Journal of Electronic Imaging

JElectronicImaging.org

Invariant representation for rectilinear rulings

George Nagy

SPIE. IS&T imaging.org

Invariant representation for rectilinear rulings

George Nagy*
Rensselaer Polytechnic Institute, Department of Electrical, Computer and Systems Engineering, JEC,
110 8th Street, Troy, New York 12180, United States

Abstract. Ruling gap ratios are an affine-invariant characterization of parallel ruling configurations in scanned documents. This report quantifies the advantage of simultaneous extraction of horizontal and vertical rulings. It demonstrates that every ruling gap ratio can be derived from a minimal set of basis ratios. The effect on the basis ratios of noise on the radial coordinates of individual rulings is analyzed and the dependence of basis-ratio variability on random-phase sampling noise is determined as a function of the spatial sampling rate. The analysis provides insight into already-presented small-scale experimental results on form classification and guidance for future work that requires the extraction of parallel lines from scanned or photographed images. *© 2014 SPIE and IS&T* [DOI: 10.1117/1.JEI.23.6.063011]

Keywords: forms processing; form classification; isothetic lines; geometric invariance; ruling gap ratio; ratio of correlated Gaussians; logarithmic quantization; edit distance.

Paper 14575P received Sep. 15, 2014; accepted for publication Nov. 3, 2014; published online Dec. 8, 2014.

1 Introduction

Horizontal and vertical rulings (or rules) are prevalent in bureaucratic forms, tables, schematic diagrams, organization charts, music scores, and engineering drawings. In some applications, the configuration of such isothetic rulings can serve as a signature for a family of similar objects. While the ruling configuration is generally obvious to a human observer, extracting it from a bi-level scanned document and representing it in a format useful for classification often encounters the following difficulties:

- Spurious lines may appear due to alignment of nonruling document components.
- Rulings may disappear because of inadequate spatial or amplitude quantization.
- Copied and scanned images may be subject to arbitrary translation, rotation, and scaling.

Selectivity for rulings over spurious lines is promoted by simultaneous extraction of the largest set of mutually parallel or perpendicular lines. The desirable geometric invariance can be secured by using ratios of distances between rulings rather than their locations within the image. The effect of missing lines can be alleviated by turning the sequence of ratios into a symbol string and measuring the similarity of ruling configurations with an edit distance. We have already demonstrated form classification based on these notions on a small dataset. The objective of this report is to provide some analytical support for the method.

In the next section, we review relevant prior work. In Sec. 3, we derive an approximation to the probability of error in ruling detection. Section 4 addresses the minimum number of ratios required to preserve the ruling configuration. Section 5 quantifies some effects of positional noise due to insufficient foreground pixels and to random-phase spatial sampling. For the sake of completeness, in Sec. 6, we briefly review our earlier experiment on form classification based on ruling gap ratios. In Sec. 7, we summarize our findings and list the remaining obstacles to developing a truly satisfactory predictive theory of ruling-based form classification.

2 Prior Work

Line segment recognition has been steadily improved during the last three decades as part of table interpretation,[1–4] form processing,[5–9] engineering drawing analysis,[10–13] and feature extraction from natural images.[14] Analysis of historical forms[15,16] became popular even as most contemporary forms migrated to the Web. The Hough transform that we use for line location has remained one of the leading methods for line and arc extraction since its rediscovery by Duda and Hart in the early 1970s.[17] It does not require edge linking and is, therefore, often preceded only by edge extraction with the venerable Prewitt filter[18] or other (Sobel, Roberts)[19] 3×3 pixel edge filters. The Hough transform was recently used with parameters similar to ours, but with a Canny filter, for strong line detection in a more general document analysis context.[20] We have found neither research addressing the extraction and quantification of rectilinear rule structures independently of other document content, nor prior application of orthogonal line filtering to Hough lines or skew detection.[21]

Our interest in spatial sampling noise was triggered by peaks in the autocorrelation function corresponding to opposite stroke edges in scanned character images.[22] The variation (noise!) due to repeated scanning was exploited by Zhou and Lopresti to decrease the optical character recognition (OCR) error.[23] Random-phase sampling noise was systematically investigated in remote sensing[24,25] and in scanned documents,[26] but pixel jitter is usually modeled as if it were independent random displacement of sensor elements.[27] The relationship between spatial and amplitude quantization in

*Address all correspondence to: George Nagy, E-mail: nagy@ecse.rpi.edu

document scanning for OCR was explored thoroughly by Barney Smith.[28]

Levenshtein introduced the edit distance for error-correcting codes in 1965.[29] The optimal Wagner-Fischer algorithm was published a decade later.[30] Many variations of the original algorithms have appeared since then.[31–34] The role of the edit distance in communications and text processing was augmented by its application to genome sequencing. Developments relevant to document image analysis include normalization methods[35] and kernel techniques for embedding the edit distance in a vector space.[36] The public-domain EDIT DISTANCE WEIGHTED program that we use was posted in 2010 by Schauerte and Fink.[37]

The current study was initiated during a phase of the multilingual automatic document classification and translation (MADCAT) project[38] aiming to categorize a small subset of the collection of Kurdish documents recovered during the Anfal uprising.[39,40] The Hough transform parameters and preliminary results on the classification of some degraded forms were presented at the 2014 SPIE conference on document recognition and retrieval[41] and some of the analysis at the 2014 workshop on statistical, structural, and syntactic pattern recognition.[42]

3 Ruling Selection

The rules are selected in five simple steps. An example of the result is shown in Fig. 1.

1. Extract near-horizontal line segments (those within θ_h degrees of the horizontal axis).
2. Extract near-vertical line segments (those within θ_v degrees of the vertical axis).
3. Add 90 deg to the angle of near-horizontal lines.
4. Histogram all angles into N bins.
5. Keep only lines with slope within one-half bin-width of the centroid of the peak bin.

The θ_h and θ_v thresholds depend on the maximum expected skew. We usually set them at 30 deg. After Step 3, all the lines are "near-vertical." The number N of histogram bins is not critical: we use 20 bins of one degree and two infinite-width bins on either side. The histogram typically contains 80 to 100 lines, of which 3 to 30 may be horizontal rulings (~30 for a form with writing lines for every entry), and 3 to 6 vertical rulings. The rest are "spurious" rulings engendered by accidental alignment of edge pixels. Let the total number of true rulings be R, and the number of spurious rulings be S, with $S > R$. If all R rulings fill into a single bin and if none of the other bins contain more than $R-1$ spurious rulings, then rule selection succeeds. What is the probability that the rule selection (and the corresponding skew determination) fails?.

We assume, optimistically, that the spurious lines are independently and randomly distributed with uniform slope probability. Then, the most probable failure configuration is R spurious rulings falling into some bin other than the true rulings' bin, and all the other $S-R$ spurious rules falling into separate bins, as shown in Fig. 2. There are

$$(N-1)\binom{N-2}{S-R}$$

equally likely configurations. All other failure configurations are at least $O(N)$ times less probable.

Once the configuration is fixed, the probability of R spurious lines in one bin with each of the other $(S-R)$ lines in a separate bin is a multinomial. A spurious line falls into any bin other than the rulings bin with probability $1/(N-1)$. Therefore, a tight lower bound on the selection error is

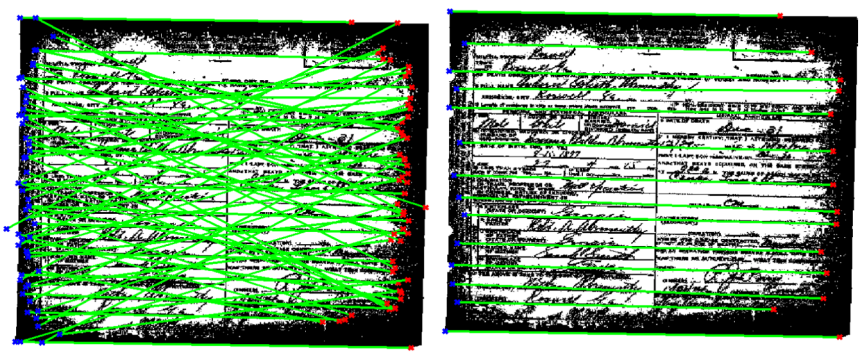

Fig. 1 Horizontal ruling lines from a skewed and noisy death certificate, extracted by ortho-filtering.

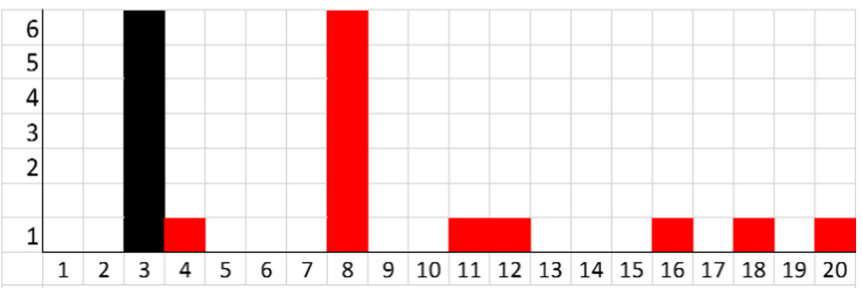

Fig. 2 Most likely configuration for error in rule selection for $N = 20$, $R = 6$, and $S = 12$.

$$P(\text{false\_max}) > \text{PeR}$$
$$= (N-1)\binom{N-2}{S-R}\binom{S}{R,1,1,\ldots,1}\left(\frac{1}{N-1}\right)^S$$
$$= \binom{N-2}{S-R}\frac{S!}{R!}\left(\frac{1}{N-1}\right)^{S-1}.$$

As seen from Table 1, PeR for $R = 6$, $S = 12$ is over 200 times less than for $R = 3$, $S = 6$ (for $N = 20$ bins). Hence, the striking advantage of combining the horizontal and vertical lines for ruling selection, especially when rulings are sparse. The assumption of independent and identically distributed (i.i.d.) angle distribution of the nonruling lines is, however, questionable. Furthermore, the number of bins (N) cannot be arbitrarily increased because the limited number of pixels constituting a ruling precludes precise estimation of its radial and angular coordinates.

The probability of any specific configuration of S indistinguishable lines in $N' = N - 1$ indistinguishable bins can be exactly computed via the sequence of conditional probabilities of generating a new configuration by adding a line to an existing configuration. (The usual terminology is balls into boxes, bins, or urns).

Let $P_S[M_{0,k_0}, M_{1,k_1}, \ldots, M_{i,k_i}, \ldots, M_{S,k_s}]$ denotes the probability of a histogram with k_0 empty bins. k_1 bins with 1 line, k_2 bins with 2 lines, and so on. There are $S = \Sigma_j(j \times k_j)$ lines altogether. The number of possible configurations, or partitions of S, is given by Sloan A00041 (OEIS) as 1, 1, 2, 3, 5, 7, 11, 15… The probabilities of nonzero bin occupancies, for S from 0 to 3, are

$$P_0[M_{0,N'}] = 1,$$

$$P_1[M_{0,N'-1}, M_{1,1}] = P_0[M_{0,N'}] \times 1,$$

$$P_2[M_{0,N'-2}, M_{1,2}] = P_1[M_{0,N'-1}, M_{1,1}] \times \frac{(N'-1)}{N'},$$

$$P_2[M_{0,N'-1}, M_{2,1}] = P_1[M_{0,N'-1}, M_{1,1}] \times \frac{1}{N'},$$

$$P_3[M_{0,N'-3}, M_{1,3}] = P_2[M_{0,N'-2}, M_{1,2}] \times \frac{N'-2}{N'},$$

$$P_3[M_{0,N'-2}, M_{1,1}, M_{2,1}] = P_2[M_{0,N'-2}, M_{1,2}] \times \frac{2}{N'} + P_2[M_{0,N'-1}, M_{2,1}] \times \frac{(N'-1)}{N'},$$

$$P_3[M_{0,N'-1}, M_{3,1}] = P_2[M_{0,N'-1}, M_{2,1}] \times \frac{1}{N'}.$$

The exact probability of skew detection error can be calculated. For $R = S$, the bound is exact. With $S = 6$, the bound underestimates the error for $R = 3$ by 19.4%, for $R = 4$ by 7.6%, and for $R = 5$ by 0.9%. For $N \gg S$, the error decreases as $O(R^{-1})$, with the constant of proportionality governed by S.

Table 1 Dominant term of the probability of false maxima in the angle histograms. The bold values show the advantage of doubling R and S.

N	R	S	PeR (%)	N	R	S	PeR (%)
20	3	3	0.27701	40	3	3	0.065746
20	**3**	**6**	**3.95461**	40	3	6	1.122005
20	3	9	6.61081	40	3	9	3.119688
20	3	12	3.33205	40	3	12	4.099149
20	6	6	0.00004	40	6	6	1.11E − 06
20	6	9	0.00242	40	6	9	7.94E − 05
20	**6**	**12**	**0.01060**	40	6	12	0.000579

An asymptotic approximation[43] of interest in load balancing of servers without central control and in estimating linked-list lengths from collisions in hashing states that, for $S = N'$, the expected maximum bin occupancy is $E[k_{\max}] \to [\ln N'/\ln(\ln N')]$; $[\ln 19/\ln(\ln 19)] \cong 2.73$. This suggests that conflicting configurations with more than three spurious lines in a bin are unlikely even with $S = N'$. In contrast to our bound, this equation offers no way of estimating the probability of error for $R > \ln N'/\ln(\ln N')$. For $N' = 19$ and $S = 6$, the exact conditional probability computation yields $E[k\max] = 1.64$.

4 Ruling Gap Ratios

If the configuration of rectangular rulings is to be used as a signature for a family of documents, then it is essential to circumvent the effects of arbitrary and unknown translation, rotation, and scaling that may occur while copying and scanning the documents. The perpendicular distance (i.e., the difference of ρ-coordinates) between a pair of parallel lines is invariant under translation and rotation, but not under scaling. However, the ratio of the perpendicular distances between rulings is invariant under all three transformations. (It is, in fact, invariant under a generalized affine transformation, which includes nonisotropic scaling and shear.).

In this section, we consider only one set of parallel lines. The results apply to both the near-horizontal and the near-vertical rulings. Information from the two sets of parallel rulings can be combined during the form classification phase. We make use only of the Hough ρ-coordinates, which are estimated from all the edge pixels of the ruling. The length and lateral position of the rulings often cannot be accurately determined because they depend only on a few end pixels.

Let ρ_a, ρ_b, ρ_c, and ρ_d be the radial coordinates of four arbitrarily selected parallel rulings L_a, L_b, L_c, and L_d. Consider all $R(R-1)/2$ pairwise distance ratios of the form $(\rho_a \text{ to } \rho_b)/(\rho_c \text{ to } \rho_d)$. How many of the possible ratios are required to characterize the disposition of R parallel rulings?

Let the extracted rulings, sorted according to their ρ coordinates, be $L_0, L_1, \ldots, L_k, \ldots, L_R$, with $\rho_k < \rho_{k+1}$. Define basis ratio $\gamma_i = (\rho_{i+2} - \rho_{i+1})/(\rho_{i+1} - \rho_i)$. Then, any pairwise inter-ruling distance $\rho_s - \rho_t$ can be written in terms of the first pair of rulings and the basis ratios γ_i as

$$\rho_s - \rho_t = (\rho_2 - \rho_1) \times \gamma_1 \times \gamma_2 \times \ldots . \\ \times \gamma_{t-1}(1 + \gamma_t\{1 + \gamma_{t+1}[\ldots(1 + \gamma_{s-2})]\}).$$

The γ_i terms on the left account for the rule-to-rule distances up to ρ_t, and the $(1 + \gamma_i)$ terms on the right for the inter-rule gaps from ρ_t to ρ_s. Therefore, any pairwise ratio can be expressed in terms of only the basis ratios as

$$\frac{\rho_s - \rho_t}{\rho_u - \rho_w} = \\ \frac{\gamma_1 \times \gamma_2 \times \ldots \times \gamma_{t-1}(1 + \gamma_t\{1 + \gamma_{t+1}[\ldots(1 + \gamma_{s-2})]\})}{\gamma_1 \times \gamma_2 \times \ldots \times \gamma_{w-1}(1 + \gamma_w\{1 + \gamma_{v+1}[\ldots(1 + \gamma_{u-2})]\})}.$$

This means that the order and relative location of any set of rectilinear lines can be captured by two basis-ratio sequences Γ_H and Γ_V. If, for example, the near-horizontal rulings have ρ-coordinates [120, 234, 890, 1242, and 1600] and the vertical rulings [210, 300, 540, and 890], then the corresponding sequences are

$$\Gamma_H = [(890 - 234)/(234 - 120), (1242 - 890)/ \\ (890 - 234), (1600 - 1242)/(1242/890)] \\ = [5.75, 0.54, \text{and } 1.02],$$

$$\Gamma_V = [(540 - 300)/(300 - 210), (890 - 540)/(540 - 300)] \\ = [2.67, 1.46].$$

For full-page rulings, the edges of the page (if detectable) can be included to ensure a minimum of three rulings of either orientation. A form with fewer rulings would not be much of a form in any case, but the null sequence is admissible for symbol matching.

5 Noisy ρ Coordinates

The location of the rulings is subject to at least two kinds of uncertainty. The first is caused by the limitation of precise location estimation from a finite number of pixels. The second is the inevitable random-phase spatial-sampling noise. We also observed variability due to the fact that forms from different print shops, or even from the same printer but different press runs, may be printed with slightly different settings. We have not, however, seen enough instances of multisource variability to attempt to model it.

5.1 *Measurement Error*

We compute the effect on the ruling gap ratios of i.i.d. Gaussian noise on the ruling locations ρ_i. The effect of this noise on the gap between two rulings is additive. We assume that the noise on a ruling at ρ_i, is $\varepsilon_i = N(0, \sigma_0)$, where $\sigma_0 \ll \rho_{i+1} - \rho_i$ (i.e., the variability is much less than the length of the gap). Therefore, the ruling gap in the numerator of a basis ratio can be considered a Gaussian random variable $X = (\rho_{i+2} + \varepsilon_{i+2}) - (\rho_{i+1} + \varepsilon_{i+1})$ distributed according to $N(\mu, \sigma)$, where $\mu = \rho_{i+2} - \rho_{i+1}$ and $\sigma^2 = 2\sigma_0^2$ (the variances of the independent noise variables at the ends of the gap add). The variable in the denominator of the ratio is $Y = (\rho_{i+1} + \varepsilon_{i+1}) - (\rho_i + \varepsilon_i)$, distributed according to $N(\eta, \sigma)$, where $\eta = \rho_{i+1} - \rho_i$ because adjacent gaps that share a ruling perturbed by ε_{i+1}, X and Y are correlated.

The distribution of the ratio of two correlated Gaussian variables is notoriously difficult to calculate because it requires several numerical integrations. We, therefore, use the approximation developed by Hinkley that guarantees a low error for our conditions.[44] To apply Hinkley's approximation, we must first derive the correlation coefficient $\rho_{X,Y}$ linking the numerator and the denominator of the basis ratio.

$$E[XY] = E\{[(\rho_{i+2} + \varepsilon_{i+2}) - (\rho_{i+1} + \varepsilon_{i+1})][(\rho_{i+1} + \varepsilon_{i+1}) - (\rho_i + \varepsilon_i)]\} \\ = E[(\rho_{i+2} - \rho_{i+1})(\rho_{i+1} - \rho_i)] - E[\varepsilon_{i+1}\varepsilon_{i+1}] = \sigma_0^2.$$

$$\text{Then } \rho_{X,Y} = \frac{\text{cov}[X, Y]}{\text{std}[X]\text{std}[Y]} = \frac{E[XY] - E[X]E[Y]}{\text{std}[X]\text{std}[Y]} \\ = \frac{\mu\eta - \sigma_0^2 - \mu\eta}{\sigma^2} = -\frac{1}{2},$$

because ε_i and ε_j are independent if $i \neq j$.

Now let $W = X/Y$. Then, the approximate CDF, pdf, and a bound on the approximation are

$$f^*(w) = \frac{b(w)d(w)}{\sqrt{2\pi}\sigma^2 a^3(w)}; \quad F^*(w) = \Phi\left[\frac{\eta w - \mu}{\sigma^2 a^2(w)}\right]; \\ |F(w) - F^*(w)| \leq \Phi\left(-\frac{\eta}{\sigma}\right),$$

where Hinkley's equations simplify, because the variances of X and Y are the same, to

$$a(w) = \frac{1}{\sigma}(w^2 + 2\rho_{X,Y} + 1)^{1/2},$$

$$b(w) = \frac{1}{\sigma^2}[\mu w - \rho_{X,Y}(\mu + \eta w + \eta^2)],$$

$$c = \frac{1}{\sigma^2}(\mu^2 - 2\rho_{X,Y}\mu\eta + \eta^2),$$

$$d(w) = \exp\left[\frac{b^2(w) - ca^2(w)}{2(1 - \rho_{X,Y}^2)a^2(w)}\right].$$

The error of approximation is the probability that the denominator is less than zero, which in our case is very small for values of $\sigma_0 \ll \eta = \rho_{i+1} - \rho_i$. Figure 3 shows the probability density function $f*(w)$ and the cumulative probability distribution function $F*(w)$ for values of $\mu = 8$, $\eta = 24$, and $\sigma = 1$ and 2, respectively (plotted at 0.05 intervals of w). Further exploration of the parameter space shows that for a fixed value of $(\mu + \eta)$ and of σ, the ratio of the standard deviation to the mean of the ratio is least when $\mu = \eta$.

5.2 *Random Phase Sampling Error*

The precise quantification of gap ratios, like that of all image features, is also hampered by the random-phase noise induced by the arbitrary placement of any document with respect to the scanner's or camera's sensor array. This noise can be reduced, but not eliminated, by increasing

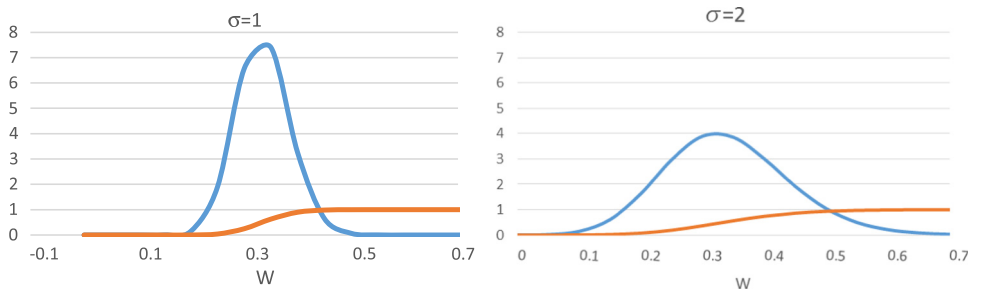

Fig. 3 Approximate pdfs and CDFs for a noisy gap ratio.

Fig. 4 Random-phase noise. Here, $L_1 = 4.2\delta$. After spatial sampling L_1 will be either three or four pixels long, depending on its position relative to the sampling grid.

the spatial sampling rate. Fortunately, the estimation of the variability of one-dimensional features (Fig. 4) like length is much simpler than that of two-dimensional features like area.[26]

The distances between rule edges are quantized to integer values by scanning. Consider gaps of length L_1 and L_2 sampled at δ-length intervals in Fig. 4. After sampling, L_1 will be of length $\lfloor L_1/\delta \rfloor$ or $\lfloor L_1/\delta \rfloor - 1$, and L_2 will be $\lfloor L_2/\delta \rfloor$ or $\lfloor L_2/\delta \rfloor - 1$. (Gap length is the number of background pixels minus 1). The ratio can take only one of three values, with their probabilities given as follows:

$$\text{Prob}(\lfloor L_1/\delta \rfloor - 1)/(\lfloor L_2/\delta \rfloor) = L_1 \bmod \delta,$$

$$\text{Prob}(\lfloor L_1/\delta \rfloor - 1)/(\lfloor L_2/\delta \rfloor - 1) = 1 - (L_1 \bmod \delta + L_2 \bmod \delta),$$

$$\text{Prob}(\lfloor L_1/\delta \rfloor)/(\lfloor L_2/\delta \rfloor - 1) = L_2 \bmod \delta.$$

In the worst case, (when, $L_1 \bmod \delta = L_2 \bmod \delta = 1/2$), the three possible values occur with probabilities of 0.25, 0.50, and 0.25. If random-phase sampling noise changes the mapping of any ratio to a symbol (Sec. 6), then identical rule configurations will result in different symbol strings and, therefore, in a nonzero edit distance between them. This is likely only for gaps of a few pixels.

6 Ratio Quantization, Edit Distance, and Classification

This paper focuses on the analytical developments of Secs. 3, 4, and 5, and it contains no new experimental results. In this section, we merely review a single experiment that was already presented in more detail.[41,42]

The pairs of sequences of horizontal and vertical ruling gap ratios of a set of noisy binarized forms were quantized and mapped into strings of symbols. The forms were then classified into preset classes by comparing the symbol strings of unknown forms with the symbols strings of reference forms from each class.

6.1 *Logarithmic Ratio Quantization*

Uniform quantization of the ratios—for edit distance computation—would map the prevalent near-unity ratios into few symbols. Logarithmic mapping of gap ratios to string symbols flattens the resulting symbol probability distribution. Therefore, the gap ratios γ are mapped into N bins k of size increasing away from that of the central bin for unity ratio:

$$k = F(\gamma \cdot K, N) = \min\left\{\max\left[\frac{(\log_{10}\gamma + K)(N-2)}{2K} + 1, 1\right], N\right\}.$$

The parameters N and K govern the logarithmic bin size. The domain of the mapping includes two semi-open intervals for very small and very large ratios (for $|\log \gamma| > K$).

The smallest gaps in a document typically correspond to the space required to print or write a word or a number. Even dense forms rarely have more than 30 lines of text; most forms have fewer than 20. The smallest gaps are likely to be those from double lines. The largest gap can be no larger than page height. Gap ratios typically range from 0.1 to 10, and the smallest significant difference is about 30%. Setting $N = 24$ and $K = 1.3$ yields 22 finite bins increasing by 30% from $\gamma = 0.05$ to $\gamma = 20$. The resulting symbol alphabet is {'1', '2', ..., '24'}.

6.2 *Edit Distance*

The metric used for classification was the Levenshtein edit distance. Schauerte's open-source program accepts arbitrary weights for the cost of the insertions, deletions, and substitutions necessary to convert one string into another, but lacking enough training data to estimate the optimal weights, we set them all equal. With more data, substitutions could be also weighted according to the size difference of the gap ratios. An example of the edit distances D_H and D_V between the horizontal and vertical basis ratio sequences (represented here by alphabetic symbols) of forms #54 and #55 is shown below:

		Assigned																
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	ERROR	TOTAL
	1	3															0	3
	2		23	2		1											3	24
	3		4	0		2											6	6
	4				37												0	37
	5					10											0	10
	6						6										0	8
True	7							4									0	4
	8								5								0	5
	9									5							0	5
	10		1			1					11						2	13
	11				1							3					1	4
	12					2	1	1					8				4	12
	13									1				0			1	1
	14														6		0	6
	15															20	0	20
		0	5	2	1	6	1	1	0	1	0	0	0	0	0	0	17	158

Fig. 5 Results from leave-one-out edit-distance-based classification of 158 MADCAT forms.

H54: ‘mdkkllilkkkkkkkkkkkkkkkkklkimks’ V54: ‘mikogog’
H55: ‘jlmgkmjiknhkmjljkkkkljkimjhlllmjh’ V55: ‘lhknemoe’

$$D_{\mathrm{H}}(54,55) = 24 \qquad D_{\mathrm{V}}(54,55) = 6.$$

The edit distance computation could take missing or spurious rules into account. When a symbol does not match, the algorithm can check whether combining adjacent gaps would reduce the edit distance. (A rule missed in one document is equivalent to a spurious rule in the other and can be analogously treated.) This check can be extended, at exponentially growing cost, to several consecutive gaps.

6.3 *Classification*

The rule detection, logarithmic gap ratio quantization, and string matching were applied as parts of the MADCAT project to a set of 158 extremely noisy scanned forms of 15 types. The sizes of the groups ranged from 1 to 37 forms. These filled-out forms contain still sensitive personnel information collected by Iraqi Government agencies, which precludes presenting them. The forms were classified by a nearest neighbor classifier with the edit distance function. The resulting error rate was 11% (17 errors). Ten errors are due to groups 3 and 12 (see Figs. 5 and 6). One error is unavoidable because group 13, with only one member, has no reference pattern for the nearest neighbor. There are six confusions between groups 2 and 3 that differ only by a single ruling. The MATLAB® program runs in 1 s per form on a 2 GHz laptop, with 83% of the time taken by the Hough transform.

7 Summary and Discussion

We explored the extraction and quantitative representation of rectilinear ruling configurations that might serve as features in some document image analysis tasks.

We explained, by formulating an approximation for the probability of error, the exponential increase in the effectiveness of sparse ruling location that can be achieved by simultaneous analysis of the distribution of the angular coordinates of all quasihorizontal and quasivertical ruling candidates. The method is simple and direct, yet apparently seldom utilized.

We proved that the sequence of the $N-2$ ratios of the distances between N consecutive parallel rulings completely characterizes the ruling configuration. In contrast, the histogram of the size of the ruling gaps or of their ratios fails to preserve their order.

The proposed basis ratios are least sensitive to measurement noise when their magnitude is near unity. Random-phase spatial quantization, on the other hand, significantly perturbs the ratios only when one of the gaps is small compared to the pixel size. The effect of this noise at a given spatial resolution can be reduced by grayscale or color conversion instead of binarization.

Since the variable-length sequences of rulings do not easily fit vector-space classification, we mapped them into symbol strings. The symmetric logarithmic transformation

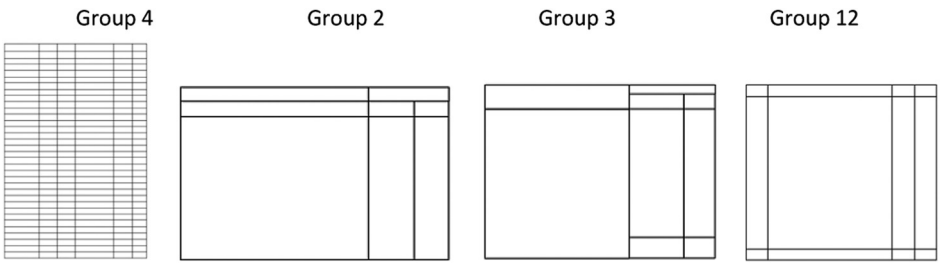

Fig. 6 Models of empty tables provided with the ground truth but not used for classification. There were no errors on tables with many rulings, like those in Group 4. There were six confusions between Groups 2 and 3, and 4 involving Group 12.

ensured equal relative quantization error regardless of ratio size.

A review of an earlier experiment shows only that the proposed processing chain can be readily implemented and that run time on a PC is fast enough for practical applications. The sample size was far too small to draw defensible conclusions regarding accuracy and generalizability. Except possibly for historical forms, there are not many examples where form classification by means of ruling configurations is the best alternative. Modern forms usually have a prominent alphanumeric or bar-coded form identifier, and in its absence OCR will usually reveal preprinted labels that are unique to each form class.

Predicting the actual error rate would require computing the joint probability distribution of all the ratios under the difficult constraint that each ratio is statistically dependent on the two preceding ratios. One possibility is the generalization of Chow's expansion.[45] One would first have to justify the i.i.d. assumption of the uniform angular distribution of the spurious rulings. The prediction will be even more problematic if the obvious possible improvements (e.g., variable weights and tracking down missing/extra rulings) in the edit-distance-based classification are implemented. None of these are feasible without a much larger labeled sample of ruled documents.

Although the analysis was prompted by a form-classification task, we have received suggestions for its potential applicability to other image analysis tasks with rectilinear line configuration, including building, street, and agricultural field location by remote sensing, conversion of hand-drawn legacy drawings and diagrams, staff line extraction in camera-based music score interpretation, and information recovery from scanned printed tables.

Acknowledgments

The author is grateful to Dr. Daniel Lopresti (Lehigh University) for cogent explanations of dynamic programming algorithms, for an introduction to the MADCAT perspectives on document interpretation, and for providing access to the experimental data. He is also indebted to Dr. Prateek Sarkar (Google) for suggestions on earlier attempts at analysis.

References

1. O. Hori and D. S. Doermann, "Robust table-form structure analysis based on box-driven reasoning," in *Proc. 3rd Int. Conf. on Document Analysis and Recognition (ICDAR'95)*, IEEE, Montréal, Canada (1995).
2. K. Itonori, "A table structure recognition based on textblock arrangement and ruled line position," in *Proc. 2nd Int. Conf. on Document Analysis and Recognition (ICDAR'93)*, pp. 765–768, IEEE, Tsukuba Science City, Japan (1993).
3. T. A. Bayer, "Understanding structured text documents by a model based document analysis system," in *Proc. 2nd Int. Conf. on Document Analysis and Recognition (ICDAR'93)*, pp. 448–453, IEEE, Tsukuba Science City, Japan (1993).
4. T. Watanabe et al., "Structure analysis of table-form document on the basis of the recognition of vertical and horizontal line segments," in *Proc. 1st Int. Conf. on Document Analysis and Recognition*, pp. 638–646, Association francaise pour la cybernetique economique et technique (AFCET), Paris (1991).
5. A. Dengel and G. Barth, "ANASTASIL: hybrid knowledge-based system for document layout analysis," in *Proc. 11th Int. Joint Conf. on Artificial intelligence*, pp. 1249–1254, Morgan Kaufman Publishers, Detroit, Michigan (1989).
6. K.-C. Fan, Y.-K. Wang, and M.-L. Chang, "Form document identification using line structure based features," in *Proc. 6th Int. Conf. on Document Analysis and Recognition, ICDAR '01*, IEEE Computer Society, Seattle, Washington (2001).
7. T. Kieninger and A. Dengel, "Applying the T-Recs table recognition system to the business letter domain," *Proc. 6th Int. Conf. on Document Analysis and Recognition. ICDAR '01*, pp. 704–708, IEEE Computer Society, Seattle, Washington (2001).
8. J. C. Handley, "Document recognition," Chapter 8 in *Electronic Imaging Technology*, E. R. Dougherty, Ed., SPIE Press, Bellingham (1999).
9. A. Dengel, "Towards understandable explanations for document analysis systems," in *10th IAPR International Workshop on Document Analysis Systems (DAS)*, M. Blumenstein, U. Pal, and S. Uchida, Eds, IEEE Computer Society, Queensland, Australia (2012).
10. A. K. Chhabra, V. Misra, and J. Arias, "Detection of horizontal lines in noisy run length encoded images: the FAST method," *Lec. Notes Comput. Sci.* **1072**, 35–48 (1996).
11. K. Tombre, "Analysis of engineering drawings: State of the art and challenges," *Lec. Notes Comput. Sci.* **1389**, 257–264 (1998).
12. Y. Yu, A. Samal, and S. Seth, "A system for recognizing a large class of engineering drawings," *IEEE Pattern Anal. Mach. Learn.* **19**(8), 868–890 (1997).
13. D. Dori and L. Wenyin "Automated CAD conversion with the machine drawing understanding system: concepts, algorithms, and performance," *IEEE Trans. Syst., Man, Cybern.* **29**(4), 411–416 (1999).
14. H. I. Koo and N. I. Cho, "Skew estimation of natural images based on a salient line detector," *J. Electron. Imaging* **22**(1), 013020 (2013).
15. B. Coüasnon and L. Pasquer, "A real-world evaluation of a generic document recognition method applied to a military form of the 19th century," in *Proc. 6th Int. Conf. on Document Analysis and Recognition*, pp. 779–783, IEEE Computer Society, Seattle, Washington (2001).
16. B. Coüasnon, "Recognition of Tables and Forms," in *Handbook of Document Image Processing and Recognition*, D. Doermann and K. Tombre, Eds., Springer, Berlin (2014).
17. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York (1973).
18. J. M. S. Prewitt, "Object enhancement and extraction," in *Picture Processing and Psychopictorics*, B. S. Lipkin and A. Rosenfeld, Eds., Academic Press, New York (1970).
19. K. R. Castleman, *Digital Image Processing*, Prentice Hall , Englewood Cliffs, New Jersey(1996).
20. M. S. Erkilinc et al., "Text, photo, and line extraction in scanned documents," *J. Electron. Imaging* **21**(3), 033006 (2012).
21. A. Amin et al., "Comparative study of skew detection algorithms," *J. Electron. Imaging* **5**(4), 443–451 (1996).
22. G. Nagy, "On the spatial autocorrelation function of noise in sampled typewritten characters," *1968 IEEE Region III Convention Record*, pp. 7.6.1–7.6.5, New Orleans, Louisiana (1968).
23. J. Zhou and D. Lopresti, "Repeated sampling to improve classifier accuracy," in *Proc. IAPR Workshop on Machine Vision Applications*, pp. 346–351, International Association for Pattern Recognition, Kawasaki, Japan (1994).
24. D. I. Havelock, "Geometric precision in noise-free digital images," *IEEE Trans. Pattern Anal. Mach. Intell.* **11**(10), 1065–1075 (1989).
25. D. I. Havelock, "The topology of locales and its effect on position uncertainty," *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(4), 380–386 (1991).
26. P. Sarkar et al., "Spatial sampling of printed patterns," *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(3), 344–351 (1998).
27. H. S. Baird, "The state of the art in document image degradation modeling," in *Digital Document Processing*, B. B. Chaudhuri, Ed., pp. 261–279, Springer, Verlag (2007).
28. E. B. Smith, "Characterization of image degradation caused by scanning," *Pattern Recogn. Lett.* **19**(13), 1191–1197 (1998).
29. V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Dokl. Akad. Nauk SSSR* **163**(4), 845–848 (1965).
30. R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *J. ACM* **21**(1), 168–173 (1974).
31. V. I. Levenshtein, "Universal bounds for codes and designs," in *Handbook of Coding Theory*, V. S. Pless and W. C. Huffman, Eds., Elsevier, Amsterdam (1978).
32. P. A. V. Hall and G. R. Dowling, "Approximate string matching," *ACM Comput. Surv.* **12**(4), 381–402 (1980).
33. D. Sankoff and J. B. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison Wesley, Reading, Massachusetts (1983).
34. G. Navarro, "A guided tour to approximate string matching," *ACM Comput. Surv.* **33**(1), 31–88 (2001).
35. A. Marzal and E. Vidal, "Computation of normalized edit distance and applications," *IEEE Trans. Pattern Anal. Mach. Intell.* **15**(9), 926–932 (1993).
36. M. Neuhaus and H. Bunke, "Edit distance-based kernel functions for structural pattern classification," *Pattern Recogn.* **39**, 1852–1863 (2006).
37. B. Schauerte and G. A. Fink, "Focusing computational visual attention in multi-modal human-robot interaction," in *Proc. 12th Conf. on Multimodal Interfaces, ICMI*, ACM, Beijing, China (2010).

38. National Institute of Standards and Technology (NIST) "Multilingual automatic document classification and translation evaluation (MADCAT)," (3December2010), http://www.nist.gov/itl/iad/mig/madcat.cfm (8 January 2014).
39. B. P. Montgomery, "The Iraqi secret police files: a documentary record of the Anfal genocide," *Archivaria* **52**, 81–82 (2001).
40. B. P. Montgomery, "Returning evidence to the scene of the crime: why the Anfal files should be repatriated to Iraqi Kurdistan," *Archivaria* **69**, 143–171 (2010).
41. G. Nagy and D. Lopresti, "Form similarity via Levenshtein distance between ortho-filtered logarithmic ruling-gap ratios," *Proc. SPIE* **9021**, 902106 (2013).
42. G. Nagy, "On parallel lines in noisy forms," *Lec. Notes Comput. Sci.* **8621**, 173–182 (2014).
43. M. Raab and A. Steger, "Balls into bins – a simple and tight analysis," *Lec. Notes Comput. Sci.* **1518**, 159–170 (1998).
44. D. V. Hinkley, "On the ratio of two correlated normal random variables," *Biometrika* **56**(3), 635–639 (1969).
45. C. K. Chow, "A recognition method using neighbor dependence," *IRE Trans. Electron. Comput.* **EC-11**, 683–690 (1962).

George Nagy is an emeritus professor of computer engineering at RPI. He received his BEng and MEng degrees from McGill University in 1959 and 1960, and his PhD degree on neural networks from Cornell University in 1962. He is the author of more than 100 journal papers and book chapters, and of a dozen SPIE conference papers. His current research includes document image processing, data extraction from web tables, and classification of Chinese calligraphy. He is a fellow of the IEEE and of the IAPR.