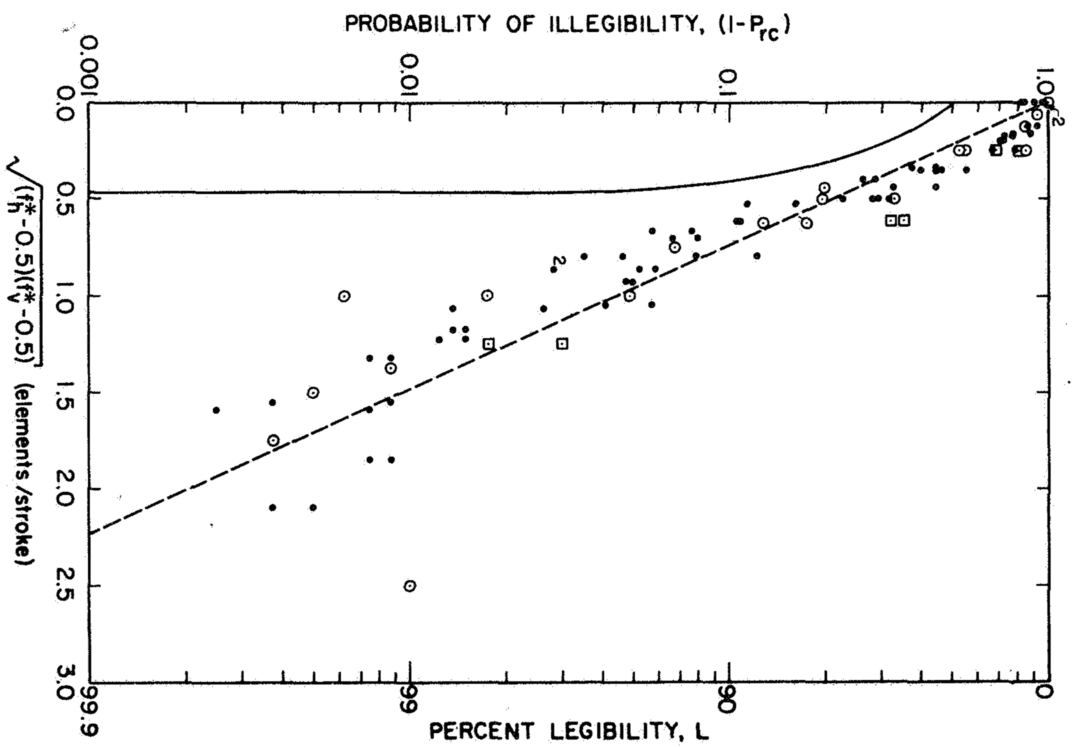
One further refinement was attempted. A second order effect was found for data points progressively farther from the 45° ray for both case A and B. To illustrate this, the farthest points were marked with squares in the resolution planes of Fig. 2.2-2 and 2.2-7. Their corresponding legibilities were also marked this way when plotted in Fig. 2.2-6 and 2.2-8. These data points line up with a negative slope that is less than for the other points on the figures. This could be accommodated in the model by changing the square root in Eq. 2.2-9 to an arbitrary exponential value, g.

For Case B, Eq. 2.2-9b was transformed into the form:

$$\ln[-c \ln(1 - P_{rc})] = g[(f_h^* - 0.5)(f_v^* - 0.5)] \qquad (2.2\text{-}17)$$



a. Uppercase



b. Lowercase

Fig. 2.2-8. REGRESSION LINE FOR ALL ELEMENTS, WITH $a_{hv} = 0.5$.

and regression coefficients calculated, using

$$X = (f_h^* - 0.5)(f_v^* - 0.5)$$

$$Y = \ln[-0.267 \ln(1 - P_{rc})]$$

$$a_0 = 0$$

$$a_1 = g$$

having used the value for c for uppercase from Table 2.2-3. As a result, the value for the exponent became:

$$g = 0.524 \pm 0.016$$

confirming that a departure from $g = 0.5$ might increase the fit to the data.

A similar analysis was performed using just the square element data to evaluate the choice of $1 - e^{-x^g}$ with power $g = 1$:

$$X = f_{hv}^* - 0.5$$

$$Y = \ln[-0.246 \ln(1 - P_{rc})]$$

$$a_0 = 0$$

$$a_1 = g$$

having used the value for c for uppercase with $b = 0.5$, from Table 2.2-2. This time the exponent came out quite close to its proper value

$$g = 1.014 \pm 0.059$$

with $g = 1.0$ well within one standard error.

In summary, the analytic model proposed for legibility is:

$$L = 100\left\{1 - \exp\left[-\frac{1}{c}\sqrt{(f_h^* - 0.5)(f_v^* - 0.5)}\right]\right\} \% \qquad (2.2\text{-}18)$$

with $c = 0.267$ and 0.312 for upper- and lowercase Mid-Century type respectively. The corresponding standard errors of estimate measured for the model were 8.74 percent and 12.37 percent. For just square elements, this model degenerates to:

$$L = 100\left\{1 - \exp\left[- \frac{1}{c}(f^*_{hv} - 0.5)\right]\right\} \qquad \% \qquad (2.2\text{-}19$$

with the same values of c for Mid-Century type and $c = 0.182$ and 0.161 having been measured respectively for upper- and lowercase Dual-Gothic type.

Chapter III

THE ENTROPY OF PRINTED MATTER WITH SPATIAL FREQUENCY

With a printed document digitized as discussed previously, attention can be focused next on the number of binary digits used to represent it. The dissection itself produces binary "elements" (with dimensions specified by the resolution grid). Information theory indicates the potential for encoding such messages into representations that, on the average, have fewer binary digits than uncoded representations [Shannon - 1948].

This compression of average message "length" stems from a-priori knowledge about the messages to be processed. Here, the emphasis on printed matter specifies messages (images) with a distinctive statistical structure suited to compression coding. The next sections deal with this statistical structure and the scan patterns to extract it. An empirical model is presented expressing the entropy of these scan patterns as a function of resolution, character size and density, and page area. Finally, a measure for efficient compression is proposed.

3.1. Source Alphabets for Line-by-Line Scanning

The problem of synthesizing a compression code is classically defined in terms of a "source," its "alphabets," and a-priori distributions over these alphabets. Minimum average length codes can be designed to match any of the given distributions [Huffman - 1952]. In this problem, the source is output from a document scanner. Elements in the output stream may be grouped in a variety of ways to form "patterns," that map into symbols of a source alphabet. The statistics of each such source alphabet and the ultimate compression it achieves, depends on the match between its scan pattern and typical input images.

The following subsections explore various scan patterns, their resulting statistics, and the compression achieved for document group C1 (described in Appendix A.2). Intuitively, the objective is to find scan patterns that define natural source alphabets--decompositions that capture the information present in a typical image with a minimum of accompanying redundancy.

3.1.1 Element alphabets

Consider the source alphabet, X:

$$X = \{x_0, x_1\}$$

The decomposition of an image for X is a mapping of single black or white output elements into the symbols x_0 or x_1 respectively. This simple mapping uses a scan pattern extending over only one element for decomposing the input image. Figure 3.1-1a illustrates decomposition of four elements in the horizontal scan of an idealized letter "H." The scan is shown producing symbols from alphabet X in the order x_0, x_1, x_1, x_1.

Another pattern, defined as covering groups of n successive elements, can decompose the image into an alphabet of 2^n symbols. This alphabet, assuming successive scan elements are independent, is the n^{th} extension of X, written X^n:

$$X^n = \{x_0^n, x_1^n, x_2^n, \ \ldots \ , x_m^n\} \quad \text{where} \quad m = 2^n - 1$$

Figure 3.1-1b illustrates decomposition of the four elements into only two symbols, using the 2^{nd} extension of X as the alphabet. Note that for line-by-line scanning "successive" elements are physically adjacent. In other schemes such as Pseudo-Random Scanning successive elements would be distributed about the image almost at random [Huang - 1964].

The encoding process consists of mapping symbols from the source alphabet one-to-one into equivalent binary "words." This is done to represent document images with (on the average) fewer binary digits. The average length of these words, L_n, can be bounded for any n^{th} extension using Shannon's First Theorem [Abramson - 1963]:

$$H[X] \leq \frac{L_n}{n} < H[X] + \frac{1}{n} \qquad \text{bits/symbol} \qquad (3.1.1)$$

Here, H[X] is the entropy of the source--defined over the probability distribution of the alphabet symbols:

$$H[X] = -\sum_i P[x_i] \log_2 P[x_i] \qquad \text{bits/symbol} \qquad (3.1\text{-}2)$$

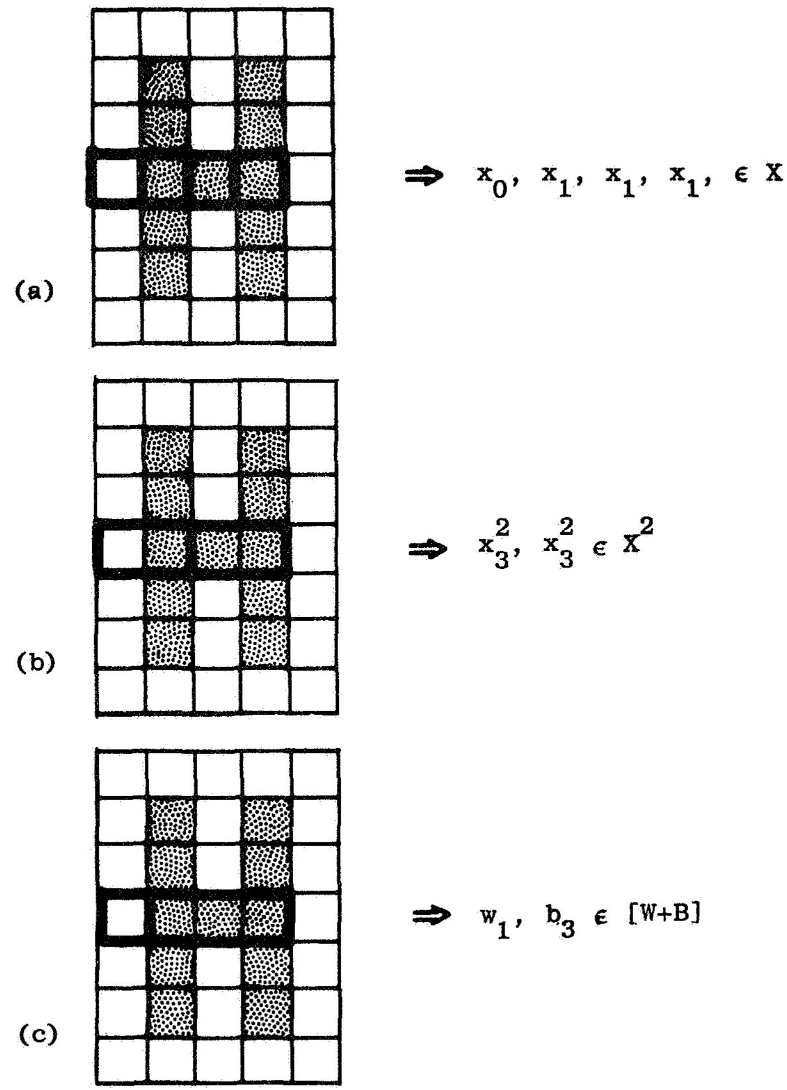


Fig. 3.1-1. EXAMPLES OF 1-D SCAN PATTERNS
(HORIZONTAL LINE-BY-LINE SCANNING.)

As an example, let us use the distribution of X for the Pica type document C1. Of the 730,281 elements sampled, 9.5 percent were black. Expressed as estimates with accompanying standard error, the symbol probabilities are:

$$\hat{P}[x_0] \quad 0.095 \pm 0.0003$$

$$\hat{P}[x_1] \quad 0.905 \pm 0.0003$$

Assuming this to be the actual distribution, the entropy is calculated and Eqs. 3.1-1 and 2 yield:

$$H[X] = 0.453 \qquad \text{bits/symbol}$$

$$0.453 \leq \frac{L_n}{n} < 0.453 + \frac{1}{n} \qquad \text{bits/symbol} \tag{3.1-3}$$

The 1st extension, for example, would require words with average length:

$$0.453 \leq \hat{L}_1 < 1.453 \qquad \text{bits}$$

The 2nd extension would require longer words with double the entropy:

$$0.906 \leq \hat{L}_2 < 1.906 \qquad \text{bits}$$

The 2^{nd} extension, however, has the source symbols grouped in twos. The average length per original symbol is more meaningful for comparison:

$$0.453 \leq \frac{\hat{L}_2}{2} < 0.953 \qquad \text{bits/symbol}$$

Notice that the upper bound has tightened.

Dependency can also be assumed within the scan patterns of X^2, and distributions estimated directly. This is in contrast to using X, assuming independence, and deriving $P\{x_i^2\}$s as the product of two $P\{x_i\}$s. Defining these groups of two as symbols from a joint source alphabet, [X,Y], with

$$X = \{x_0 = w, x_1 = b\}, \text{ at time } t$$

$$Y = \{y_0 = w, y_1 = b\}, \text{ at time } (t+1)$$

A new joint entropy with four terms may be estimated:

$$H[X,Y] = -\sum_i \sum_j P[x_i,y_j] \log_2 P[x_i,y_j] \quad \text{bits/symbol} \qquad (3.1\text{-}4)$$

The resultant entropy calculation yields:

$$H[X,Y] = 0.741 \qquad \text{bits/symbol}$$

$$0.741 \leq \hat{L}_1 < 1.741 \qquad \text{bits/symbol}$$

The dependency present in groups of two can also be utilized by computing the conditional distributions for an element, given the preceding scan element. The source is then modeled as a 1st order Markov chain (Fig. 3.1-2) with possible states b and w for the successive trials X and Y. The conditional entropy for each possible x_i is calculated using:

$$H[Y|X=x_i] = -\sum_j P[y_j|x_i] \log_2 P[y_j|x_i] \quad \text{bits/symbol} \qquad (3.1\text{-}5)$$

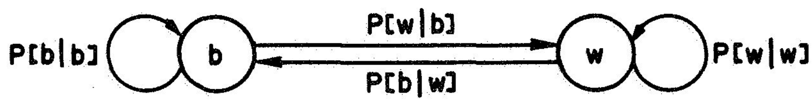


Fig. 3.1-2. MARKOV MODEL FOR CONDITIONAL SOURCE.

This model is treated as two cases, depending on the prior state:

$$H[Y|X=x_0] = 0.930 \qquad \text{bits/symbol}$$

$$H[Y|X=x_1] = 0.232 \qquad \text{bits/symbol}$$

Similarly, Shannon's First Theorem is applied for each case:

$$0.930 \leq \hat{L}_1 < 1.930 \ , \qquad \text{given } X=x_0$$

$$0.232 \leq \hat{L}_1 < 1.232 \ , \qquad \text{given } X=x_1$$

The conditional entropy, given the entire set X, is a weighted average:

$$H[Y|X] = P[x_0]\, H[Y|X{=}x_0] + P[x_1]\, H[Y|X{=}x_1] \text{ bits/symbol} \tag{3.1-6}$$

with value:

$$H[Y|X] = 0.288$$

The average word length becomes:

$$0.288 \leq L_1 < 1.288$$

Although the joint and conditional entropies both take dependency between two elements into consideration, they utilize it in different ways. An alphabet based on the joint entropy sends scanner output for groups of two elements; entropy must be halved to compute the bits per element. The conditional approach only sends output for one element, using one of two alphabets depending on the prior state. The two can be reconciled using the identity:

$$H[X,Y] = H[X] + H[Y|X]$$

which can be verified using the entropies in the previous examples. The conditional entropy is just the additional information required to complete the task of sending both scan elements.

A good measure for comparison of source alphabets is the compression, C_n, defined as:

$$C_n = \frac{np}{L_n} \quad \text{bits/symbol} \tag{3.1-7}$$

This is simply the reciprocal of previous expressions, adjusted by the number, p, of image elements in a source symbol or pattern. Combining with Eq. 3.1-1 gives a bounded relationship for the compression:

$$\frac{p}{H[X] + \frac{1}{n}} < C_n \leq \frac{p}{H[X]} \quad \text{elements/bit}$$

Compression values for the previous examples are as follows:

For [X]:

$$p = 1 \qquad \text{elements/symbol}$$

$$0.69 < C_1 \leq 2.21 \qquad \text{elements/bit}$$

For $[X^2]$:

$$p = 2 \qquad \text{elements/symbol}$$

$$1.05 < \hat{C}_1 \leq 2.21 \qquad \text{elements/bit}$$

For [X,Y]:

$$p = 2 \qquad \text{elements/symbol}$$

$$1.15 < \hat{C}_1 \leq 2.70 \qquad \text{elements/bit}$$

For $[Y|X]$:

$$p = 1 \qquad \text{elements/symbol}$$

$$0.776 < \hat{C}_1 \leq 3.46 \qquad \text{elements/bit}$$

My purpose in elaborating in such detail with examples, is to convey some of the "feel" for these definitions to be used throughout the next sections. Attention will be focused on estimates of the upper bound for compression, and its dependency on the scan pattern chosen. Notice that the using of n^{th} extensions serves merely to tighten the lower bound. This expresses the improvement actually obtained by using n^{th} extension codes to approach the upper limit for C. These theoretical coding techniques are well developed. However, the matching of source alphabets to significant populations in the real world is the challenge. Insight into the real population is required to find "natural" source alphabets that match the inherent structure and maximize the upper bound for compression.

3.1.2. Run-length alphabets

One natural source alphabet for line-by-line scanning uses variable scan patterns that cover the length of a "run" [Laemmel - 1951]. A run is an adjacent series of output elements all of the same "color"-- black or white.

The run-length source alphabet [W+B] may be defined in terms of two subalphabets, one for white and the other for black runs

$$W = \{w_1, w_2, w_3, \ldots\}$$

$$B = \{b_1, b_2, b_3, \ldots\}$$

where the total set has been labeled [W+B], using "+" to denote the union between two subalphabets. The length of run is denoted by the subscript. Figure 3.1-1c illustrates a scan producing set members w_1 and b_3, and emphasizes how these scan patterns very in length to match runs.

The estimated distributions for the two subalphabets are shown in Fig. 3.1-3. Although these are discrete densities, the estimated values have been joined by straight lines to emphasize their shape. The horizontal scale has been stretched out to indicate any runs across the full page width. Such runs indicating all-white lines, can be seen in the distribution for P[W]. They have been labeled, along with other recognizable characteristics of the printed page, using the following notation:

w_s = stroke width

w_{ic} = inter-character width

w_{iw} = inter-word width

w_{ec} = end-of-sentence width (peculiar to the test documents C1 and C2)

w_m = margin width

w_p = page width

If runs are independent, [Capon - 1959] showed that they can be modeled as if generated by the Markov source for [Y|X] in Fig. 3.1-2. The distribution of runs is just the geometric distribution for first-passage

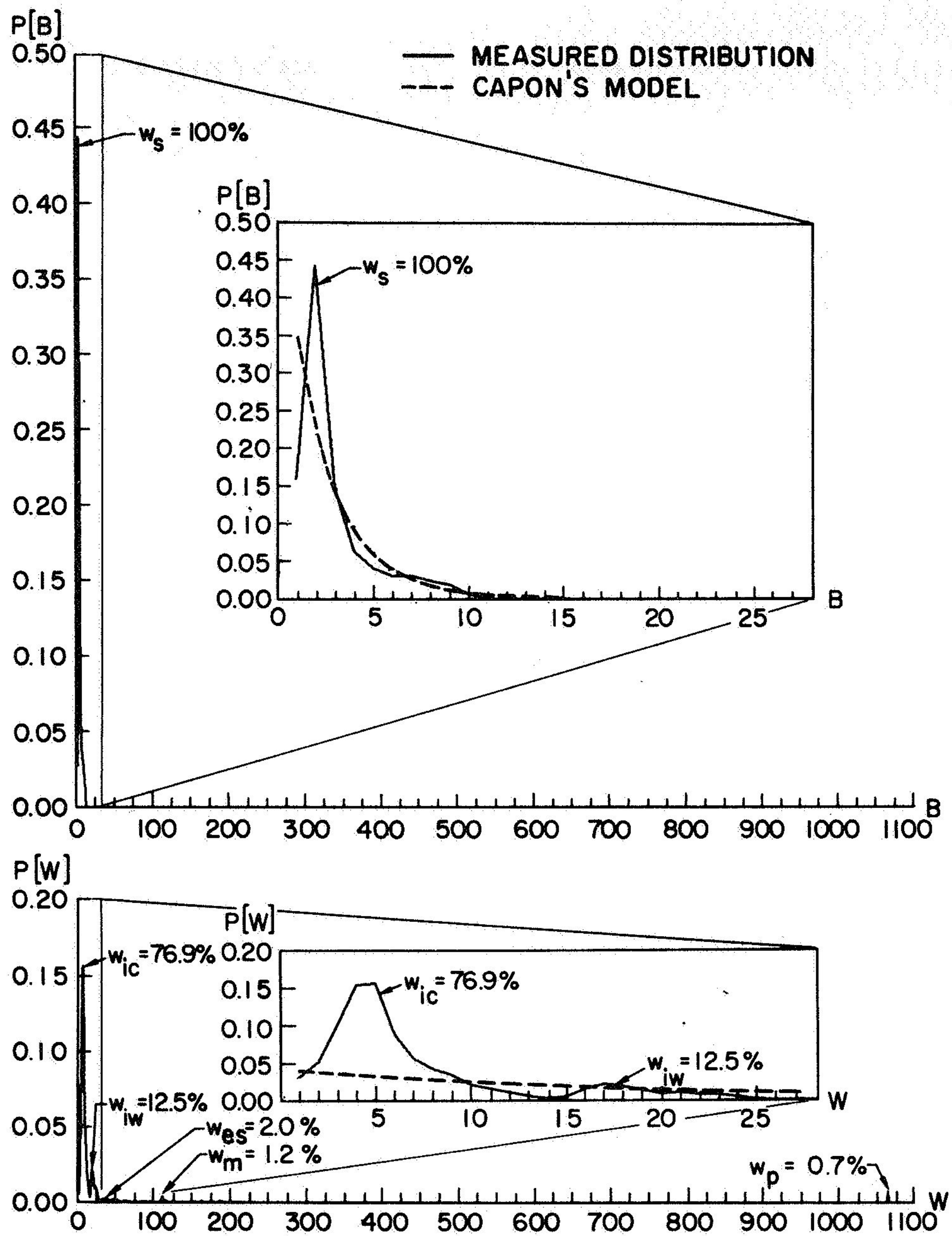


Fig. 3.1-3. INDEPENDENT RUN-LENGTH DISTRIBUTIONS.

times of the Markov states. The exploded views give better detail for the character-sized features. Percentages have been given to indicate that frequency of occurrences (6.7 percent of the distribution for P[W] is down in the noise).

Using these distributions, the entropy for each subalphabet of run-lengths was estimated:

$$\hat{H}[B] = 2.54 \qquad \text{bits/symbol}$$

$$\hat{H}[W] = 4.67 \qquad \text{bits/symbol}$$

The weighted averages of scan pattern lengths, p_{ave}, are needed for computing compression:

$$\hat{p}_{ave}\,[B] = \sum_i i\,\hat{P}[b_i] = 2.88 \quad \text{elements/symbol}$$

$$\hat{p}_{ave}\,[W] = \sum_i i\,\hat{P}[w_i] = 26.62 \quad \text{elements/symbol}$$

The bounds on compression estimated for each subalphabet using Eq. 3.1-6 are:

$$0.81 < \hat{C}_1[B] \leq 1.13 \qquad \text{elements/bit}$$

$$4.70 < \hat{C}_1[W] \leq 5.71 \qquad \text{elements/bit}$$

For the total alphabet, these can be combined as a weighted average of the <u>bits</u> each subalphabet generates per input element (hence, reciprocals are required):

$$\frac{1}{C_n[W+B]} = \frac{P[a_1]}{C_n[W]} + \frac{P[a_0]}{C_n[W]} \qquad \text{bits/element} \qquad (3.1\text{-}8)$$

where the weights are just the probability of black or white elements as measured previously for the alphabet, X. The combined compression is:

$$3.24 < \hat{C}_1[W+B] \leq 4.14 \qquad \text{elements/bit}$$

expressed here for suitable comparison with the preceding results.

Compressions for the various one-dimensional scan patterns are summarized in Table 3.1-1. Accompanying the data for Pica type document C1, is data for an identical document that used Dual-Gothic type.

Table 3.1-1

COMPRESSION FOR SOURCE ALPHABETS
(ONE-DIMENSIONAL)

Source Alphabet	Document C1 (Pica)	Document C2 (Dual-Gothic)
[X]	2.21 elements/bit	2.51 elements/bit
[X,Y]	2.70 " "	3.00 " "
[Y|X]	3.46 " "	3.76 " "
[W+B]	4.14 " "	4.73 " "
Blackness	9.5%	8.0%
Resolution (absolute)	8 × 8 mil	8 × 8 mil
Resolution (normalized)	1.62 × 1.62	1.25 × 1.25

If a binary source can be modeled as a two-state, 1st-order Markov chain, the first passage times for the states represent runs [Capon - 1959]. The first-passage times and hence the runs for the two states are simply geometrically distributed. Thus, only one parameter characterizes each distribution, and these turn out to be relatively easy to measure (distributions calculated by Capon's Model are superimposed on the actual run distributions in Fig. 3.1-3 for comparison.) The compression for the Markov source, $H[Y|X]^{-1}$, was shown by Capon to equal the compression obtained by instead encoding its run-lengths (if compared over a large sample). He also proved that for this model, the runs are independently distributed.

As may be seen in Table 3.1-1, $C_1[W+B]$ is greater than $C_1(Y|X)$ for both type fonts. Capon also found the same to be true in comparison with measurements by [Deutsch - 1956]. However, having used a sample over 100 times as large (730,281 vs 5,075), these variations are unlikely to be due to sample size. The lower entropy per element estimated from actual run-length frequencies, implies greater than 1^{st} order dependence between scan elements. In the next section, the assumption of independence between runs will also be examined.

3.2. Extracting the Dependency in Images of Printed Matter

The compression attainable by encoding the dependency of entire runs prompted further thought about the characteristics of dissected print. After scanning through a few strokes, the width of subsequent strokes for that document should be easy to guess. The regularity in character structure allows further predictions to be made. Following a stroke might be an intra-character space--like between the vertical sides of an H, D, or O. Other possibilities would be the space between letters or words.

What is important is that organizing the image into runs gets at these salient features. A run increases its length until the stroke or space that it consists of has been completed. In a sense, the run increases until that stroke or space has been "measured." This is evident in Fig. 3.1-3, where features from the character structure were readily identified. By contrast, blocking consecutive scan elements into fixed-length words attacks the dependency, but in a somewhat random manner. Criticism of the approach used to match the alphabet to source characteristics is not based on just aesthetic reasons. Organizing the scan data in a meaningful way is usually rewarded with greater compression as well as insight.

Having captured the structure of print in using runs, and having recognized that prediction of structure is possible, the next step is to investigate whether runs are predictable. Another way of looking at this question is to ask whether dependency exists between runs.

3.2.1. A Markov Model for the dependence between runs

To characterize possible dependence between runs, a finite Markov model was ~~used~~ as shown in Fig. 3.2-1.

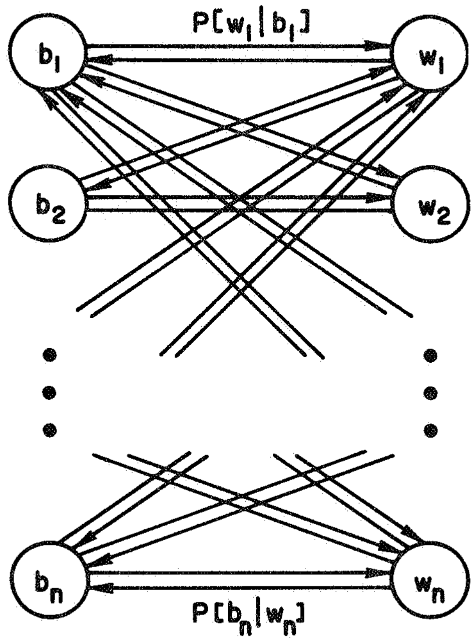


Fig. 3.2-1. MARKOV MODEL OF DEPENDENCY BETWEEN RUNS

Here I have assumed the 1st-order Markov property, that dependence is limited to the immediately preceding state:

$$P[X_i = x_i | X_{i-1} = x_{i-1}, \ldots, X_0 = x_0] = P[X_i = x_i | X_{i-1} = x_{i-1}] \qquad (3.2\text{-}1)$$

where X_i is a random variable representing the state of the model at time i. (A good reference on Markov chains, including their application to information sources, can be found in [Ash - 1965]. For a statistical reference on estimation of transition probabilities, see [Mood and Graybill - 1963].) The Markov chain is shown with a state for each possible run of black or white. Its transition matrix has the form:

$$
P[W{+}B \mid W{+}B] = \left[\begin{array}{cccc|cccc}
0 & 0 & \cdots & 0 & p_{w_1 b_1} & p_{w_1 b_2} & \cdots & p_{w_1 b_n} \\
0 & 0 & & 0 & p_{w_2 b_1} & p_{w_2 b_2} & & p_{w_2 b_n} \\
\vdots & & & \vdots & \vdots & & & \vdots \\
0 & 0 & \cdots & 0 & p_{w_n b_1} & p_{w_n b_2} & & p_{w_n b_n} \\
\hline
p_{b_1 w_1} & p_{b_1 w_2} & \cdots & p_{b_1 w_n} & 0 & 0 & \cdots & 0 \\
p_{b_2 w_1} & p_{b_2 w_2} & & p_{b_2 w_n} & 0 & 0 & & 0 \\
\vdots & & & & \vdots & & & \vdots \\
p_{b_n w_1} & p_{b_n w_2} & & p_{b_n w_n} & 0 & & \cdots & 0
\end{array}\right]
$$

$$
= \left[\begin{array}{c|c} 0 & P[B|W] \\ \hline P[W|B] & 0 \end{array}\right] \qquad (3.2\text{-}1)
$$

where in the 1^{st} and 3^{rd} quadrants:

$$P[B|W] = \{p_{b_i w_j} = P[w_j|b_i] \geq 0\}$$

$$P[W|B] = \{p_{w_i b_j} = P[b_i|w_j] \geq 0\}$$

and in the 2^{nd} and 4^{th} quadrants;

$$P[W|W] = \{p_{w_i w_j} = P[w_j|w_i] \equiv 0\} \equiv 0$$

$$P[B|B] = \{p_{b_i b_j} = P[b_j|b_i] \equiv 0\} \equiv 0$$

Subscripts i and j correspond to consecutive increments in time.

The values in the 2^{nd} and 4^{th} quadrants are zero, since runs alternate in color. This is illustrated in Fig. 3.2-1 by the absence of self-loops or transitions between runs of the same color. A consequence of this is that the stochastic matrix:

$$P[W+B|W+B]$$

can be broken into two stochastic submatrices:

$$P[B|W] = \{p_{b_i w_j}\} \tag{3.2-2a}$$

$$P[W|B] = \{p_{w_i b_j}\} \tag{3.2-2b}$$

such that:

$$\sum_j p_{b_i w_j} = 1$$

$$\sum_j p_{w_i b_j} = 1$$

for the 1st and 3rd quadrants respectively, and $i,j = 1,2, \ldots, n$.

In order to estimate the terms for one of the Markov matrices, frequencies of occurrance, N_{ij}, have to be recorded for the various transitions. The conditional probabilities are then estimated using:

$$p_{ij} = \hat{P}[X_j|X_i] = \frac{N_{ij}}{N_{i\cdot}} \tag{3.2-3}$$

where:

$$N_{i\cdot} = \sum_j N_{ij}$$

This statistic is an unbiased, consistent, maximum-likelihood estimator (M.L.E.), and for large samples is approximately distributed with a multivariate normal distribution. The sample variance for the estimator is distributed as:

$$\bar{S}^2_{p_{ij}} = \frac{1}{N_{i\cdot}} p_{ij}(1 - p_{ij}) \tag{3.2-4}$$

The large values for $N_{i.}$ required to estimate each row in the stochastic matrices cause the total number of samples, N, to be very large:

$$N = \sum_i N_{i.}, \quad \text{for } i = 1,2, \ldots, n$$

The maximum value, n, for i and j is determined by the page width and horizontal element dimension:

$$n = \frac{w_p}{d_h}, \quad \text{maximum run length}$$

For documents C1 and C2 using 8 mil elements:

$$n_{C1} = n_{C2} = \frac{8.5}{0.008} \approx 1062$$

For document C2 using 5 mil elements:

$$n_{C2} = \frac{8.5}{0.005} = 1700$$

The array sizes are n^2 and thus require on the order of 10^6 computer words for storage.

This strain on computational facilities was resolved by reducing the matrix sizes down to some value, K. States $i = 0$ and $j = 0$ were "runs" of zero length, representing the left and right margin, respectively. Runs of $(K-1)$ or larger were grouped together, while runs of $(K-2)$ or less were left as before. Inspection of independent distributions like Fig. 3.1-3, revealed that black runs do not exceed about 10, and 16, for documents C1, and C2 (5 mils) respectively. Therefore it was arbitrarily decided to set $K = 20$. (For simplicity, data for the boundary conditions $i,j = 0$ and $i,j = (K-1)$ have been deleted in the following discussions.)

The size of N was arrived at empirically, by increasing the sample size until the row distributions stabilized. This could be readily seen using computer plots such as Fig. 3.2-2. The size of N ultimately became the data from half of a standard 8½ × 11 inch page (the exact regions for each test document are indicated in Appendix A.2):

$$N = \frac{h_p w_p}{d_h d_v} \qquad \text{elements}$$

where:

$$h_p = \text{height scanned on a page}$$

For the documents C1, and C2 (5 mils), these sample sizes were:

$$N_{C1} = 730{,}281 \qquad \text{elements}$$

$$N_{C2} = 1{,}870{,}000 \qquad \text{elements}$$

The successive curves in Fig. 3.2-2 are estimated distributions for the 1st quadrant rows of Eq. 3.2-1.

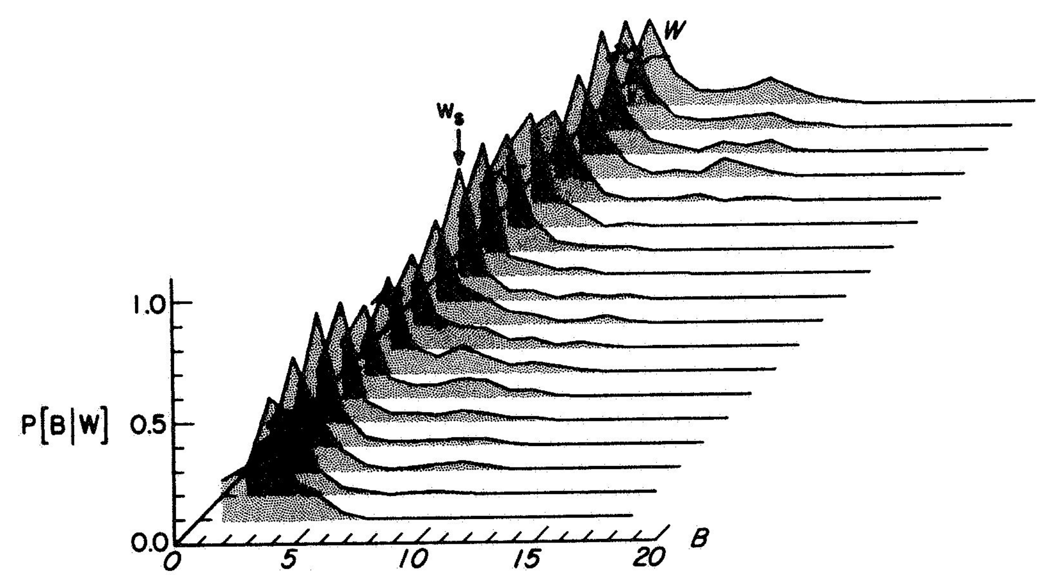


Fig. 3.2-2. CONDITIONAL PROBABILITY DISTRIBUTIONS FOR THE 1st QUADRANT MARKOV SUBMATRIX, P[B|W].

These discrete distributions were again represented as curves, by joining their values with straight lines. They each represent a distribution:

$$P[B|w_i] \, , \quad \text{for } i = 1,2, \ldots, 18$$

that must sum to one (including terms beyond the figure, out to $j = n$)

$$\sum_j P[b_j|w_i] = 1, \quad \text{for } i = 1,2, \ldots, n$$

and are dependent on some prior white run of length 1 to 18.

At first glance, the results were disappointing. There was almost no dependence to be seen, as evidenced by the similarity between planes. In other words, the distributions for black runs appeared almost independent of the preceding white run. These distributions for $P[B|W]$ just resembled the distribution for $P[B]$ in Fig. 3.1-3 (they both were derived from the same sample of document C1). The obvious salient feature was the stroke width, w_s (expressed in terms of consecutive black elements, B, rather than its dimension in mils).

On the other hand, the distribution displayed in Fig. 3.2-3 showed marked dependency.

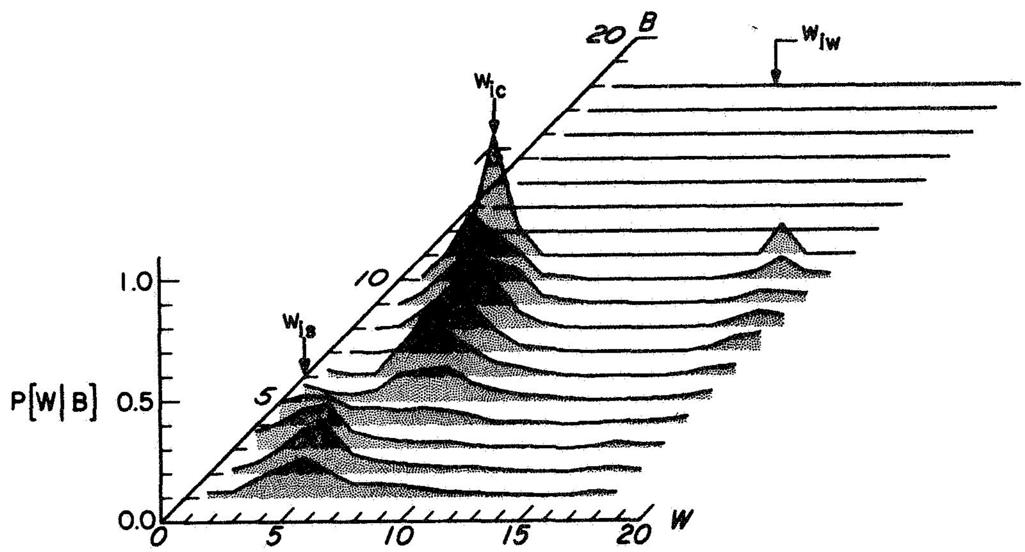


Fig. 3.2-3. CONDITIONAL PROBABILITY DISTRIBUTION FOR THE 3rd QUADRANT MARKOV SUBMATRIX, $P[W|B]$.

Here the rows are:

$$P[W|b_i], \quad \text{for} \quad i = 1,2, \ldots, 18$$

The inter-character width, w_{ic}, and inter-word width, w_{iw}, from Fig. 3.1-3 are easily recognized. Unlike the topography for w_s in the preceding figure, these regions are not parallel to the B axis. Instead they tend toward dependency with B of the form:

$$B + W = \text{constant} \tag{3.2-5}$$

Such variation is due to spatial quantization of the original characters. Whereas the width of a character may be standard, it will be digitized with varying combinations of B and W, depending on relative phase with the scan element. This tradeoff between B and W for an adjacent stroke and space is such that their sum should stay constant.

A feature heretofore lumped together with w_{ic} (in Fig. 3.1-3) is labeled w_{is}, and defined as:

$$w_{is} = \text{inter-stroke width}$$

Its characteristic was not apparent in the marginal distribution of whole runs, P[W], because w_{ic} and w_{is} overlap. One can visualize this overlap as viewing the topology in Fig. 3.2-3 while standing on the W axis. Notice that w_{is} should be an especially appropriate example of Eq. 3.2-5, being that it is postulated in terms of stroke and inter-stroke widths. This is borne out by the topology for w_{is}, which comes close to following a line with -45° slope in the B-W plane.

3.2.2. χ^2 test for dependence between runs

Encouraged by the apparent dependency in Fig. 3.2-3, my next step was to test this hypothesis statistically. The classical methods using likelihood ratios to test independence in contingency tables [Neyman and Pearson - 1933], [Wilks - 1935], have more recently been applied to the testing of finite Markov chains [Bartlett - 1951]. Among the problems considered are the estimation of transition probabilities, the testing of goodness of fit, and testing of the order of the chain. A good survey of statistical methods for Markov chains along with an extensive bibliography, can be found in [Billingsley - 1961].

In this application, my prime interest is to test the hypothesis that the chain is of a given order. It is recognized that results for the finite, discrete chain are only first approximations to such hypotheses for the original stochastic process. One long observation of the chain is made, under the assumption that the basic process is stationary and ergodic.

The likelihood ratio used to test the hypothesis that runs are independent (that the chain is of zero order) is:

$$\lambda = \frac{\left(\prod_i N_{i\cdot}^{\;N_{i\cdot}}\right)\left(\prod_j N_{\cdot j}^{\;N_{\cdot j}}\right)}{N^N\left(\prod_{i,j} N_{ij}^{\;N_{ij}}\right)} \qquad (3.2\text{-}6)$$

For large samples, this is distributed as a chi square:

$$\chi^2 = \sum_{i,j} \frac{[N_{ij} - (N_{i\cdot} N_{\cdot j}/N)]^2}{(N_{i\cdot} N_{\cdot j}/N)}, \quad \text{for} \quad i,j = 1,2, \ldots, k \qquad (3.2\text{-}7)$$

with $(k-1)^2$ degrees of freedom. It differs from $-2 \log \lambda$ by terms of the order:

$$\frac{1}{\sqrt{N}}$$

Since the model can be broken into two stochastic submatrices, each submatrix was tested individually. This was also done to isolate the possibility that the matrix $P[B|W]$ might be of zero order (a possible conclusion from Fig. 3.2-2). Document C1 was tested, using a large sample (over the area indicated in Appendix A.2) so that:

$$N[B|W] = 19{,}342$$

$$N[W|B] = 19{,}059$$

causing terms differing from -2 log λ of the order:

$$\frac{1}{\sqrt{N[B|W]}} = 0.0072$$

$$\frac{1}{\sqrt{N[W|B]}} = 0.0073$$

The computed values for the statistic were:

$$\chi^2[B|W] = 2589$$

$$\chi^2[W|B] = 4361$$

Although the value for matrix $P[B|W]$ was noticeably lower, they both exceeded α levels of 0.05 and 0.025 with ease. With $K = 20$ there were 381 degrees of freedom, such that [Hald - 1952]:

$$\chi^2_{0.050}(381) \geq 406$$

$$\chi^2_{0.025}(381) \geq 415$$

were the levels for the test.

The rows of the two matrices were also tested against the null hypothesis that they came from the same population as their marginal distribution. In other words if runs are truly independent, grouping them on the basis of the preceding run should not make them depart from a zero-order distribution. The latter has to be estimated too, of course; but coming from the larger sample, the estimated marginal distribution is considered adequate. The results of this test are summarized in Table 3.2-1.

Table 3.2-1

χ^2 DATA TO TEST ROW DISTRIBUTIONS

i	$N_{i.}$ [B|W]	$\chi^2_{i.}$ [B|W]	$N_{i.}$ [W|B]	$\chi^2_{i.}$ [W|B]
1	678.	443.	3301.	233.
2	1077.	345.	8524.	317
3	2045.	242.	2896.	721.
4	2924.	179.	1241.	446.
5	3055.	209.	807.	323.
6	1757.	321.	613.	264.
7	1050.	188.	592.	437.
8	798.	60.	441.	282.
9	649.	43.	362.	157.
10	411.	26.	150.	172.
11	303.	43.	9.	33.
12	229.	40.	0.	0.
13	140.	45.	0.	0.
14	58.	42.	0.	0.
15	111.	35.	0.	0.
16	308.	47.	0.	0.
17	431.	53.	0.	0.
18	352.	88.	0.	0.

Except for the empty rows in matrix [W|B] (beyond $W = 10$), all rows had a total frequency of at least:

$$N_{i.} \geq 58$$

such that:

$$\frac{1}{\sqrt{N_{i.}}} = 0.131$$

The α-level for (K-1)=19 degrees of freedom:

$$\chi^2_{0.05}(19) \geq 30.1$$

was only missed by the 10^{th} row of matrix P[B|W].
It had value:

$$\chi^2_{i.}[B|W = w_{10}] = 26$$

which would exceed the 0.2 α-level for 19 d.f. Perhaps its distribution too closely resembled the estimated marginal. Considering that only one out of the twenty-eight rows fell below the test level, this could also be attributed to Type II error.

Higher orders of dependency were not explored due to practical limits on computation. The evidence of first order dependence was sufficient to encourage further study on matching alphabets. Source alphabets based on this dependence produced noticeable increases in compression, as will be shown next.

3.2.3. Alphabets for the dependence between runs

Just as comparisons were made earlier between sources with alphabets [X], [X,Y], and [Y|X]; comparison may now be made between alphabets [W+B], [W+B,W+B], and [W+B|W+B]. In this case, source alphabets based on the marginal, joint, or conditional entropies of runs rather than elements are being considered.

Figures 3.2-4 and 5 illustrate joint distributions for stochastic submatrices of the matrix:

$$P[W+B, W+B] = \left[\begin{array}{c|c} 0 & P[B,W] \\ \hline P[W,B] & 0 \end{array}\right] \tag{3.2-8}$$

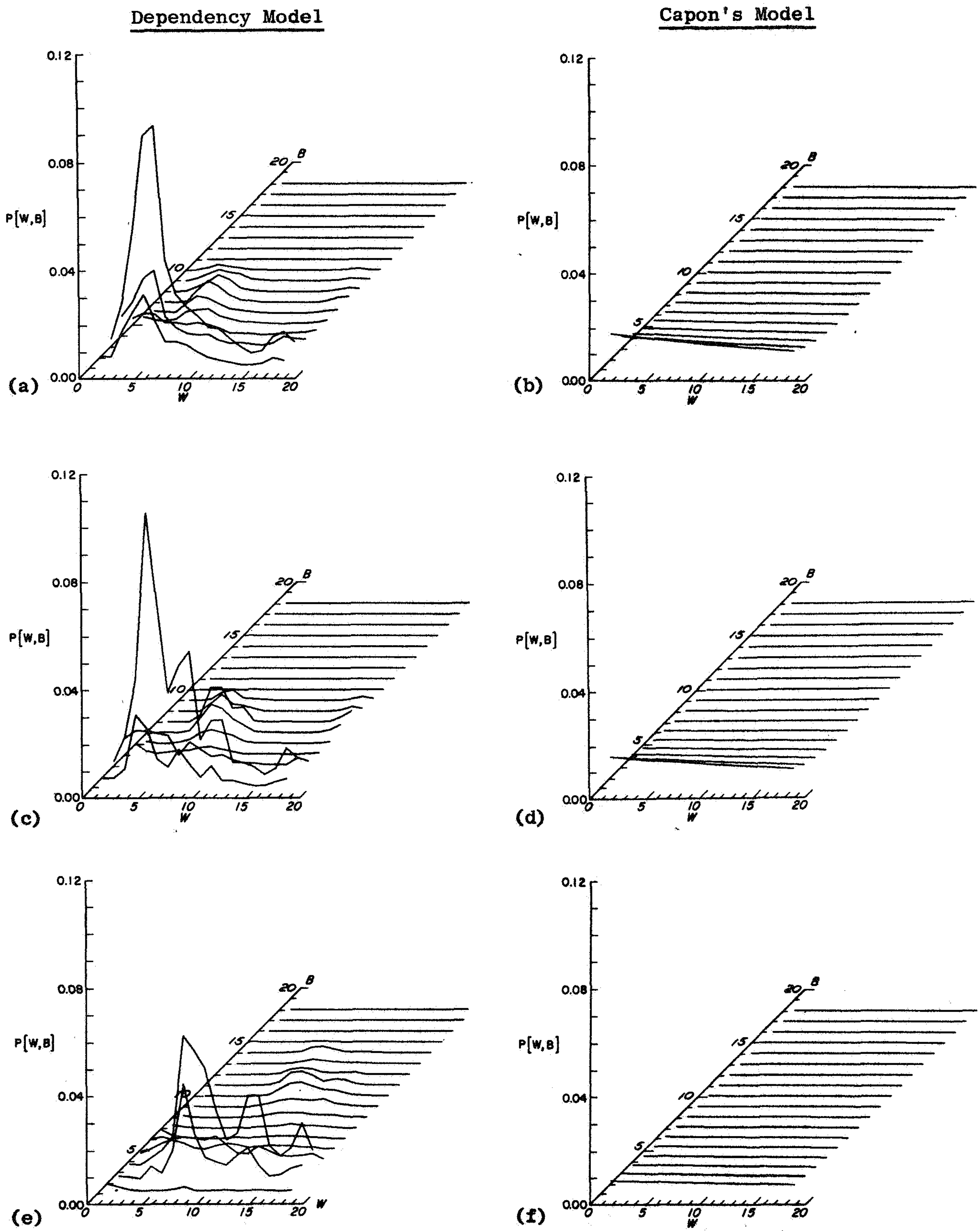


Fig. 3.2-4. JOINT PROBABILITY DISTRIBUTIONS FOR P[W,B].

where analogous to the definitions used with Eq. 3.2-1:

$$P[B,W] = \{P[b_i,w_j] \geq 0\}$$

$$P[W,B] = \{P[w_i,b_j] \geq 0\}$$

$$P[W,W] = \{P[w_i,w_j] \equiv 0\} \equiv 0$$

$$P[B,B] = \{P[b_i,b_j] \equiv 0\} \equiv 0$$

The left and right hand columns in both figures juxtapose estimated distributions with distributions derived using Capon's model (for the same raw data). The three rows in each figure are derived from the raw data of test documents C1, C2, and C2 (5 mils) respectively. Figure 3.2-4 enables comparison between their P[W,B] distributions, while Fig. 3.2-5 displays all the P[B,W] distributions. The approximate nature of a model based on 1st order dependence between elements rather than runs is apparent from the fit of Capon's model to the measured distribution.

In studying these figures, one should keep in mind that the top two rows are for documents scanned with a normalized resolution frequency of 1.25×1.25. The bottom row document has been scanned at 2.00×2.00. Comparing the bottom two rows which represent the same Dual-Gothic document scanned at different resolutions, one can see the same topography--just scaled differently.

The top two rows allow comparison at the same resolution between distributions for Pica and Dual-Gothic type. The resemblance between distributions indicates the existence of fundamental properties independent of the particular font being used. This is encouraging. A common distribution facilitates the synthesis of a mutually compatible compression code. One major purpose of this study was to explore whether the spatial frequency characteristics of different type fonts would not indeed be similar. This is somewhat analogous to mask-matching optical character recognition using a Van der Lugt filter in the spatial frequency domain [Goodman - 1968]. However, rather than depending on consistent frequency distributions for individual characters, one need only worry about

Dependency Model   Capon's Model

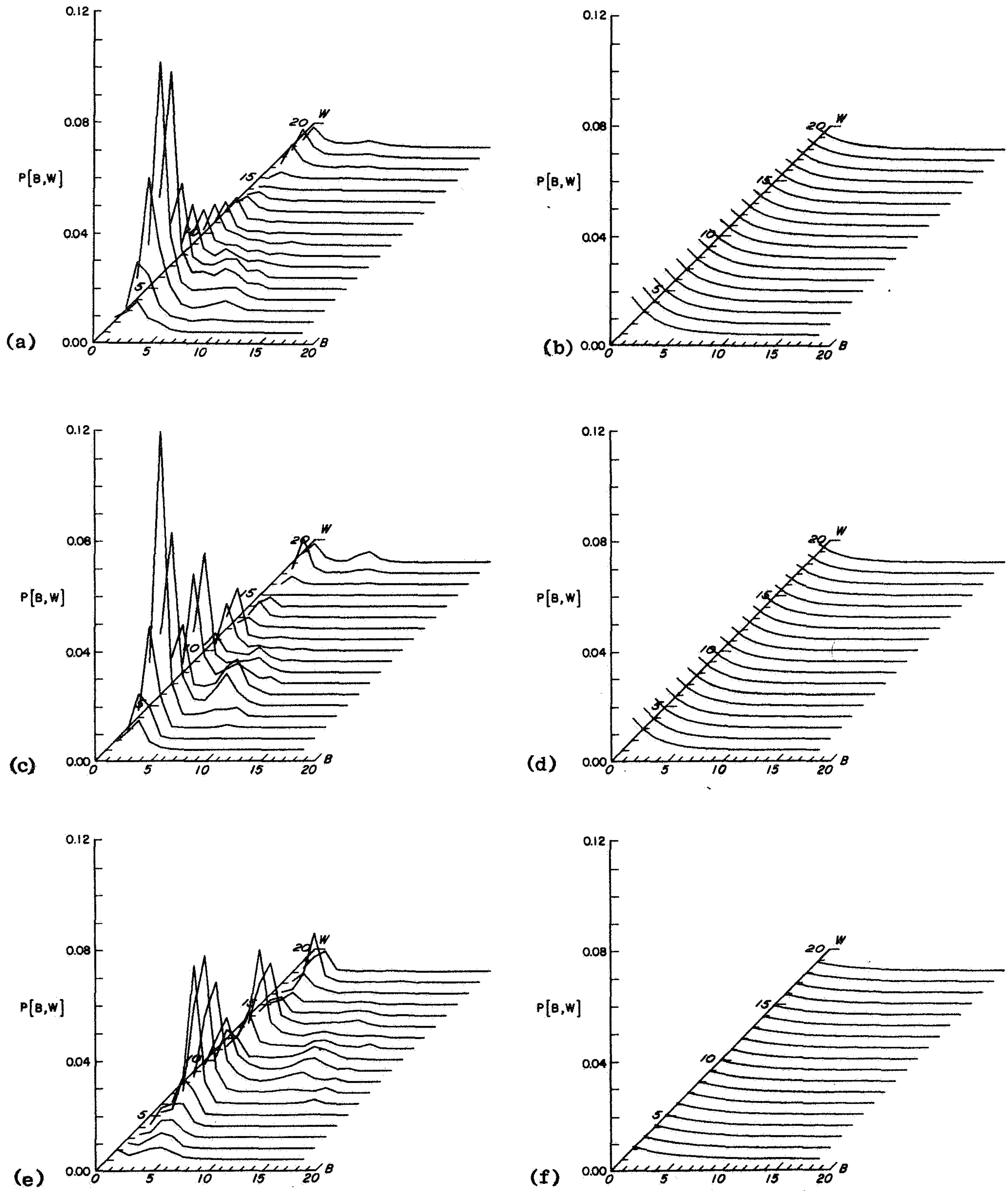


Fig. 3.2-5. JOINT PROBABILITY DISTRIBUTIONS FOR P[B,W]

consistency for the aggregate distribution of all characters. If the latter is consistent over a number of fonts, it alludes to the existence of a code matched for most printed matter (such as all standard size typewriting and printed text).

In section 3.1, an inequality derived from the grouping axiom [Ash - 1965] was illustrated when joint and conditional element alphabets were compared:

$$H[Y|X] \leq H[X,Y] \tag{3.2-9}$$

This is applicable here for run alphabets:

$$H[W+B|W+B] \leq H[W+B,W+B]$$

In order to evaluate conditional source alphabets, the estimates of conditional distributions have been extended to both sizes of Dual-Gothic type and plotted for comparison in Fig. 3.2-6. (The similarity between distributions even appears to have increased, an illusion probably due to smaller vertical scale.) A separate code dependent on the prior state, must be designed for each of the conditional submatrices $P[W|B]$ and $P[B|W]$. To capitalize on this resemblance, a common distribution can be defined as a linear combination of individual distributions:

$$P_0[X] = \sum_k a_k P_k[X] \ , \quad k = 1,2, \ldots, r \tag{3.2-10}$$

where:

- k, indicates the k^{th} distribution over X
- a_k, are coefficients that sum to unity
- X, is the source alphabet (here, a conditional row like: $[W|B=b_9]$)

The a_k are just weights, reflecting the a-priori incidence of the different distributions.

W Subalphabet　　　　　　B Subalphabet

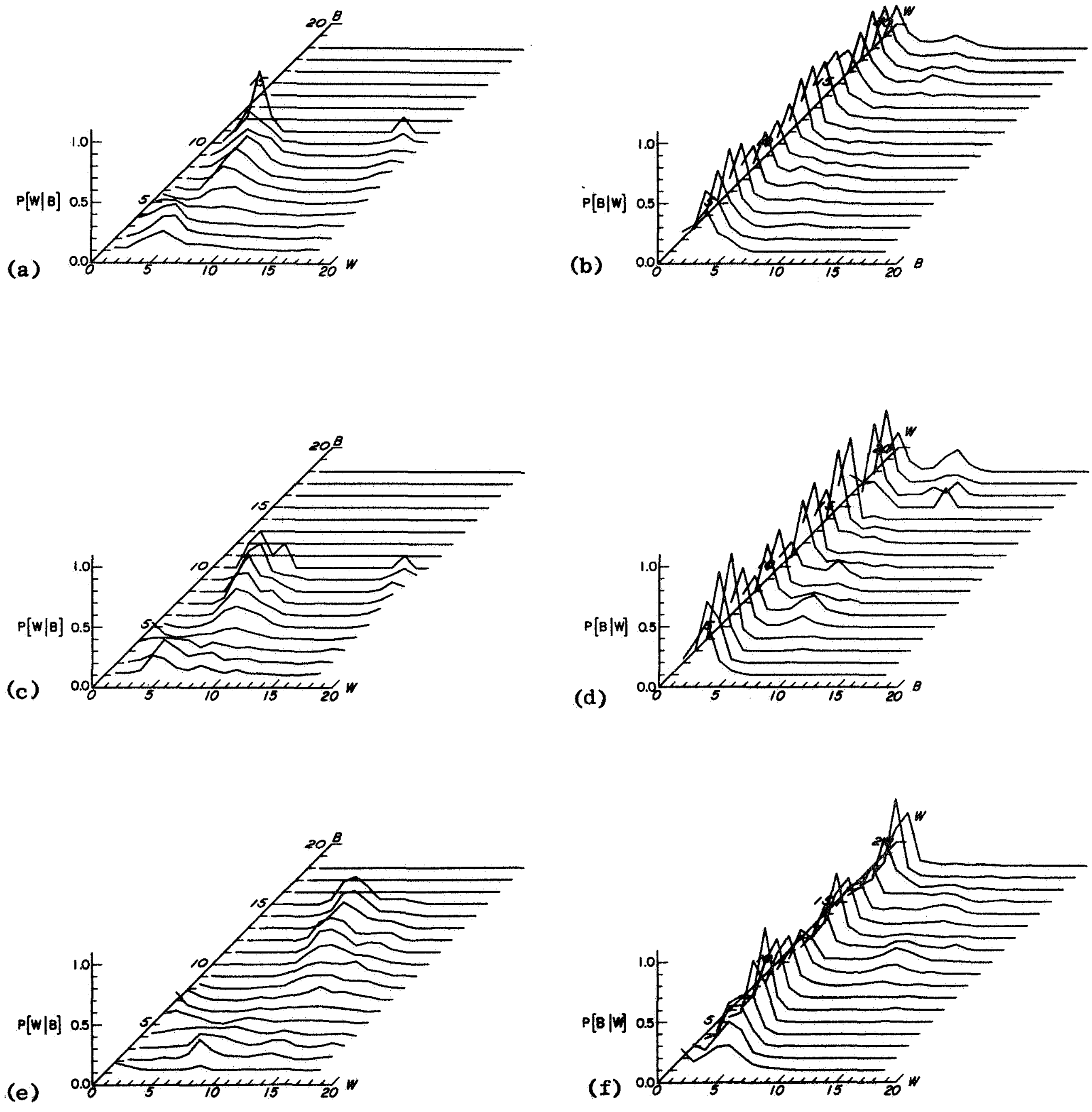


Fig. 3.2-6. MARKOV (CONDITIONAL) PROBABILITY DISTRIBUTIONS.

Since $H[X]$ is a convex function [Ash - 1965]:

$$H_0[X] \geq \sum_k a_k H_k[X] \qquad \text{bits/symbol} \qquad (3.2\text{-}11)$$

where, $H_0[X]$ is the entropy for $P_0[X]$. Physically, this means that an encoder based on $P_0[X]$, is inferior to a system that recognizes each incoming distribution and matches it with its own encoder. In the adaptive system individual encoders approach $H_k[X]$ bits per symbol, and their average rate would approach the right hand side of Eq. 3.2-11.

For a single encoder, only one distribution can be guaranteed an optimal code. To match a multiplicity of distributions, $P_0[X]$ should be best. To argue this, define:

$$Q_{0k}[X] = -\sum P_0[x_j] \log_2(P_k[X_j]) \qquad \text{bits/symbol} \qquad (3.2\text{-}12a)$$

where for $P_k[X] = P_0[X]$, by definition:

$$Q_{00}[x] = H_0[X] \qquad (3.2\text{-}12b)$$

Theorem 3-1: $Q_{0k}[X]$ is a convex function.
In equation form:

$$Q_{00}[X] \leq \sum_k a_k Q_{0k}[X] \qquad (3.2\text{-}13)$$

with equality iff $P_k[X] = P_0[X]$

Proof:

From Lemma 14.1 of [Ash - 1965]:

$$H_0[X] \leq Q_{0k}[X] \qquad (3.2\text{-}14)$$

with equality iff $P_c[X] = P_0[X]$.

Case A. All $P_k[X] \neq P_0[X]$:

$$H_0[X] < Q_{0k}[X] \tag{3.2-14a}$$

Taking the weighted positive sum of both sides of Eq. 3.2-14:

$$\sum_k a_k \; H_0[X] < \sum_k a_k \; Q_{0k}[X]$$

and since $H_0[X]$ is independent of k:

$$H_0[X] < \sum_k a_k \; Q_{0k}[X] \tag{3.2-13a}$$

Case B. All $P_k[X] = P_0[X]$:

$$H_0[X] = Q_{0k}[X] \tag{3.2-14b}$$

Taking the weighted positive sum of both sides of Eq. 3.2-14b:

$$\sum_k a_k \; H_0[X] = \sum_k a_k \; Q_{0k}[X]$$

and again using $H_0[X]$ independent of k:

$$H_0[X] = \sum a_k \; Q_{0k}[X] \tag{3.2-13b}$$

Combining Eq. 3.2-13a and b yields:

$$H_0[X] \leq \sum_k a_k \; Q_{0k}[X] \tag{3.2-15}$$

with equality iff all $P_k[X] = P_0[X]$.

Apply Eq. 3.2-12b to Eq. 3.2-15, and the theorem is complete:

$$Q_{00}[X] \leq \sum_k a_k \, Q_{0k}[X]$$

with equality iff all $P_k[X] = P_0[X]$ Q.E.D.

The single-encoder-multiple-distribution problem should have a code based on $P_0[X]$. After all, this is the resultant distribution the encoder will see. If another code is used, for example matched to a specific input distribution $P_k[X] \neq P_0[X]$, $Q_{0k}[X]$ for the overall distribution $P_0[X]$ will not be minimum. Thus, writing Theorem 3-1 in terms of expectations:

$$E_0\{-\log_2(P_0[x_j])\} < E_0\{-\log_2(P_k[x_j])\} \tag{3.2-16}$$

(where $E_0\{\ \}$ represents an expected value over distribution $P_0[X]$). For optimal coding [Ash - 1965] code word lengths, $\lambda_k[x_j]$, satisfy the inequality:

$$-\log_2(P_k[x_j]) \leq \lambda_k[x_j] < -\log_2(P_k[x_j]) + 1$$

From this the expectation of code word lengths must approximately follow:

$$E_0\{\lambda_0[x_j]\} < E_0\{\lambda_k[x_j]\} \tag{3.2-17}$$

(This certainly holds when the inequality in Eq. 3.2-16 exceeds unity.)

Conditional source entropies for documents C1, and C2 have been tabulated in Table 3.2-3 for comparison with the previous source alphabets in Table 3.1-1. Computations have been made both for K = 20, and K = 40. (Beyond K, since run dependency was lumped together, all runs were treated as equiprobable. This only causes entropy estimates to err on the high side. As K increased, appropriate additional decrease in entropy was observed.) Further exploration into one-dimensional coding was discontinued, when the following results for two-dimensional coding became evident.

Table 3.2-3

COMPRESSION FOR SOURCE ALPHABETS
(ONE-DIMENSIONAL RUNS)

Source Alphabet	Document C1 (Pica)	Document C2 (Dual-Gothic)	Document C2 (Dual-Gothic)
$[W+B \mid W+B]_{K=20}$	4.260 elements/bit	4.901 elements/bit	6.578 elements/bit
$[W+B \mid W+B]_{K=40}$	4.419 " "	5.218 " "	7.153 " "
Blackness	9.5%	8.0%	7.2%
Resolution (absolute)	8 × 8 mil	8 × 8 mil	5 × 5 mil
Resolution (normalized)	1.62 × 1.62	1.25 × 1.25	2.00 × 2.00

3.3. Source Alphabets in Two Dimensions

The simple source alphabets used for line-by-line scanning are readily extended to two dimensions. Alphabets with higher than 1^{st} order Markov dependency are of interest, since a two-dimensional pattern has many neighbors. Table 3.3-1 shows results for source alphabet patterns ("X"s) some of which are dependent on nearby neighbors ("0"s).

Table 3.3-1

COMPRESSION FOR SOURCE ALPHABETS

(TWO-DIMENSIONAL)

Source Alphabet	Document C2 (Dual-Gothic)	Document C2 (Dual-Gothic)
$\begin{bmatrix} X \end{bmatrix}$	2.50	2.69
$\begin{bmatrix} & X & X \\ X & & \\ X & & \end{bmatrix}$	2.32	4.25
$\begin{bmatrix} 0 & 0 & 0 \\ 0 & X & \\ & & \end{bmatrix}$	5.80	10.80
$\begin{bmatrix} & 0 & 0 \\ 0 & X & X \\ 0 & X & X \end{bmatrix}$	6.57	11.15
$\begin{bmatrix} & 0 & \\ 0 & X & 0 \\ & 0 & \end{bmatrix}$	10.77	23.80
Blackness	8.0%	7.2%
Resolution (absolute)	8 × 8 mil	5 × 5 mil
Resolution (normalized)	1.25 × 1.25	2.00 × 2.00

The pattern in the fifth row of Table 3.3-1 achieves the best results. However, dependency has been utilized from all four sides. For practical scanning schemes, this is equivalent to non-causal prediction (using elements received at some future time). As a result, the other patterns are restricted to dependency on elements that would already be received in line-by-line scanning.

In the fourth row, a source alphabet with four elements is shown with dependency based on four preceding elements. This pattern could be applied with conditional Huffman coding [Huffman - 1952], one code for each of the 16 prior states.

In the third row, 4^{th} order Markov dependence is used in a causal scheme. Patterns of this type have been explored for two-level data by [Wholey - 1961]. Notice that conditional Huffman coding for a one-bit alphabet is impractical. However, predictive coding [Elias - 1955] has been shown to achieve average message lengths approaching the conditional entropy for such a source.

The second row is just an extra pattern, for which data was readily available in measuring for the fourth row. Finally, the first row is just the basic single element entropy, provided for future comparison.

Comparing columns illustrates how an increase in resolution increases compression for the same pattern. The rates of increase tend to improve for the better patterns too. This resolution dependency is explored next.

3.4. The Influence of Resolution on Source Alphabets

The use of compression to judge encoding schemes suffers from a common problem. The results of a particular measurement are hard to generalize. Moreover, the same coding scheme applied to the same image will indicate greater compression as the resolution is increased. (This property can give rise to glowing reports and hinders objective evaluation). This section attempts not only to display variations in compression, but to model them for a large body of present-day documents. Namely, the images of printed matter, found so predominantly in the commercial image-processing market.

3.4.1. Entropy and compression variations with spatial frequency

The simple scan patterns in two-dimension presented in section 3.3, were next checked for their variation with frequency. In doing so, another sample of Dual-Gothic type was used. Rather than the random sample of print in document C2, documents B1-B5 from the legibility study were used. This was done to display illegibility (straight dashed line) and entropy (solid curves) from the same sample, in order to observe the tradeoff for various patterns (Fig. 3.4-1).

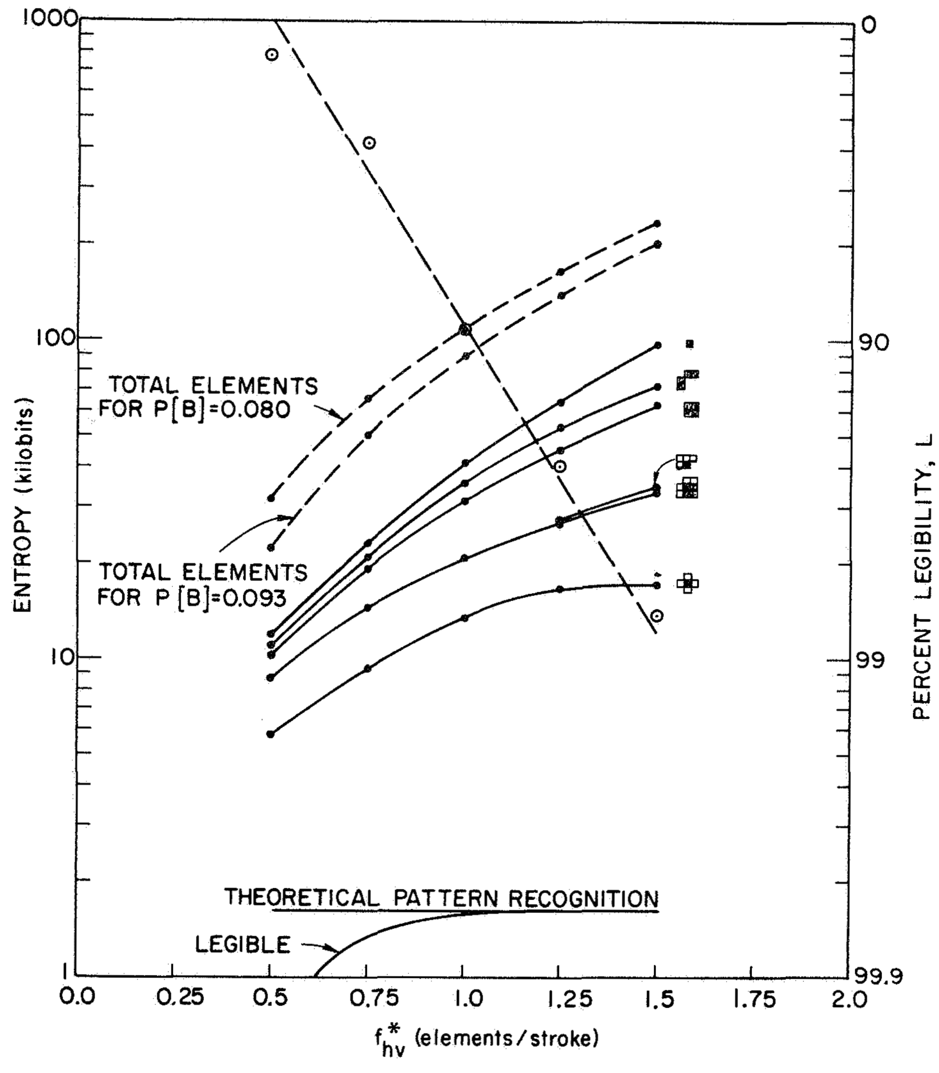


Fig. 3.4-1. ENTROPY VS RESOLUTION FOR TWO-DIMENSIONAL SCAN PATTERNS.

The set of scan patterns are like successive attempts to grasp the true information in the page. They can only approach what a person or pattern recognition system can do. The latter is represented by a flat curve at the base of the figure, and is based on an assumption of 4.03 bits per alphanumeric character [Abramson - 1963]. A few more bits per page were added to this pattern recognition model to account for character addressing, character size, and the type font. However, their effect turned out to be insignificant. An entropy curve based on pattern recognition is also displayed taking legibility into consideration. The two pattern recognition curves tend to overlap at normalized frequencies greater than one element per stroke.

At the other end of the scale, the total elements in the sample, assuming maximum uncertainty, were plotted as the dashed curve marked with P[B] = 0.093. This represents an upper bound for the entropy in the sample at one bit per element:

$$NH_{max}[X] = N \qquad \text{bits/page} \qquad (3.4\text{-}1)$$

It is interesting how the ineffectual scan patterns tend to follow the upper bound based on totally random page elements. Conversely, the effective scan patterns are not only better, but tend to be flat with increasing resolution. Apparently they are "getting at," or "recognizing" some of the information in the page; and after a point, increased resolution has little to contribute (as should be the case for pattern recognition).

For these effective patterns it seems that some form of "recognition" is taking place; perhaps recognition of some fundamental feature in characters, like their strokes. This is what the run-length patterns in one-dimension tended to do. They captured the intervals that make up the structure of a character, and seemed less sensitive to the particular font in use.

This stroke recognition by good two-dimensional patterns, is borne out by the correlation of results in Fig. 3.4-1 with measurements taken from document C2. The characters in document C2 are distributed differently but this seems to have little effect on the performance of the good patterns. They apparently get their compression from features of characters rather than the total character (as will be seen in Fig. 3.4-2).

To reconcile the data from documents B1-B5 and C2, it was necessary to make allowance for the percentage of black vs white elements in the sample (as will be seen in section 3.4.2, this percentage is of fundamental significance). It was discovered that the patterns predict all-white so well that addition of white elements has little effect on the page entropy. This can be explained using the definition of conditional entropy (Eq. 3.106 generalized):

$$H[Y|X] = \sum_i P[x_i]\, H[Y|x_i] \tag{3.4-2}$$

For the term, $H[Y|x_w]$, in which prior state, x_w, is all-white, the entropy is very low. That is, predicting the following state, Y, to be all-white has a high probability of success, and:

$$H[Y|x_w] \approx 0 \tag{3.4-3}$$

As a result, additions to the number of white page elements only affect the probability weights in Eq. 3.4-2, which have been estimated using:

$$\hat{P}[x_i] = \frac{N_i}{N}$$

For the addition of some more white elements, N_w, the new total becomes:

$$N' = N + N_w$$

and for only the all-white state, x_w, does the numerator also change:

$$\hat{P}[x_w] = \frac{N_i + N_w}{N + N_w}$$

This change in $\hat{P}[x_w]$ has little effect in Eq. 3.4-2, when multiplied with its corresponding negligible entropy (Eq. 3.4-3). For the rest:

$$P[x_i] = \frac{N_i}{N + N_w}, \quad \text{for} \quad x_i \neq x_w$$

producing little overall effect for small changes in total N:

$$N_w \ll N$$

As a result, the data in Fig. 3.4-1 can be adjusted to 8.0 percent white by simply adding more white elements, resulting in a new dashed curve for total page elements:

$$N'H_{max} = N' \qquad \text{bits/page} \qquad (3.4\text{-}4)$$

Compression, defined previously for individual symbols from a source alphabet (Eq. 3.1-7), can now be defined for a page

$$C = \frac{N}{NH} \qquad \text{symbols/bit} \qquad (3.4\text{-}5)$$

In Fig. 3.4-1, this is simply the ratio between any curve for a scan pattern and the dashed line representing the number of page elements (for a given percentage of black, P[B]).

When percent black is accounted for, the compression measurements for documents B1-B5 and C2 tend to correspond. A measure of performance independent of the black-white percentages can be defined by dividing pattern compression by single element compression:

$$C_R = \frac{C}{C[X]} = \text{relative compression} \qquad (3.4\text{-}6)$$

where the single element compression is:

$$C[X] = \frac{N}{H[X]} = \frac{N}{-\ P[B]\ \log_2 P[B] + P[w]\ \log_2 P[w]} \qquad (3.4\text{-}7)$$

For example, the relative compression for document C2 with 8.0 percent black, using the scan pattern in the fifth row of Table 3.3-1, is:

$$C_R = \frac{10.77}{2.50} = 4.31$$

The values for document C2 in Table 3.3-1 are plotted with "X"s in Fig. 3.4-2, whereas measurements for documents B1-B5 are plotted with dots. For comparison, relative compression for the run-length schemes of Table 3.2-3 have also been included, (labeled with the amount of run dependence, K = 20 or K = 40, that they utilized). The idealized curves for pattern recognition of page information have also been included.

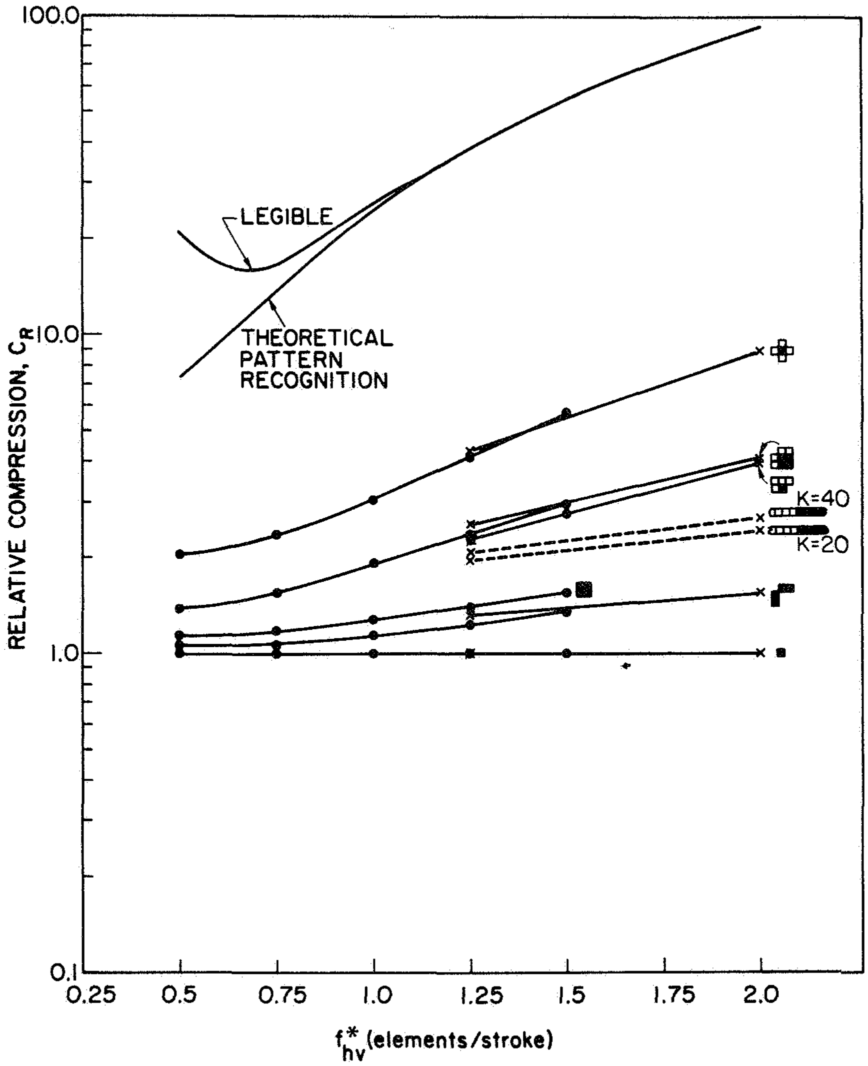


Fig. 3.4-2. RELATIVE COMPRESSION VS RESOLUTION FOR TWO-DIMENSIONAL SCAN PATTERNS.

3.4.2. A model for the entropy of simple scan patterns

While reflecting on the curves in Fig. 3.4.2, one should keep in mind that the entropy measurements are only crude approximations to the basic information in the underlying stochastic process. An image of printed matter contains a statistical distribution of black and white that can be looked upon with different resolutions and with different patterns, but the fundamental process is the same. In Fig. 3.4-1 the curve for theoretical pattern recognition stays constant with frequency (even at a resolution where legibility is low, the original image has its information--the system simply isn't receiving it).

[Vitushkin - 1961] has rigorously defined the information capacity of an image dissection scheme in terms of ϵ-entropy, or resolution-dependent entropy. For a uniformly dissected page, the maximum uncertainty for the number of binary elements used, is essentially the ϵ-entropy for the binary process. This will be recognized as the dashed curves in Fig. 3.4-1 which represent the total number of elements for different sized pages.

[McLachlan - 1958] attacked this question in terms of the total number of combinations, 2^N, that N bits can assume. (Notice that this is just the antilog_2 of Vitushkin's definition). If this combinatorial approach is applied to a page, known a-priori to have a fixed percentage of black elements, an upper information bound can be obtained. Assuming the number of black elements to be, M, then the page entropy for any pattern should obey the inequality:

$$NH_{P[B]} \leq \text{antilog}_2 \left[\frac{N!}{(N-M)!M!}\right] \tag{3.4-8}$$

using Sterling's inequality, for large N:

$$N! \approx C^{-N} N^{(N-\frac{1}{2})} (2\pi)^{\frac{1}{2}} \tag{3.4-9}$$

Eq. 3.4-8 becomes:

$$NH_{P[B]} \leq \text{antilog}_2 \left[\frac{N^N}{(N-M)^{N-M} M^M (2\pi)^{\frac{1}{2}}}\right]$$

where the powers of $\frac{1}{2}$ have been neglected. Evaluating the antilog_2 and neglecting the term $\log_2(2\ )^{\frac{1}{2}} \approx 0.8$ yields:

$$NH_{P[B]} \leq N\left[-\left(\frac{N-M}{N}\right)\log_2\left(\frac{N-M}{N}\right) - \left(\frac{M}{N}\right)\log_2\left(\frac{M}{N}\right)\right]$$

which is recognized as the single element entropy for a page:

$$NH_{P[B]} \leq NH_{P[B]}[X], \quad \text{for} \quad \hat{P}[B] = \frac{M}{N} \tag{3.4-10}$$

It is the compression for this single element that has been used to define the relative compression used in Fig. 3.4-2. Referring back to Fig. 3.4-1, the curve for single-element entropy does bound all other measurements. Furthermore, it corresponds closely to the shape of nearby curves that seem to have little grasp of the information inherent in the page. Conversely, as patterns improve, they flatten out with frequency and tend to imitate the pattern recognition curve.

To model the intervening curves, a linear combination of these two seems appropriate. Thus:

$$NH_{P[B]} = \alpha_1 NH_{P[B]}[X] + \alpha_0 H[PR] \tag{3.4-11}$$

where:

α_0, α_1, are the linear coefficients

and:

H[PR], is the theoretical entropy with pattern recognition

A least-squares fit to this model was performed for curves from Fig. 3.4-1, and uses the linear regression program of section 2.2. The results are summarized in Table 3.4-1.

A second fit was obtained, assuming that

$$\alpha_0 = 1 - \alpha_1 \tag{3.4-12}$$

Then the regression becomes:

$$N(H_{P[B]} - H[PR]) = \alpha_1 (NH_{P[B]}[X] - H[PR]) \tag{3.4-13}$$

and the equation is constrained to have an intercept of zero. The resulting values are also shown in Table 3.4-1, along with the standard deviation from the regression line, $\overline{S}_T$, and coefficient of determination, $\overline{R}^2_T$. Here, T, stands for the total page entropy:

$$T_{P[B]} = NH_{P[B]} \tag{3.4-14}$$

Table 3.4-1

ENTROPY MODEL COEFFICIENTS

Scan Pattern	N	$\alpha_0 H[PR]$	α_1	α_0	$\overline{R}^2_T$	$\overline{S}_T$
$\begin{bmatrix} & X & X \\ X & & \\ X & & \end{bmatrix}$	5	4,943 ± 2,040	0.706 ± 0.036	3.007	0.992	2,493
		0 ± 1,823	0.773 ± 0.033	0.227	0.994	4,077
$\begin{bmatrix} & & \\ & X & X \\ & X & X \end{bmatrix}$	5	5,206 ± 1,659	0.605 ± 0.029	3.167	0.993	2,028
		0 ± 1,731	0.674 ± 0.031	0.326	0.994	3,870
$\begin{bmatrix} & 0 & 0 \\ 0 & X & X \\ 0 & X & X \end{bmatrix}$	5	7,627 ± 1,781	0.278 ± 0.031	4.639	0.963	2,177
		0 ± 2,299	0.375 ± 0.042	0.625	0.964	5,141
$\begin{bmatrix} & 0 & \\ 0 & X & 0 \\ & 0 & \end{bmatrix}$	5	6,415 ± 1,624	0.124 ± 0.029	3.902	0.860	1,986
		0 ± 1,838	0.198 ± 0.033	0.902	0.922	4,109

In addition to resolution, print size, and page density, the model needs to include page area:

$$A = Nd^2_{hv} \qquad \text{mils}^2 \tag{3.4-15}$$

rather than N. Inserting Eq. 3.4-12, 14, and 15 into Eq. 3.4-11, the model becomes:

$$T_{P[B]} = \alpha_1\left(\frac{A}{d^2_{hv}}\right) H_{P[B]}[X] + (1-\alpha_1)\ H[PR] \tag{3.4-16}$$

H[PR] is actually based on the number of characters:

$$H[PR] = 4.03\ N_c \qquad (3.4\text{-}17)$$

where:

N_c = the number of characters

However, N_c should be a function of character density, P[B], the page area, A, and the area of a character, a_c:

$$N_c = \frac{P[B]\ A}{a_c} \qquad (3.4\text{-}18)$$

The character area is proportional to the size of its stroke squared:

$$a_c = K_c\ w_s^2 \qquad (3.4\text{-}19)$$

where:

K_c, is a proportionality constant

This constant has been measured for Dual-Gothic type to be:

$$K_c = 21.5$$

Combining the last four equations produces the final form:

$$T_{[PB]} = \alpha_1 \left(\frac{A}{d_{hv}^2}\right) H_{P[B]}[X] + (1 - \alpha_1) \left(\frac{4.03}{K_c}\right) \frac{P[B]\ A}{w_s^2} \text{ bits} \qquad (3.4\text{-}20)$$

The model is based on resolution, character size, page density and page area:

$$T_{P[B]} \approx T(d_{hv}, w_s, P[B], A, \alpha_1, K_c) \qquad \text{bits}$$

along with constants α_1 and K_c for the scan pattern and type font.

For α_1 large, the first term predominates and the scan patterns follow the curve for single-element entropy. For α_1 small, the second term predominates and the entropy tends to stay constant with frequency.

As is apparent from Fig. 3.4-1, the total page entropy increases with resolution although accompanied by increasing compression. A better measure than just compression is needed to express the net results for a given scan pattern, especially when using different resolutions. Let the values for compression, entropy, and other parameters be measured at the Nyquist interval for a stroke, $f^*_{hv} = 1$, and these values used as a standard for comparison. All departure from conditions at $f^*_{hv} = 1$ can then be embodied in terms of resolution efficiency, which is developed next.

For printed matter, frequency variations are better expressed in normalized form to account for the scaling effect of character size. This can be applied to Vistushin's concept of ϵ-entropy using instead:

$$\epsilon^* = \frac{\epsilon}{w_s} = \frac{d_{hv}}{w_s} = \frac{1}{f^*_{hv}}$$

By also restricting the stochastic process to be binominally distributed with parameter P[B] and using a given scan pattern, the ϵ^*-entropy is then:

$$T_{P[B]}(\epsilon^*) = \epsilon^*\text{-page entropy} \qquad \text{bits/page} \qquad (3.4\text{-}21)$$

For a single-element scan pattern, [X], and P[B] = 0.5, Eq. 3.4-21 degenerates to the basic ϵ^*-entropy for the binary stochastic process (Eq. 3.4-1):

$$[T_{0.5}(\epsilon^*)]_{[X]} = N(\epsilon^*) \qquad \text{bits/page}$$

This maximum page entropy, when relative to ϵ^*-entropy at the Nyquist interval for strokes, defines a coefficient of expansion:

$$\gamma_E(\epsilon^*) = \left[\frac{T_{0.5}(\epsilon^*)}{T_{0.5}(1)}\right]_{[X]} = \text{coeff. of page expansion}$$

$$= \frac{N(\epsilon^*)}{N(1)} = \frac{w_s^2}{d_{hv}^2} = \frac{1}{(\epsilon^*)^2}$$

The entropy for one symbol from the alphabet (averaged over the page) becomes:

$$H_{P[B]}(\epsilon^*) = \epsilon^*\text{-symbol entropy} \quad \text{bits/symbol} \qquad (3.4\text{-}22)$$

with corresponding compression:

$$C_{P[B]}(\epsilon^*) = \frac{1}{H_{P[B]}(\epsilon^*)} = \epsilon^*\text{-symbol compression} \quad \text{symbols/bit} \qquad (3.4\text{-}23)$$

A coefficient of compression is just defined by dividing through by the compression at $\epsilon = w_s$ (i.e., $f^*_{hv} = 1$):

$$\gamma_C(\epsilon^*) = \frac{C_{P[B]}(\epsilon^*)}{C_{P[B]}(1)} = \text{coeff. of symbol compression}$$

For the parameters related to a specific scan pattern, a measure embodying all the changes with resolution can now be defined as:

$$\eta(\epsilon^*) = \frac{\gamma_C(\epsilon^*)}{\gamma_E(\epsilon^*)} = \text{resolution efficiency} \qquad (3.4\text{-}24)$$

Notice the boundary condition at $\epsilon^* = 1$ that:

$$\eta(1) = 1$$

By using the value for ϵ^*-symbol entropy at $\epsilon^* = 1$ and Eq. 3.4-24, the following identities can now be used to define the basic resolution-dependent parameters:

$$H_{P[B]}(\epsilon^*) = H_{P[B]}(1)\left[\frac{(\epsilon^*)^2}{\eta(\epsilon^*)}\right] \quad \text{bits} \qquad (3.4\text{-}25)$$

$$C_{P[B]}(\epsilon^*) = \frac{1}{H_{P[B]}(1)}\left[\frac{\eta(\epsilon^*)}{(\epsilon^*)^2}\right] \qquad (3.4\text{-}26)$$

$$T_{P[B]}(\epsilon^*) = \left(\frac{A}{w_s^2}\right)\frac{H_{P[B]}(1)}{\eta(\epsilon^*)}$$

$$= T_{P[B]}(1)\ \eta(\epsilon^*) \quad \text{bits} \qquad (3.4\text{-}27)$$

The efficiency characteristic will fall off sharply for poor patterns. Conversely, for good patterns it will fall slowly; and for perfect pattern recognition:

$$\eta_{PR}(\epsilon^*) = 1$$

To illustrate this, consider the following example. At a resolution of $f^*_{hv} = 2.0$, the 4th order Markov pattern (3rd row, Table 3.3-1) has compression:

$$C = 10.80$$

At a resolution of $f^*_{hv} = 1.25$ The non-causal pattern (5th row, Table 3.3-1) has compression:

$$C = 10.77$$

If compression values alone were reported, one would perhaps assume the patterns to be equivalent (or the 4th order Markov pattern even to be better!). For a fair comparison of the patterns, the same resolution should be used. Thus, for $f_{hv}^* = 1.25$, the 4th order Markov pattern yields:

$$C = 5.80$$

which is almost half the compression for the non-causal pattern.

Furthermore, at $f_{hv}^* = 1$, the ϵ^*-efficiency is unity for all patterns. But at $f_{hv}^* = 1.25$, for the 4^{th}-order Markov pattern:

$$\eta(\epsilon^*) = 0.78$$

while for the non-causal pattern:

$$\eta(\epsilon^*) = 0.85$$

Notice how efficiency for the better pattern drops more slowly as normalized resolution increases from the value $f_{hv}^* = 1$. This is the measure of a pattern's performance. As Eq. 3.4-26 illustrates, the greater compression accompanying high resolution is discounted by $\eta(\epsilon^*)$ for comparison with the performance at $\epsilon^* = 1$.

In Fig. 3.4-1, the two measures of page information--legibility and entropy--are compared for a particular font and "scan" pattern. Actually, illegibility is plotted to emphasize the tradeoff encountered in a system design. If cost functions with resolution are available for both parameters, their cost-weighted superposition can now be computed. Minimizing this total cost function will yield the optimal operating resolution for the system.

Chapter IV

POSSIBLE EXTENSIONS

Areas for further study may be found in the process of sampling and reconstructing digital images, in the measurement of document quality, in devising scan patterns with lower entropy, and in the implementation of codes for these scan patterns.

The effect of apertures shaped other than rectangular, has not been evaluated for its impact on printed matter. Black/white decision levels other than 50 percent of area integration could be explored, along with the feedback necessary to match 0 and 100 percent levels to the black and white on incoming documents. Quantizing two-level input data with multiple levels (grey-scale) may be explored for improving quality.

Objective quality measures are needed to evaluate grey-scale images. The desirability of having visible stair-steps in rectangularly reconstructed images, needs to be evaluated also. Round print elements may be an improvement due to the absence of sharp corners.

The possibilities for scan patterns have by no means been exhausted. Two-dimensional equivalents to "runs" need to be investigated. The literature in pattern recognition should be monitored for results that may carry over to image compression. Another related area is in optical image processing.

Clever implementations may be found for encoding good source alphabets, or good coding schemes may specify source alphabets to be tested. Adaptive coding schemes may also be investigated for simplicity, good ϵ-compression, and high resolution efficiency.

In general, searching through fundamental problems related to future image processing systems may turn up rewarding areas for further study.

APPENDIX

A.1. Experimental System Description

The block diagram in Fig. A.1-1 summarizes the author's experimental system, used to simulate the processing of digital images.

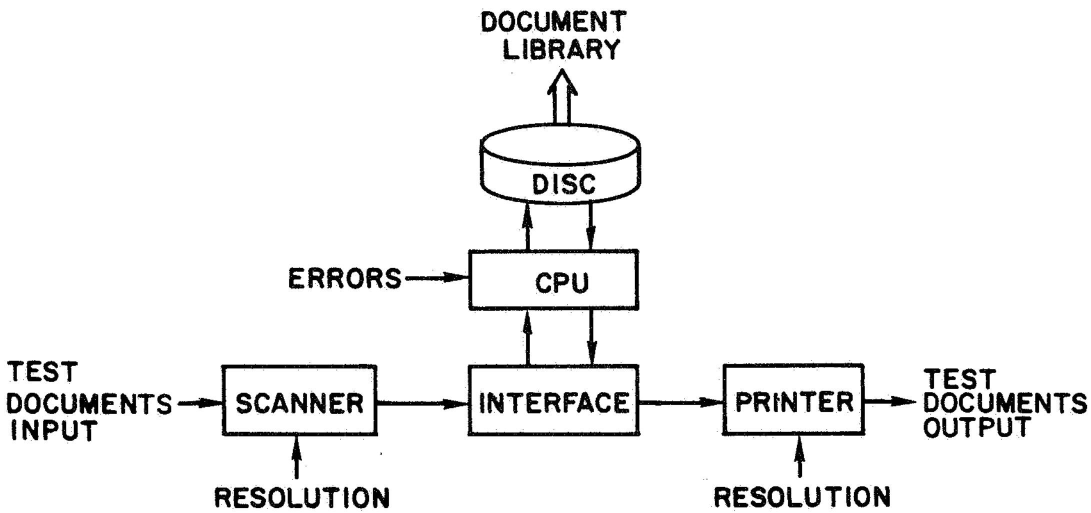


Fig. A.1-1. EXPERIMENTAL IMAGE-PROCESSING SYSTEM

Input-output is through a facsimile scanner/printer, modified to permit two-dimensional variations in resolution. An IBM 1620 is attached through interface electronics, to do the simulation and to permit image storage on 1311 disc packs. These digital images can then be transferred to tape and preserved in a document library for future study. When simulation of compression coding, transmission errors, and decoding is complete, images are put out on the printer for evaluation of their quality.

The mechanical scanner/printer can be seen in Fig. A.1-2. On the left side, a test document is mounted on a drum under scanning optics. As the drum rotates, the carriage with optics moves slowly to the right due to the "pitch" of a lead screw. This combined motion produces line-by-line scanning (which is vertical for a document oriented as shown).

On the right side, an output document can be seen under an electrostatic print stylus. After charge deposition, output documents are placed in a bath of liquid toner and "fixed" on a hot plate. Electrostatic printing is used to simulate hardcopy output for a practical

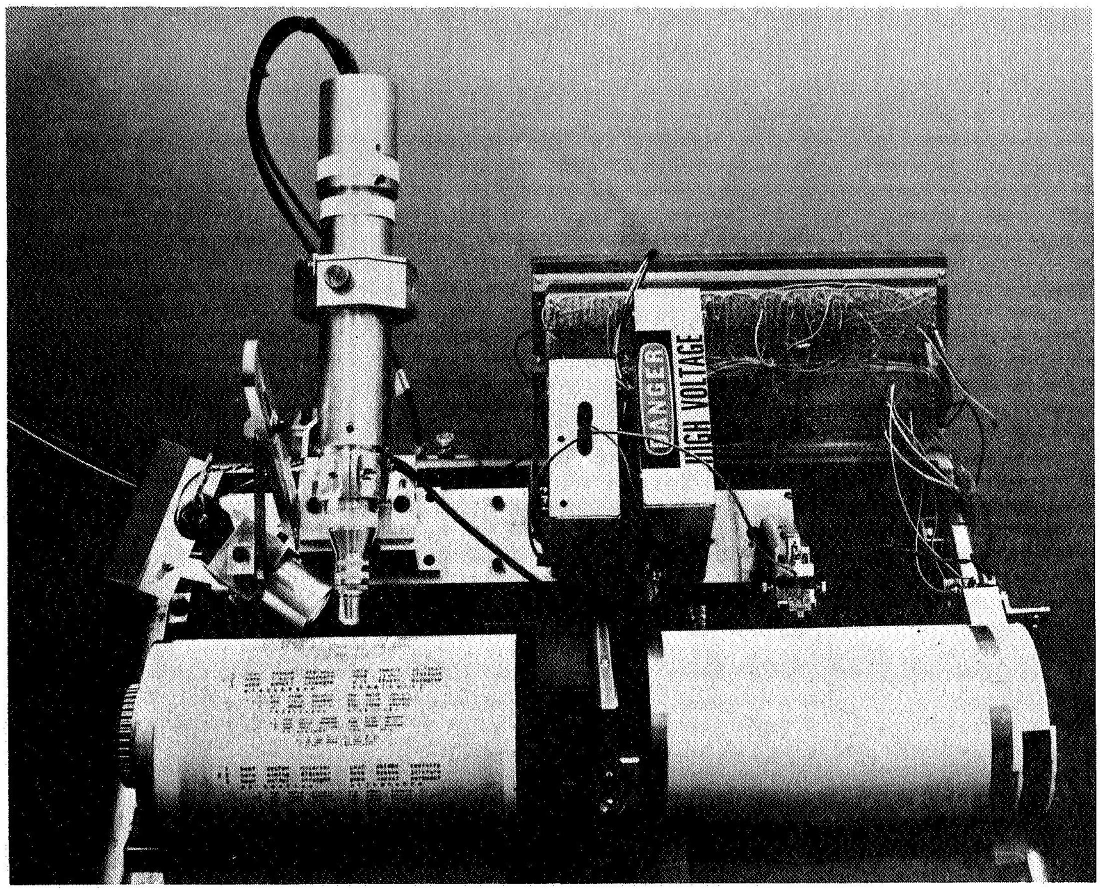


Fig. A.1-2. MECHANICAL SCANNER/PRINTER.

system. This process is carefully controlled, to maintain consistent output quality for experiments measuring the effects of resolution.

Although it is possible to transmit real-time from scanner to printer, this mode is avoided (except for testing). The synchronization errors between the CPU and mechanical I/O must be simulated. Interface clock rates are slaved to the drum, using signals from an optical clock track for frequent reset of a free-running multivibrator.

To digitize the image, area integration is first performed, followed by sampling and a 50 percent decision level. This results in a two-level dissection of the document into rectangular binary "elements," (illustrated in Fig. 2.1-1 of the text).

Area integration transverse to the direction of scan is mechanical. Five interchangeable scanning slits are used, one for each transverse resolution. Scanning-slit substitutions are accompanied with matching changes in the lead screw drive. Parallel to the scan, integration is electronic; with sample and squelch rates changeable to also obtain five different resolutions. Thus, 25 rectangular scan and patch resolutions are available. (Their values are specified in the tables of Appendix A.3.)

For the printer, five print heads are used to change transverse resolution (along with the same lead screw drive shared with the scanner). Resolution parallel to the scan is again determined in the electronics. The printing is done using a contact stylus, continuously "on" for contiguous black elements.

The position accuracy achieved in scanning and printing was theoretically estimated to be ±0.5 mil. Visual measurement from output data was limited to ±0.5 mil due to fuzzy edges from printing electrostatically. Such measurements, however, tended to confirm that the desired resolution values had been achieved. (See Fig. A.1-3). Since the same hardware is shared for scanning and printing, these measurements apply as well to the scanning process.

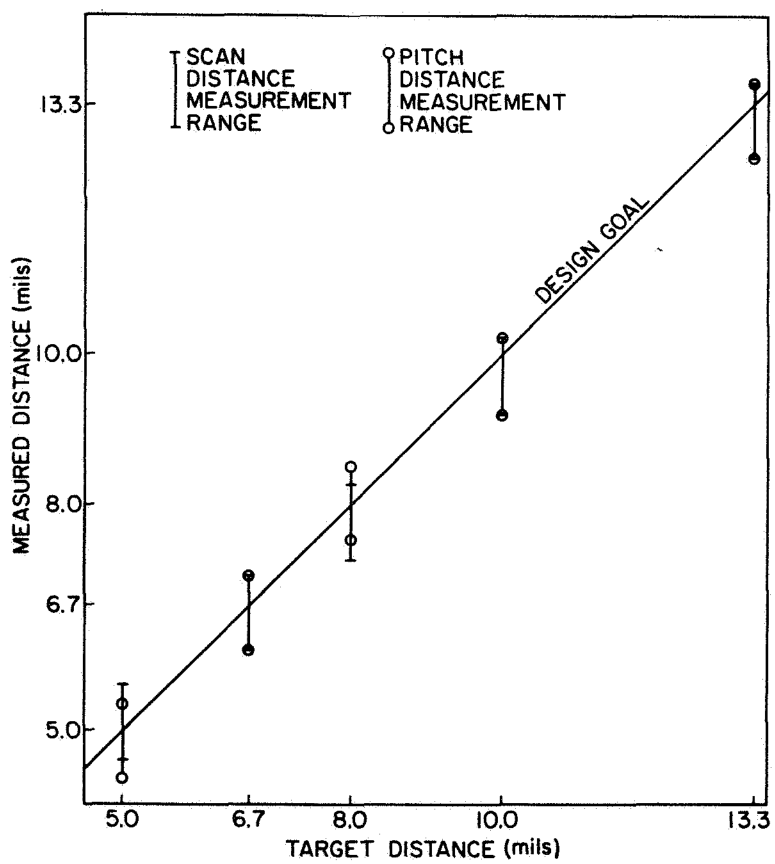


Fig. A.1-3. SPATIAL POSITION ACCURACY.

These visual measurements were made from microphotographs taken of the same region in each test document (Fig. 2.1-1 in the text, again, illustrates one such picture). To measure positional accuracy, the width of a "step" is used (to cancel out edge effects). To isolate a pair of edge effects, the width of a one bit "stroke" is measured. Then the difference between "step" and "stroke" widths indicate the amount of two edge effects.

Second order effects were present due to the electrostatic printing process. It had been observed that in printing, black areas ended up slightly larger after charge deposition and toning. To compensate for this in the pitch direction, slightly narrower print heads were used. This compensation was insufficient however, since visual measurements indicate an average of 0.7 mil widening in the pitch direction at each black edge.

In the scan direction, a single-shot was used to delete a fixed amount of print voltage "on"-time, for every black run. This was to compensate for the "spread" effect at the two edges of a black run. In practice, measurements indicate that an average of 0.5 mil was lost at every black edge due to over-compensation for "spread" in the scan direction.

These second order printing variations should have small effect on legibility. In all cases the desired dot to be printed was present. Only its shape was distorted from square to rectangular. The same applies to variations in toning. Cases where a print dot is visible but lighter in color should have only a second-order effect on legibility.

Images stored in the document library were subsequently used for the entropy studies in Chapter III. (Fig. A.1-4). An IBM 360 model

DOCUMENT LIBRARY

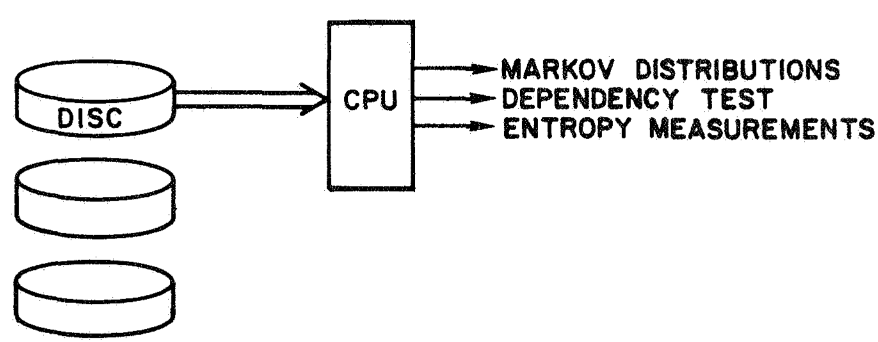


Fig. A.1-4. OFF-LINE IMAGE ANALYSES.

40 was used to estimate the Markov distributions, and compute χ^2 statistics, for these images. Different source alphabets could be readily programmed and entropy measurements made.

A.2. Test Documents

The first two sets of documents, A and B, were designed by human factors engineers R. L. Erdman, and A. S. Neal, associates at IBM. These test documents played an integral part in their measurements of document legibility, to evaluate throughput from the experimental system [Arps, et al - 1966]. The raw data from these legibility measurements is summarized in Appendix A.3.

The test document in Fig. A.2-1 was used not only for basic legibility measurements, but was useful in tuning up the experimental system. The upper part with four sizes of type constitutes document A1.

Figures A.2-2 through A.2-6 represent five sizes of Dual-Gothic type, aligned vertically as well as well as horizontally. This font was selected for its uniformity and stroke width as well as lack of serifs.

Documents C1 and C2 were from a random selection (picked by a secretary just requested to type a page full of text). This sample contains some 1400 letters. In Fig. A.2-7, its letter frequencies are compared with larger random samples by [Pratt - 1942] and [Dewey - 1923].

The documents were generated on a typewriter capable of changing fonts. In this way, the position of characters in the two documents were identical. Document C1 was made using common Pica type (Fig. A.2-8). Then, for comparison with documents B1-5, Document C2 was made using Dual-Gothic type (Fig. A.2-9).

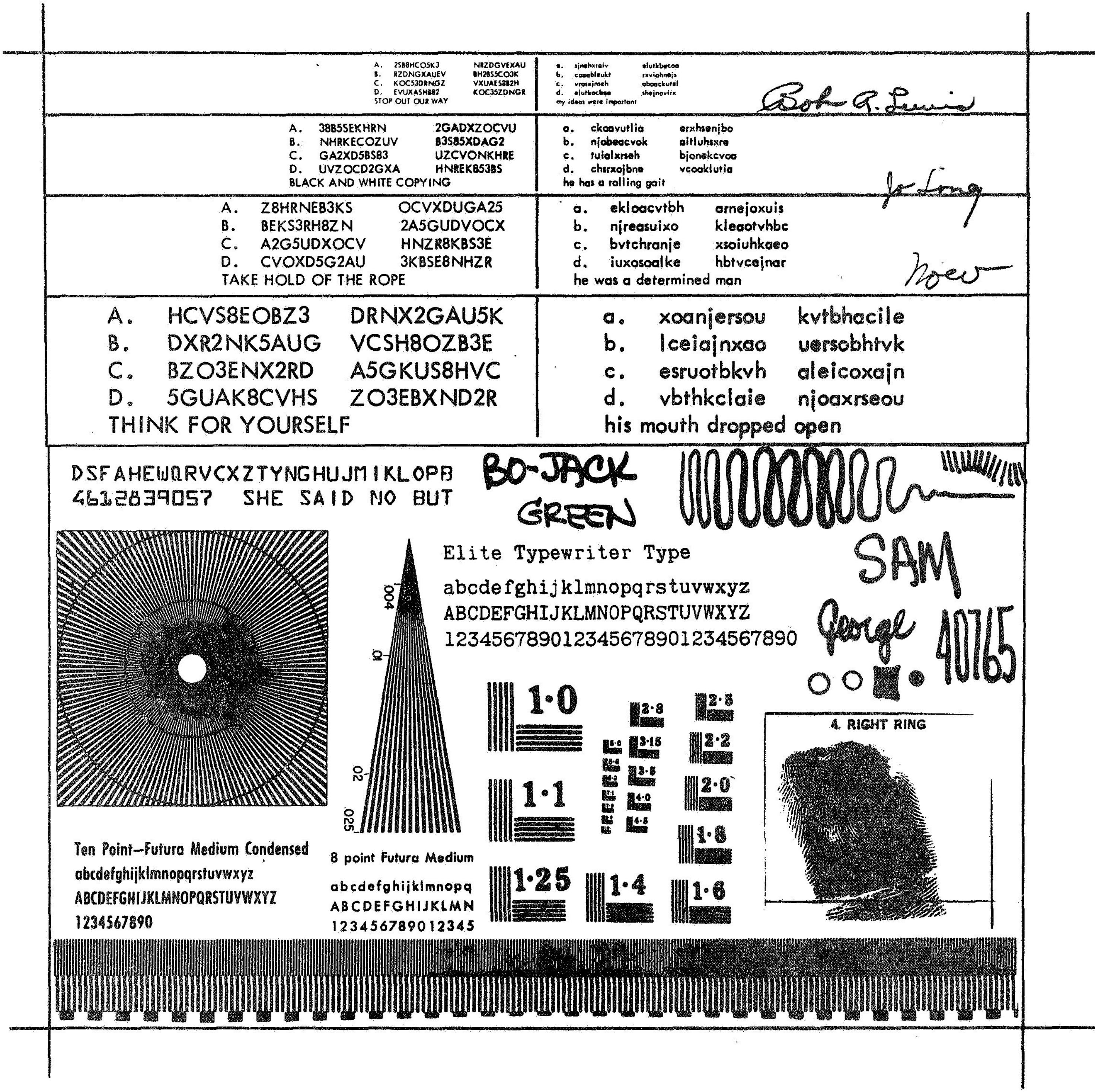


Fig. A.2-1. TEST DOCUMENT A1: MID-CENTURY TYPE AT 5.0, 7.5, 10.0, AND 15.0 MIL STROKE-WIDTHS.

AUBR8ZCS3H	ENXDOVGK52
UGE2KS8HOV	3XBNZC5ARD
G53DAHKVZC	OBEXS8RU2N
5RONUVACS8	ZE3BHK2GDX
R2ZXGCU8HK	S3OEVAD5NB
2DSB58GKVA	HOZ3CUNRXE
DNHERK5ACU	VZSO8GX2B3
NXV32ARU8G	CSHZK5BDEO
XBCODU2GK5	8HVSAREN3Z
BE8ZNGD5AR	KVCHU23XOS
E3KSX5NRU2	AC8VGDOBZH
3OAHBRX2GD	U8XC5NZESV
OZUVE2BD5N	GKA8RXS3HC
ZSGC3DENRX	5AUK2BHOV8
SH58ON3X2B	RUGADEVZCK
HVRKZXOBDE	2G5UN3CS8A
VC2ASBZEN3	D5RGXO8HKU
C8DUHES3XO	NR25BZKVAG
8KNGV3HOBZ	X2DRESACU5
KAX5COVZES	BDN23HU8GR

ovbuarehki	eoljtaxsnc
juvooxstha	iecnblarek
nouejarbtl	askevcoxih
eeosnoxvbc	lrhiukjaat
iserejauvk	cxtaohnolb
arsxinoouh	kabletejcv
lxraaejeot	hovcsbinku
caxolinseb	tjukrvaeho
koajcaersv	bnohxulite
hjonklixru	veetaocabs
tnjehcaaxo	uisboeklvr
benitkloae	oarvjshcux
vieabhcjos	elxunrtkoa
uailvtknjr	scaoexbheo
olacubhenx	rkoeiavtsj
eclkovtiea	xhjsaoubrn
skcheubaio	atnrljovxe
rhktsovlaj	obexcneuai
xthbreucln	jviakesooa
abtvxsokce	nuaohirejl

Fig. A.2-2. DOCUMENT B1: DUAL-GOTHIC TYPE, 12.0 MIL STROKE WIDTH.

KARHXCDVUB	OGS3Z8EN25	jcnvxaeueb	hliootskar
AU2VB8NCGE	Z5HOSK3XDR	nkeuaoiosv	tcajebrhlx
UGDCEKX853	SRVZHAOBN2	ehioojaeru	bklnsvxtca
G5N83ABKRO	H2CSVUZEXD	itaejnlsxo	vhceruabko
5RXKOUEA2Z	VD8HCGS3BN	ablsnecrae	utkixoovhj
R2BAZG3UDS	CNKV85HOEX	lvcreikxos	obhaaejutn
2DEUS5OGNH	8XACKRVZ3B	cukxiahajr	evtlosnobe
DN3GHRZ5XV	KBU8A2CSOE	kohaaltonx	subcjreevi
NXO5V2SRBC	AEGKUD8HZ3	hetolcbjea	rovknxisua
XBZRCDH2E8	U35AGNKVSO	tsbjckvnio	xeuheaarol
BES28NVD3K	GORU5XACHZ	brvnkhueaj	asotiolxec
E3HDKXCNOA	5Z2GRBU8VS	vxuehtoiln	orebajcask
3OVNAB8XZU	RSD52EGKCH	uaoitbeace	jxsvlnkorh
OZCXUEKBSG	2HNRD35A8V	ooeabvslki	narucehjxt
ZS8BG3AEH5	DVX2NORUKC	ejslvurcha	eoxokitnab
SHKE5OU3VR	NCBDXZ2GA8	snrcuoxktl	ijaehabeov
HVA3RZGOC2	X8ENBSD5UK	rexkoeahbc	anostlviju
VCUO2S5Z8D	BK3XEHNRGA	xiahesotvk	lejrbcuano
C8GZDHRSKN	EAOB3VX25U	aaotsrjbuh	cinxvkolee
8K5SNV2HAX	3UZEOCBDRG	oljbrxnvot	kaeauhecis

Fig. A.2-3. DOCUMENT B2: DUAL-GOTHIC TYPE, 10.0 MIL STROKE WIDTH.

5K8HU3DGOB	RNASVX2CEZ	nukojeblha	vcxoitsrea
RAKVGON5ZE	2XUHCBD83S	eohensvcto	ukajabrxil
2UAC5ZXRS3	DBGV8ENKOH	ietserukbj	ohonlvxaac
DGU8RSB2HO	NE5CK3XAZV	asbrixohvn	etjecuaolk
N5GK2HEDVZ	X3R8AOBUSC	lrvxaaetue	sbnikoojch
XR5ADV3NCS	BO2KUZEGH8	cxualosboi	rveahejnkt
B2RUNCOX8H	EZDAGS35VK	kaoocjrvea	xuiltsnehb
ED2GX8ZBKV	3SNU5HORCA	hoejknxusl	aoacbreitv
3ND5BKSEAC	OHXGRVZ28U	tjsnheaorc	oelkvxiabu
OXNREAH3U8	ZVB52CSDKG	bnretloexk	jschuaalvo
ZBX23UVOGK	SCERD8HNA5	vexibajsah	nrktoolcue
SEBDOGCZ5A	H832NKVXUR	uiaavlnrot	exhbejckos
H3ENZ58SRU	VKODXACBG2	oaolucexjb	iatvsnkher
VO3XSRKH2G	CAZNBU8E5D	eljcokianv	aoburehtsx
CZOBH2AVD5	8USXEGK3RN	scnkehaoeu	ljvoxitbra
8SZEVDUCNR	KGHB35AO2X	rkehstljio	cnueaabvxo
KHS3CNG8X2	A5VEORUZDB	xhitrbcnae	keosolvuaj
AVHO8X5KBD	URC3Z2GSNE	atabxvkels	hlerjcuoon
UCVZKBRAEN	G28OSD5HX3	oblvauhicr	tasxnkoeje
G8CSAE2U3X	5DKZHNRVBO	jvcuootakx	blraehesni

Fig. A.2-4. DOCUMENT B3: DUAL-GOTHIC TYPE, 8.0 MIL STROKE WIDTH.

BGCS2VUZE3	D8RXK5HNOA
E58HDCGS30	NK2BARVXZU
3RKVN85HOZ	XADEU2CBSG
O2ACXKRVZS	BUN3GD8EH5
ZDU8BA2CSH	EGXO5NK3VR
SNGKEUD8HV	35BZRXAOC2
HX5A3GNKVC	ORES2BUZ8D
VBRUO5XAC8	Z23HDEGSKN
CE2GZRBU8K	SDOVN35HAX
83D5S2EGKA	HNZCXORVUB
KONRHD35AU	VXS8BZ2CGE
AZX2VNORUG	CBHXESD853
USBDCXZ2G5	8EVA3HNKRO
GHEN8BSD5R	K3CUOVXA2Z
5V3XKEHNR2	AO8GZCBUDS
RCOBA3VX2D	UZK5S8EGNH
28ZEUOCBDN	GSARHK35XV
DKS3GZ8ENX	5HU2VAORBC
NAHO5SK3XB	RVGDCUZ2E8
XUVZRHAOBE	2C5N8GSD3K

xbleoatnao	ecurkvisjh
avcielbeoj	skoxhuarnt
oukascvijn	rheatolxeb
johlrkuane	xtsobecaiv
netcxholei	abrjvskoau
esbkatecia	ovxnurhjlo
irvhobskal	juaeoxtnce
axutjvrhlc	nooieabeks
laobnuxtck	eejasovihr
coeveoabkh	isnlrjuatx
kjsuieovht	arecxnolba
hnroasjutb	lxikaeecvo
texelrnobv	caaholskuj
biascxeevu	koltjarhon
vaorkaisuo	hjcbnlxtee
uljxhoaroe	tnkvecabsi
ocnatjlxes	behuikovra
ekeobncasr	vitoahjuxl
shijvekorx	uabeltnoac
rtanuihjxa	olvscbeeok

Fig. A.2-5. DOCUMENT B4: DUAL-GOTHIC TYPE, 6.0 MIL STROKE WIDTH.

KZXG6DOVUN	5H2SRECAB3
ASB5KNZCGX	RVDH238UEO
UHERAXS85B	2CNVDOKG3Z
GV32UBHKRE	D8XCNZA5OS
5COOGEVA23	NK88XSURZH
R8ZN53CUDO	XAEK8HG2SV
2KSXRO8GNZ	BU3AEV5DHC
DAHB2ZK5XS	EGOU3CRNV8
NUVEOSARBH	35ZGO82XCK
XGC3NHU2EV	ORS5ZKDB8A
B58OXVGD3C	Z2HRSANEKU
ERKZ8C5NO8	SDV2HUX3AG
32ASE8RXZK	HNCDVGBOU5
ODUH3K2BSA	VX8NC5EZGR
ZNGVOADEHU	CBKX8R3S52
SX5CZUN3VG	8EA8K2OHRD
HBR8SGXOC5	K3UEADZV2N
VE2KH5BZ8R	AOG3UNSCDX
C3DAVRESK2	UZ5OGXH8NB
8ONUC23HAD	GSRZ5BVKXE

otunaejcek	aivihxbosr
eboelsnkih	oauctavjrx
svelcrehat	jlokbounxa
rusakxltlb	ncehvjoeao
xorlhaabcv	ekstuneloj
aexctolvku	lhrboesajn
osakbjcuho	atxveirine
jrohvnkote	lbausaxcel
nxjtuehebs	cvoorlakia
eanboitsvr	kujexcohal
ioeveabrux	honsakjtlc
ajluslvxoa	teerohnbck
lnaorcuaeo	bslxjtevkh
celexkoosj	vraanbiuht
klcsahejrn	uxloevaotb
hakrotsnxe	oacjiulebv
tlhxjbreai	eoknaocsvu
bctanvxioa	sjhelekruo
vkboeuaajl	rnticshxoe
uhvjioolnc	xebakrtaes

Fig. A.2-6. DOCUMENT B5: DUAL-GOTHIC TYPE, 4.0 MIL STROKE WIDTH.

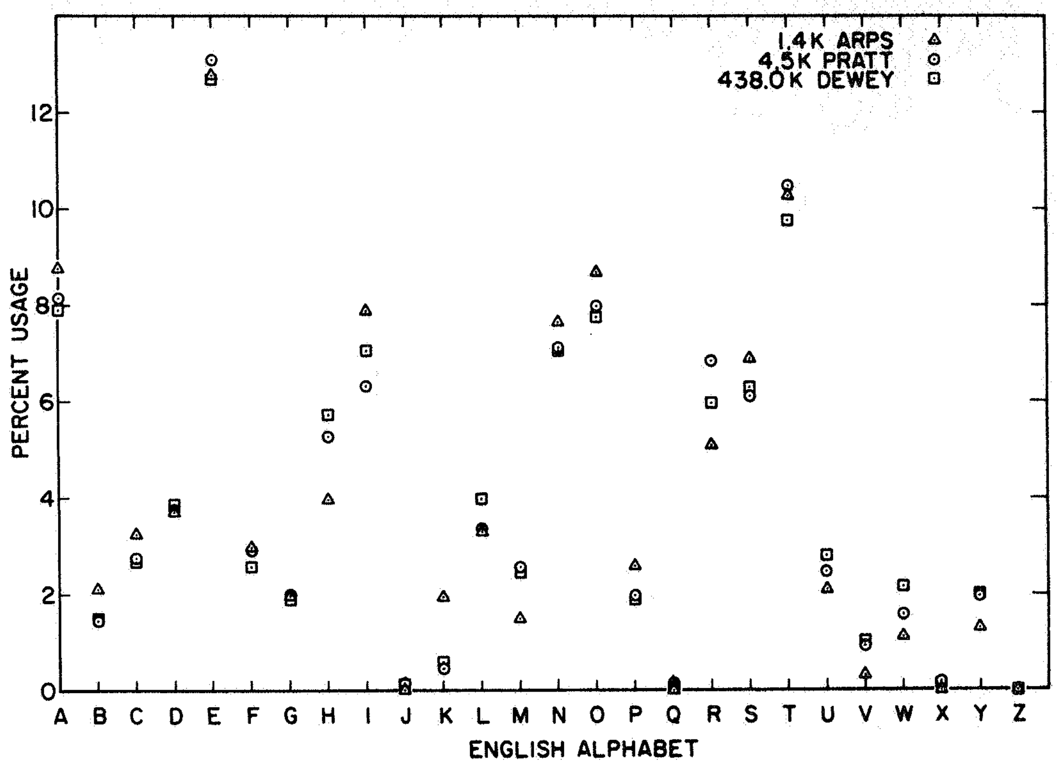


Fig. A.2-7. LETTER FREQUENCY FOR DOCUMENTS C1 AND C2.

The following cautions may seem unimportant to some; but it is surprising how many shorthand writers sidestep these definite aids to efficiency. Date the notebook every day. Much needless searching through notebooks has been caused by neglect to do this. Have one place for the notebooks on your desk. Keep notebooks for one year. Then discard the oldest when filing the latest. It is not unusual to be asked to retranscribe notes a year old. If taking dictation from different persons, put the initials of each at the beginning of each block of dictation--or use separate books. Cross off dictation immediately after transcribing each page. This is an important safeguard against the possible omission of some part of the work. Keep a rubber band around the notebook to mark the end of the finished notes, in order not to be constantly fluttering pages looking for a place to write when called for dictation. If dictation, such as a telegram, is taken on a separate piece of paper, date that paper and file it away with the old notebooks. It may be necessary to refer to it again. Do not be extravagant with notebooks by making large, careless notes. If it is difficult to keep the notes a natural size, use a notebook with unglazed paper. The pencil will not glide so easily over this paper, and the strokes will be retarded, making smaller, neater notes. Check the notebook at the end of each day to see that nothing important is left undone. Failure to do this has caused trips back to the office at night. Serious consequences might arise if, for instance, a wire is not dispatched as it should be. Pen and Pencils. Keep a fountain pen filled and three dictation pencils always in readiness in one certain place on the desk. Sharpen pencils after taking dictation, not after being called to take dictation. The best shorthand pencil is the medium soft No. 2. The No. 1 is too soft to hold a point; and

Fig. A.2-8. TEST DOCUMENT C1: PICA TYPE, 13.0 MIL STROKE WIDTH.

The following cautions may seem unimportant to some; but it is
surprising how many shorthand writers sidestep these definite aids
to efficiency. Date the notebook every day. Much needless search-
ing through notebooks has been caused by neglect to do this.
Have one place for the notebooks on your desk. Keep notebooks
for one year. Then discard the oldest when filing the latest.
It is not unusual to be asked to retranscribe notes a year old.
If taking dictation from different persons, put the initials of
each at the beginning of each block of dictation--or use separ-
ate books. Cross off dictation immediately after transcribing
each page. This is an important safeguard against the possible
omission of some part of the work. Keep a rubber band around the
notebook to mark the end of the finished notes, in order not to
be constantly fluttering pages looking for a place to write when
called for dictation. If dictation, such as a telegram, is taken
on a separate piece of paper, date that paper and file it away
with the old notebooks. It may be necessary to refer to it again.
Do not be extravagant with notebooks by making large, careless
notes. If it is difficult to keep the notes a natural size, use
a notebook with unglazed paper. The pencil will not glide so
easily over this paper, and the strokes will be retarded, making
smaller, neater notes. Check the notebook at the end of each day
to see that nothing important is left undone. Failure to do this
has caused trips back to the office at night. Serious conse-
quences might arise if, for instance, a wire is not dispatched as
it should be. Pen and Pencils. Keep a fountain pen filled and
three dictation pencils always in readiness in one certain place
on the desk. Sharpen pencils after taking dictation, not after
being called to take dictation. The best shorthand pencil is the
medium soft No. 2. The No. 1 is too soft to hold a point; and

Fig. A.2-9. TEST DOCUMENT C2: DUAL-GOTHIC TYPE, 10.0 MIL STROKE WIDTH.

A.3. Human Factors Data on Legibility vs Resolution [Arps, et al - 1966].

To cancel out systematic machine errors due to the direction of scan, all resolution combinations were tested twice, once with a document scanned vertically and again with a horizontally scanned document. The variations due to scan direction were not significant, so that observations were combined for pairs of documents at the same resolution combination. (These combinations are specified as "horizontal" and "vertical" with respect to the lines of type on a document).

Legibility results are given in Tables A.3-1 and A.3-2 for upper- and lowercase Mid-Century type respectively. These values are for 4 sizes of type at all 25 resolution combinations. They were each estimated using the sample mean from 80 observations (where 80 subjects read 10 characters apiece):

$$\hat{L}_i(f_h, f_v, w_s) = \frac{1}{n} \sum_{j=1}^{n} L_{ij}(f_h, f_v, w_s) \qquad \%$$

where:

$L_{ij}(f_h, f_v, w_s)$ = an observation of legibility %

$\hat{L}_i(f_h, f_v, w_s)$ = the sample mean value of legibility at one resolution combination %

f_h = horizontal spatial frequency elem./inch

f_v = vertical spatial frequency elem./inch

w_s = character size in terms of stroke width mils

$n = 80$ = number of observations

The standard deviation of the sample mean was also estimated for each of the resolution combinations using the unbiased statistic:

$$\hat{S}_{L_i}(f_h, f_v, w_s) = \sqrt{\frac{1}{n(n-1)} \sum_{j=1}^{n} [L_{ij}(f_h, f_v, w_s) - \hat{L}_i(f_h, f_v, w_w)]^2} \qquad \text{mils}$$

The standard deviation for the measurements of uppercase Mid-Century type are given in Table A.3-3.

Note that in reading the entries from any box of a Table, they correspond to character sizes in descending order. For example, in the upper left box of Table A.3-1, the four entries of percent legibility are:

92.5	for 15 mil stroke-width (114 mils high)
38.25	for 10 mil stroke-width (76 mils high)
5.25	for 7.5 mil stroke-width (59 mils high)
0.0	for 5.0 mil stroke-width (38 mils high)

Table A.3-4 gives legibility values for upper- and lowercase Dual-Gothic type. These values were measured at only one resolution combination, 8.0×8.0 mils/elem. (equivalent to the center box in Figs. A.3-1 and A.3-2). However, additional data is given for the various sizes of type corrupted by a model for transmission errors.

Table A.3-5 gives corresponding standard deviation values for the Dual-Gothic legibility data.

Table A.3-1

UPPERCASE, MID-CENTURY TYPE (FOUR SIZES)
LEGIBILITY AT HORIZONTAL VS VERTICAL RESOLUTION COMBINATIONS
FOR 15.0, 10.0, 7.5, AND 5.0 MIL STROKE-WIDTHS

HORIZONTAL RESOLUTION

Spatial Distance, d_h, in Mils/Elem.
(Spatial Frequency, f_h, in Elem./Inch)

VERTICAL RESOLUTION Spatial Distance, d_v, in Mils/Elem. (Spatial Frequency, f_v, in Elem./Inch)	13.3 (75)	10.0 (100)	8.0 (125)	6.7 (150)	5.0 (200)
13.3 (75)	92.5 38.25 5.25 0.0	94.625 69.25 20.75 0.25	96.75 80.0 42.75 0.25	97.625 90.0 55.125 3.0	96.375 86.0 54.50 1.75
10.0 (100)	96.125 73.875 30.375 0.0	97.0 90.25 72.875 0.625	99.5 94.25 84.625 10.5	99.625 94.50 89.25 30.5	99.125 97.625 89.875 35.875
8.0 (125)	95.625 79.75 36.25 1.0	98.875 95.125 85.0 6.5	99.25 96.75 90.5 31.25	99.125 96.625 95.625 54.5	99.875 98.875 94.25 59.0
6.7 (150)	97.75 84.0 47.875 3.5	99.75 95.125 89.5 23.375	99.625 98.5 93.0 50.75	100.0 99.375 94.5 72.125	99.875 99.125 95.875 83.875
5.0 (200)	98.25 85.75 54.25 3.375	99.875 98.375 88.125 22.0	99.625 98.875 92.375 44.75	99.875 98.75 95.5 81.0	100.0 99.75 97.875 86.75

Table A.3-2

LOWERCASE, MID-CENTURY TYPE (FOUR SIZES)
LEGIBILITY AT HORIZONTAL VS VERTICAL RESOLUTION COMBINATIONS
FOR 15.0, 10.0, 7.5, AND 5.0 MIL STROKE-WIDTHS

HORIZONTAL RESOLUTION

Spatial Distance, d_h, in Mils/Elem.
(Spatial Frequency, f_h, in Elem./Inch)

VERTICAL RESOLUTION
Spatial Distance, d_v, in Mils/Elem.
(Spatial Frequency, f_v, in Elem./Inch)

	13.3 (75)	10.0 (100)	8.0 (125)	6.7 (150)	5.0 (200)
13.3 (75)	82.5 14.25 0.75 0.0	87.75 44.325 6.875 0.25	95.0 67.25 11.625 0.25	94.25 70.75 29.75 2.25	98.25 67.875 19.375 1.375
10.0 (100)	92.125 55.5 13.5 0.125	95.125 80.5 44.375 0.875	98.5 89.375 62.625 4.0	99.25 92.0 73.875 9.0	99.75 94.75 77.375 15.125
8.0 (125)	95.25 55.5 21.75 0.125	98.625 89.125 55.375 2.0	99.125 93.25 80.25 15.125	99.625 97.375 83.875 22.25	99.125 98.625 94.25 33.0
6.7 (150)	95.875 68.375 27.25 0.25	99.125 93.375 71.25 7.25	99.125 97.375 88.625 26.375	99.625 98.25 87.125 47.375	99.5 98.75 96.5 60.125
5.0 (200)	97.0 64.75 31.75 0.5	99.25 94.125 72.0 3.75	99.25 97.375 92.375 21.375	99.625 98.5 95.375 53.5	99.0 99.5 99.375 67.0

Table A.3-3

UPPERCASE, MID-CENTURY TYPE (FOUR SIZES)
STANDARD DEVIATION AT HORIZONTAL VS VERTICAL RESOLUTION COMBINATIONS
FOR 15.0, 10.0, 7.5, AND 5.0 MIL STROKE-WIDTHS

HORIZONTAL RESOLUTION

Spatial Distance, d_h, in Mils/Elem.
(Spatial Frequency, f_h, in Elem./Inch)

VERTICAL RESOLUTION
Spatial Distance, d_v, in Mils/Elem.
(Spatial Frequency, f_v, in Elem./Inch)

	13.3 (75)	10.0 (100)	8.0 (125)	6.7 (150)	5.0 (200)
13.3 (75)	6.264 21.275 9.274 0.0	6.038 14.275 19.320 0.0	6.721 11.904 18.581 3.019	5.025 10.862 18.940 7.647	4.975 12.443 22.263 6.925
10.0 (100)	6.740 20.975 19.859 1.571	6.038 9.274 18.500 2.436	3.556 6.363 12.729 11.374	1.571 7.462 12.917 18.755	1.118 4.039 11.034 17.603
8.0 (125)	5.460 19.093 27.141 1.571	2.193 6.319 14.048 16.374	3.091 5.223 9.264 20.337	1.911 3.593 8.916 25.347	1.911 3.556 11.933 22.331
6.7 (150)	4.837 12.629 26.908 6.825	1.912 6.329 9.908 23.216	3.258 5.725 7.605 23.106	0.0 2.436 8.845 15.236	1.118 3.328 6.917 16.809
5.0 (200)	5.335 17.760 25.052 6.517	3.258 4.837 9.345 22.258	1.118 3.179 7.593 23.090	1.118 2.843 6.880 14.539	0.0 1.571 4.691 10.997

Table A.3-4

LEGIBILITY, DUAL-GOTHIC TYPE (FIVE SIZES)
UPPER- AND LOWERCASE, WITH AND WITHOUT ERRORS
FOR 12.0, 10.0, 8.0, 6.0, AND 4.0 MIL STROKE-WIDTHS

	Upper-Case	Lower-Case
Without Errors	99.206 99.175 95.394 75.019 28.462	99.575 99.331 98.069 85.800 37.712
With Errors	98.337 98.887 88.137 55.319 13.531	98.625 95.944 89.031 58.812 22.269

HORIZONTAL AND VERTICAL RESOLUTION

$d_h = d_v = 8.0$ mils/elem.
$(f_h = f_v = 125$ elem./inch)

Table A.3-5

STANDARD DEVIATION, DUAL-GOTHIC TYPE (FIVE SIZES)
UPPER- AND LOWERCASE, WITH AND WITHOUT ERRORS
FOR 12.0, 10.0, 8.0, 6.0, AND 4.0 MIL STROKE-WIDTHS

	Upper-Case	Lower-Case
Without Errors	2.101 2.113 4.471 10.307 12.147	1.505 2.040 3.273 9.240 17.461
With Errors	3.084 4.424 9.450 16.714 9.635	2.682 5.058 9.246 16.821 13.782

HORIZONTAL AND VERTICAL RESOLUTION

$d_h = d_v = 8.0$ mils/elem.
$(f_h = f_v = 125$ elem./inch)

A.4 References

1. Abramson, N., Information Theory and Coding, McGraw-Hill Book Co., New York, 1963.

2. Arps, et al, "Digital Facsimile: Transmission Economy and Copy Quality," Laboratory Memorandum ASDJ-M-097, ASDD Los Gatos Laboratory, IBM, Los Gatos, California May 31, 1966.

3. Arps, et al, "Optimal Resolution for Image Processing of Printed Matter," 1968 IEEE Int'l Conf. on Communications, pp. 838-843, June 1968.

4. Ash, R. B., Information Theory, John Wiley & Sons, New York, 1965.

5. Bartlett, M. S., "The frequency goodness of fit test for probability chains," Proc. Camb. Phil. Soc., vol. 47, pp. 86-95, 1951.

6. Billingsley, P., "Statistical methods in markov chains," Ann. Math. Stat., vol. 32, pp. 12-40, 1961.

7. Bracewell, R., The Fourier Transform and Its Applications, McGraw-Hill Book Co., New York, 1965.

8. Capon, J., "A probabilistic model for run-length coding of pictures," IRE Trans. on Information Theory, vol. IT-5, pp. 157-163, December 1959.

9. Cornog, D. Y. and Rose, F. C., Legibility of Alphanumeric Characters and Other Symbols, National Bureau of Standards Misc. 262-2, February 10, 1967.

10. Deutsch, S., "A note on some statistics concerning typewritten or printed material," IRE Trans. on Information Theory, vol. IT-3, pp. 1470148, June 1957.

11. Dewey, G., Relativ Frequency of English Speech Sounds, Harvard University Press, Cambridge, 1923.

12. Elias, P., "Predictive coding," IRE Trans. on Information Theory, vol. IT-1, pp. 16-33, March 1955.

13. Ezekiel, M., and Fox, K. A., Methods of Correlation and Regression Analysis, John Wiley & Sons, New York, 1963.

14. Goodman, J. W., Introduction to Fourier Optics, McGraw-Hill Book Co., San Francisco, 1968.

15. Hald, A., Statistical Tables and Formulas, John Wiley & Sons, New York, 1952, pp. 40-43.

16. Huang, T. S., "Pseudorandom scanning," IEEE Trans. on Communication Technology, vol. COM-12, pp. 105-106, September 1964.

17. Huffman, D. A., "A method for the construction of minimum-redundancy codes," Proc. IRE, vol. 40, pp. 1098-1101, September 1952.

18. Laemmel, A. E., "Coding processes for bandwidth reduction in picture transmission," Report R 246-51 Microwave Res. Inst., Polytechnic Inst. of Brooklyn, Brooklyn, N. Y., August 1951.

19. McLachlan, D., "Description mechanics," Information and Control, vol. 1, pp. 240-266, 1958.

20. Mood, A. M., and Graybill, F. A., Introduction to the Theory of Statistics, McGraw-Hill Book Co., New York, 1963.

21. Neyman, J., and Pearson, E. S., "On the use and interpretation of certain test criterion for purposes of statistical inference," Biometrika, vol. 20A, pp. 175-240, 263-294, 1928.

22. Ostle, B., Statistics in Research, The Iowa State University Press, Ames, Iowa, 1963.

23. Pratt, F., Secret and Urgent, The Story of Codes and Ciphers, Blue Ribbon Books, Garden City, N. Y., 1942.

24. Shannon, C. E., "A Mathematical Theory of Communication," Bell System Technical Journal, vol. 27, pp. 379-423, 623-656, July, October, 1948.

25. Shurtleff, D. A., "Studies in Television Legibility: A Review of the Literature," Information Display, pp. 40-45, January 1967.

26. Tinker, M. A., Legibility of Print, Iowa State University Press, Ames, Iowa, 1963.

27. Vitushkin, A. G., Theory of the Transmission and Processing of Information, Pergamon Press, New York, 1961.

28. Whittaker, E. T., On the functions which are represented by the expansion of the interpolation theory, Proc. Roy. Soc. Edinburgh, Sect. A, vol. 35, pp. 181, 1915.

29. Wholey, J. S., "The Coding of Pictorial Data," IRE Trans. on Information Theory, vol. IT-7 , pp. 99-104, April 1961.

30. Wilks, S. S., "The likelihood test of independence in contingency tables," Ann. Math. Stat., vol. 6, pp. 190-196, 1935.

A.5 Bibliography

Two extensive bibliographies that have already been published in the area of redundancy reduction are:

1. Cherry, E. E., and Gouriet, G. G., "Some possibilities for the compression of television signals by recoding," Proc. IEE, vol. 100, pt. 3, pp. 9-18, January 1953.

2. Pratt, W. K., "A bibliography on television bandwidth reduction studies," IEEE Trans. on Information Theory, vol. IT-13, pp. 114-115, January 1967.

Since the latter was written, a special issue of the IEEE Proceedings has been published on Redundancy Reduction:

3. Cutler, C. C., Proc. IEEE Special Issue on Redundancy Reduction, vol. 55, pp. 251-406, March 1967.

In addition to the references in Appendix A.3 and the three sources above, the following recent additions need to be made:

4. Davisson, L. D., "The theoretical analysis of data compression systems," Proc. IEEE, vol. 56, pp. 176-186, February 1968.

5. Limb, J. O., "Entropy of quantised television signals," Proc. IEE, vol. 115, pp. 16-20, January 1968.

6. Ott, G., "Compact encoding of stationary markov sources," IEEE Tran. on Information Theory, vol. IT-13, January 1967.

7. Richards, G. P., and Bisignani, W. T., "Redundancy reduction applied to coarse-fine encoded video," Proc. IEEE, vol. 55, pp. 1707-1717, October 1967.