G. Nagy, On Parallel Lines in Noisy forms, Procs. IAPR SS&SPR, Juunsu, Finland, Fränti et al. (Eds.): S+SSPR 2014, LNCS 8621, pp. 173–182, 2014.

On Parallel Lines in Noisy Forms

George Nagy

Electrical, Computer and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY, USA
nagy@ecse.rpi.edu

Abstract. Quantification of the rectilinear configuration of typeset rules (lines) opens the way to form classification and content extraction. Line detection on scanned forms is often accomplished with the Hough transform. Here it is followed by simultaneous extraction of the dominant perpendicular sets of extracted lines, which ensures rotation invariance. Translation and scale invariance are attained by using minimal horizontal and vertical sets of distance ratios ("rule gap ratios") instead of rule-edge locations. The ratios are logarithmically mapped to an alphabet so that the resulting symbol strings can be classified by. edit distance. Some probability distributions associated with these steps are derived. Analytical considerations and small-scale experiments on scanned forms suggest that this approach has potential merit for processing degraded forms.

Keywords: forms, tables, rules, distance ratio, rotation invariance, scale invariance, random-phase noise, edit distance.

1 Introduction

Many documents exhibit an isothetic configuration consisting of orthogonal sets of parallel components. Line segments are explicit in ruled tables and forms, and implicit in parallel rows of text and justified margins and gutters. Rectilinear structures are also common in artifacts like cultivated fields, cities, buildings and machines: in fact, their presence is one of the prime clues for distinguishing man-made from natural. Although parallel lines play a role in other image processing and computer vision tasks as well, here we address only scanned or photographed form images. Fig. 1 shows examples of forms that offer a rich line structure but may have been scanned at too low resolution or are too noisy for OCR-based classification.

The "near-horizontal" lines shown in Fig. 1 were extracted by the Hough transform in rho-theta format. The line configurations include both rectilinear *rules* (the printing and publishing term for typeset lines), and s*purious lines* induced by accidental alignments of diverse page content. The images display various rule configurations, with the members of each class sharing essentially the same rule configuration but exhibiting different spurious lines. The task at hand is classifying new images into predefined classes. Since the forms are captured by a scanner or a camera, their position, scale and skew angle within the image are unknown.

P. Fränti et al. (Eds.): S+SSPR 2014, LNCS 8621, pp. 173–182, 2014.

The ratios of the distances between pairs of rules (*rule gap ratios*) are geometrically invariant features. (Invariant features are more commonly used in scene image analysis than in document recognition.) The ordered sets of horizontal and vertical ratio values are converted to a pair of symbol strings that characterize the ruling configuration of the underlying form. The forms are then classified according to the (1,1,1) edit distance between new images and existing class representative. So we

1. Distinguish isothetic rules from spurious lines formed by accidental alignments;
2. Compute the minimum set of algebraically independent rule gap ratios;
3. Map the ordered horizontal and vertical gap ratios into symbol strings;
4. Classify the unknown images based on the edit distance between symbol strings.

In the following sections we review prior work, examine each of the above steps, and give an example of their application to a set of degraded and mutilated form images.

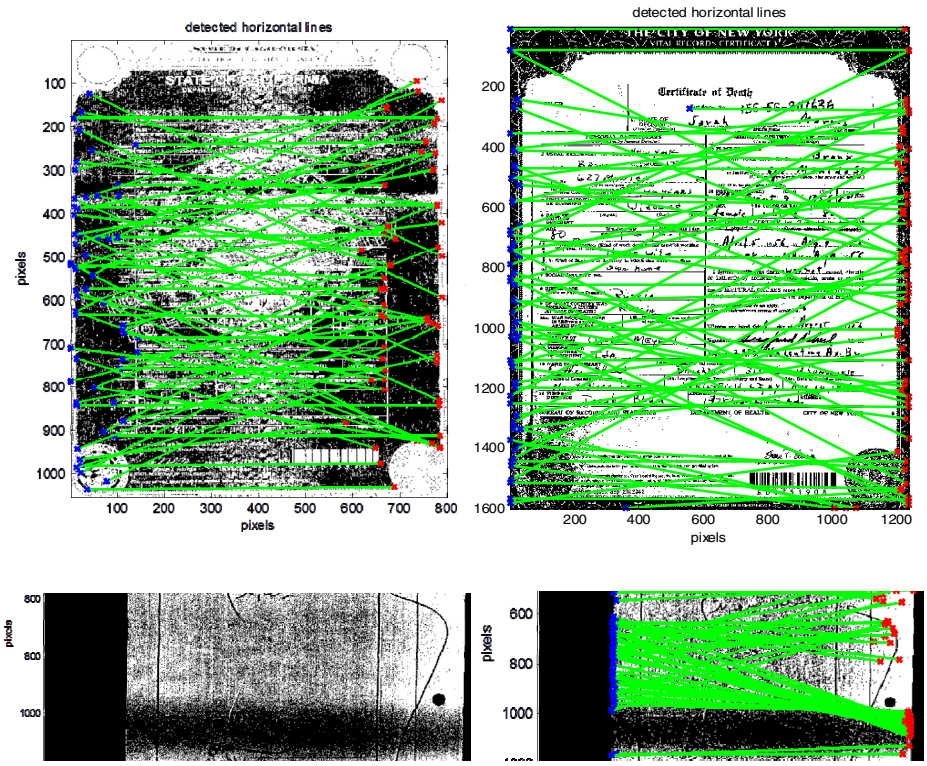

Fig. 1. Examples of forms with explicit isothetic rule structure. The two forms on top are from web archives. The partial form image on the bottom is from our classification experiment. Here only lines (shown in green) within ±30° of horizontal are extracted.

2 Prior Work

Line segment recognition has been steadily improved during the last three decades as part of table interpretation, form processing, and engineering drawing analysis. Historical form analysis became popular even as most contemporary forms migrated to the web. The Hough transform 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 seventies [1]. It does not require edge linking and is therefore often preceded only by edge extraction with the venerable Prewitt filter [2]. Other 3×3 pixel edge filters (Sobel, Roberts) yield similar results. 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.

Our interest in spatial sampling noise was triggered by peaks in the autocorrelation function corresponding to opposite stroke edges in scanned character images [3]. The variation (noise!) due to repeated scanning was exploited by Zhou and Lopresti to decrease OCR error [4]. Random-phase sampling noise was systematically investigated in remote sensing [5,6] and in scanned documents [7], but pixel jitter is usually modeled as if it were independent random displacement of sensor elements [8]. The relationship between spatial and amplitude quantization in scanning was explored thoroughly by Barney Smith [9].

Levenshtein introduced the edit distance for error-correcting codes in 1965 [10]. The optimal Wagner-Fischer algorithm was published a decade later [11]. Many variations of the original algorithms have appeared since then [12,13,14]. 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 [15] and kernel techniques for embedding the edit distance into a vector space [16]. The public-domain EDIT DISTANCE WEIGHTED program that we use was posted in 2010 by B. Schauerte [17].

The current study was initiated during a phase of the MADCAP project [18] concerned with categorization of a small subset of the collection of Kurdish documents recovered during the Anfal uprising [19,20]. The Hough transform parameters and preliminary results on classification of some degraded forms were presented at the 2014 SPIE Conference on Document Recognition and Retrieval [21].

3 Orthogonal Line Extraction

The accidental alignments of handwriting, stamps, binder holes, checkmarks and other non-rule pixels may give rise to far more spurious lines than the number actual rules on the page (Fig. 2). Since in contrast to the randomly distributed spurious lines all the nominally horizontal (or vertical) rules have the same angle, an obvious way to distinguish them is to histogram all the line angles. Then the lines in the most populated bin will be the rules. This stratagem fails only if too many spurious lines fall into some other bin. Below we calculate the dominant term of the probability of such an event as a proxy for the actual probability.

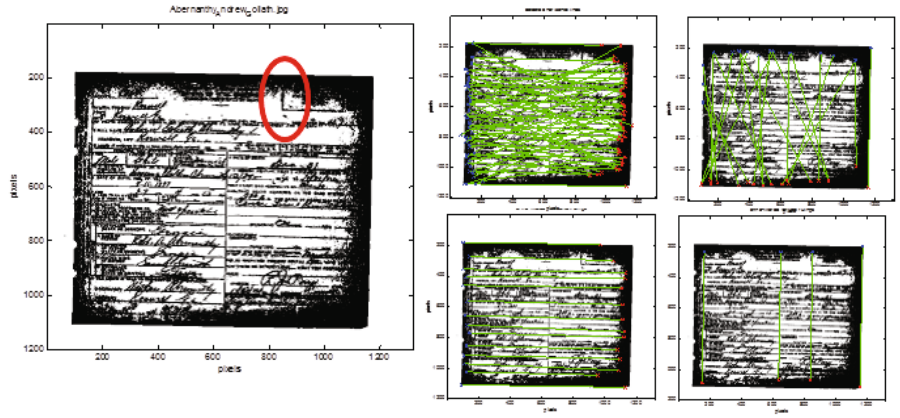

Fig. 2. A low-resolution, noisy and skewed death certificate. Near-horizontal and near vertical lines extracted by the Hough Transform and rules retained by orthogonal filtering.

Extreme skew is unlikely, therefore only lines within ±20° of the nominal and x- and y-axes need be extracted. Let there be R rules and S spurious lines on a page. Their angles are sorted into a histogram with N uniformly spaced bins ($N > R+S$). The rules are parallel and therefore fall into the same bin, but the skew detection will be incorrect if R or more of the spurious lines fall into some other bin. Under a (questionable!) i.i.d. assumption, the most probable such case is that R of the S spurious lines fall into a single bin and that each of the others occupies one bin. This can happen in as many ways as there are of picking single-occupancy bins. Therefore a lower bound on the probability that at least R of the S spurious lines fall into the same bin is:

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

Table 1. Dominant term of the probability of false maxima in the angle histogram

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

The shaded cells of Table 1 show that while the probability of a false maximum for 3 rules and 6 spurious lines is at least an appreciable 3.9%, doubling the number of lines reduces the dominant term to 0.01%. This can be achieved by adding 90° to the theta coordinate of every line within 20° of the vertical axis and histogramming all the line angles together. In the image of Fig. 2, every visible vertical rule is found,

including the edge of the box at the top right of the form marked with a red oval, with no false positives. Simultaneous identification of orthogonal lines pays off.

4 Rule Gap Ratios

No further use of the theta coordinates is made. The computation of the rule gap ratios requires only sorting the Hough rho coordinates of each set of extracted and ortho-filtered parallel lines and subtracting them pairwise to find the successive horizontal and vertical edge-to-edge rule gaps. Given N parallel rules, there are $O(N^2)$ pairs of rules and $O(N^4)$ possible ratios. It is clear, however, that there cannot be more than *N-2* algebraically independent ratios from which the value of all the others can be calculated. We choose as *basis ratios* the ratios of consecutive gaps, defined for horizontal or vertical lines located at $x_1, x_2, \ldots, x_i, \ldots, x_N$ (w.r.t. an arbitrary origin) as:

$$R_i = \left(x_{i+1} - x_i\right) / \left(x_{i+2} - x_{i+1}\right)$$

There are *N-2* such ratios, and any other ratio of line segments can be recovered from them. The proof is conceptually simple but notationally tedious, so we give an example instead. Let the three distances between four lines be *a*, *b*, and *c* (Fig. 3).

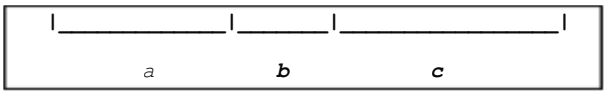

Fig. 3. Ratios of rule gaps

The two basis ratios are $R_1 = a/b$, and $R_2 = b/c$. An arbitrary ratio such as $(b+c)/(a+b)$ can be expressed in terms of the basis ratios as:

$$\frac{b+c}{a+b} = \frac{1}{(a/b)+1} + \frac{1}{(b/c)(a/b+1)} = \frac{1}{R_1+1} + \frac{1}{R_2(R_1+1)}$$

The general formula that proves the sufficiency of the basis ratios is:

$$\frac{x_s - x_t}{x_u - x_v} = \frac{\left(R_1 \times R_2 \times \ldots \times R_{t-1}\right)\left(1 + R_t\left[1 + R_{t+1}\left[\ldots\left[1 + R_{s-1}\right]\right]\right]\right)}{\left(R_1 \times R_2 \times \ldots \times R_{v-1}\right)\left(1 + R_v\left[1 + R_{v+1}\left[\ldots\left[1 + R_{u-1}\right]\right]\right]\right)}$$

The rule configuration of a page is preserved by the two sets of translation, scale and rotation invariant basis ratios. Lines are considered to be of infinite extent. If end-point information is required, it is kept separately. The accuracy of the rule gap ratios is affected by edge location variability and by random-phase sampling noise.

4.1 Edge Location Variability

Some applications must cope with forms reprinted at different times and by different printers. Even if the variability of the line and line-edge locations as a fraction of page

size is small, it may have a significant effect on the gap ratios. Each gap ratio is a function of the position of three (parallel) rules. What is the probability density function (pdf) of the ratio as a function of the variability of the edges?

The only line-segment ratio we found discussed in the literature is that resulting from of splitting a unit-length line segment by a uniformly distributed point L, which results in ratio $W = L/(1-L)$ [22]. The probability density of W,

$$f(w) = 1/(1+w)^2,$$

is skewed because its range is zero to infinity but its mean must be 0.5.

We extended the calculation of the pdf of $W = L/(1-L)$ to two *independent* (non-adjacent) gaps of lengths L_1 and L_2 distributed uniformly: $L_1 \in x_0 \pm a$ and $L_2 \in y_0 \pm a$. The resulting piecewise rational polynomial functions provide further insight:

$$f(w) = 0 \qquad \text{if } w \le \frac{x_0 - a}{y_0 + a} \text{ or } \frac{x_0 + a}{y_0 - a} < w;$$

$$f(w) = \frac{1}{8a^2}\left[(y_0 + a)^2 - \left(\frac{x_0 - a}{w}\right)^2\right] \qquad \text{if } \frac{x_0 - a}{y_0 + a} < w \le \frac{x_0 - a}{y_0 - a};$$

$$f(w) = \frac{y_0}{2a} \qquad \text{if } \frac{x_0 - a}{y_0 - a} < w \le \frac{x_0 + a}{y_0 + a};$$

$$f(w) = \frac{1}{8a^2}\left[\left(\frac{(x_0 + a)}{w}\right)^2 - (y_0 - a)^2\right] \qquad \text{if } \frac{x_0 + a}{y_0 + a} < w \le \frac{x_0 + a}{y_0 - a}$$

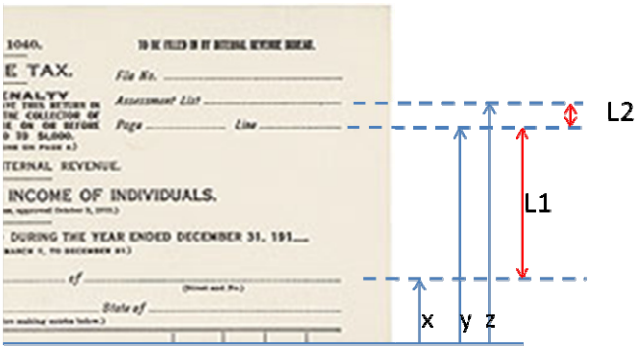

Fig. 4. The rule edges x, y, and z are uniformly distributed over a range of 2a. What is the pdf of the gap ratio $L_1/L2_?$?

What we must consider, however, is the more difficult three-variable case of a basis ratio formed by three *adjacent* edges located at *x, y,* and *z,* where x is uniformly and independently distributed over $x_0 \pm a$, *y* over $y_0 \pm a$, and *z over* $z_0 \pm a$ (Fig. 4). The gaps are $L_1 = y\text{-}x$ and $L_2, = z\text{-}y$. The basis ratio is $W = L_1/L_2$, as in Fig. 3,

The gap lengths L_1 and L_2 are the difference of uniformly and independently distributed variables and therefore have a simple triangular distribution centered on the mean difference. But analytical formulation of the joint pdf of L_1 and L_2 is complicated by the statistical dependence induced by the shared edge y. After deriving the lengthy formula we must still resort to simulation to compute the pdf of the ratio W.

The effect on the ratio of edge variability is illustrated in Fig. 5 for $x_0 = 1$, $y_0 = 4$, $z_0 = 10$, and three values of a. Large values of a correspond to high rule edge variability. W ranges from $(y_0\text{-}x_0\text{-}2a)/(z_0\text{-}y_0\text{+}2a)$ to $(y_0\text{-}x_0\text{+}2a)/(z_0\text{-}y_0\text{-}2a)$. As a approaches zero, the distribution converges to a delta function located at the nominal value of the ratio.

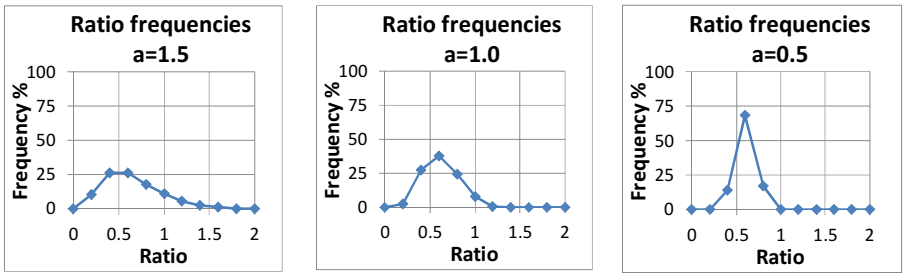

Fig. 5. Frequency distribution of gap ratio between variable edge locations

4.2 Random-Phase Sampling Noise

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 the spatial sampling rate.

The distances between rule edges are quantized to integer values by scanning. As a one-dimensional analogy, consider rule gaps of length L_1 and L_2 sampled at δ-length intervals (Fig. 6). 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 minus1.) The ratio can take only one of three values: $(\lfloor L_1/\delta \rfloor -1)/(\lfloor L_2/\delta \rfloor)$, $(\lfloor L_1/\delta \rfloor -1)/(\lfloor L_2/\delta \rfloor -1)$, and $(\lfloor L_1/\delta \rfloor)/(\lfloor L_2/\delta \rfloor -1)$. In the worst case, when, $L_i = \lfloor L_i \rfloor + \delta/4$,: the three possible values occur with probabilities of 0.25, 0.50, 0.25. If random-phase sampling noise changes the mapping of any ratio to a symbol (cf. §5), then identical rule configurations will result in different symbol strings and therefore in non-zero edit distance between them.

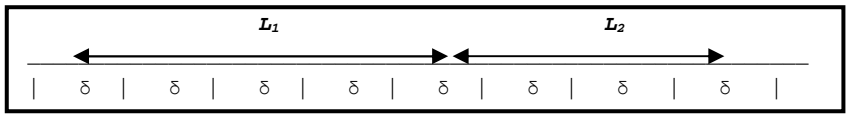

Fig. 6. Random-phase noise. Here $L_1 = 4.2\delta$. After spatial sampling L_1 will be either 3 or 4 pixels long, depending on its position relative to the sampling grid.

5 Ratio Quantization and Edit Distance

The smallest gaps in a document typically correspond to the space required to print or write a word or a number. Even densely printed documents have no more than 60 lines of text; most forms have fewer than 20. The smallest gaps are likely to be those from double rule dashes. 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%.

Uniform quantization of the ratios – for edit distance computation – would map the prevalent near-unity ratios into very few symbols. Logarithmic mapping of gap ratios to string symbols flattens the resulting symbol probability distribution. Therefore gap ratio R is mapped into bin k, where k ranges from 1 to N:

$$k = F(R;K,N) = \min\left(\max\left(\left\lceil \frac{(\log_{10} R + K)(N-2)}{2K} \right\rceil + 1, \quad 1 \right), \quad N \right)$$

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 $|\log R| > K$). Setting N =24 and K=1.3 yields 22 finite bins increasing by 30% from R=0.05 to R = 20. The resulting symbol alphabet is {'1', '2', …, '24'}.

The metric used for classification is 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.

The edit distance computation could take into account missing or spurious rules. 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 treated analogously.) This check can be extended, at exponentially growing cost, to several consecutive gaps.

6 Plausible Applications

Deteriorated and poorly-scanned forms abound in historical census, military and municipal records. Some of the recent interest in such documents is due to genealogical research (including its medical implications). Even contemporary forms may be degraded by repeated photocopying, reduced resolution for web display, or batch scanning with a page-feed scanner without adequate skew and binarization control.

Modern form identification is generally based on a barcode or some Form Identification Number (FIN) prominently printed at the top or near one of the corners. In their absence, OCR'd forms can be identified using preprinted text specific to each type of form. Both the FIN and the preprinted labels usually exhibit enough redundancy to tolerate OCR errors. The ruling-based classification discussed here is appropriate only for forms that cannot be OCR'd and have an isothetic rule structure without too many other aligned edges. In principle the method could be applied hierarchically, possibly via the quad tree [23], to forms with highly localized rules.

The rule detection, logarithmic gap ratio quantization and string matching were applied as part of the MADCAT project to a set of 158 extremely noisy scanned forms of 15 types (Fig. 7). These filled-out forms contain personnel information collected by Iraqi government agencies and regrettably only redacted or partial images can be presented. 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. One error is unavoidable because Group 13, with only one member, has no reference pattern for Nearest Neighbor. There are 6 confusions between groups 2 and 3 that that differ only by a single ruling. The Matlab program runs in 1 second per form on a 2 GHz laptop, with 83% of the time taken by the Hough transform.

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

7 Envoy

In the expectation of future large-scale endeavors on degraded but rule-rich corpora, we examined some benefits and drawbacks of three related ideas:

- Simultaneous orthogonal filtering of Hough lines to eliminate of spurious lines.
- Extracting gap ratios of parallel rules for geometric invariance.
- Classifying the ratios by edit distance, bridging statistical and structural methods.

Acknowledgment. The author thanks Prof. Daniel Lopresti (Lehigh University) for access to the MADCAT data and for suggestions on edit distance computation. He is also grateful for the close reading of the manuscript and recommendations of Dr. Prateek Sarkar (Google, Inc.) and of one of the referees.

References

1. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley (1973)
2. Prewitt, J.M.S.: Object Enhancement and Extraction. In: Lipkin, B.S., Rosenfeld, A. (eds.) Picture Processing and Psychopictorics. Academic Press (1970)

3. Nagy, G.: On the Spatial Autocorrelation Function of Noise in Sampled Typewritten Characters. In: 1968 IEEE Region III Convention Record, New Orleans, United States, pp. 7.6.1–7.6.5 (1968)
4. Zhou, J., Lopresti, D.: Repeated Sampling to Improve Classifier Accuracy. In: Proc. IAPR Workshop Machine Vision Applications, Kawasaki, Japan, pp. 346–351 (1994)
5. Havelock, D.I.: Geometric Precision in Noise-Free Digital Images. IEEE Trans. Pattern Analysis and Machine Intelligence 11(10), 1,065–1,075 (1989)
6. Havelock, D.I.: The Topology of Locales and Its Effect on Position Uncertainty. IEEE Trans. Pattern Analysis and Machine Intelligence 13(4), 380–386 (1991)
7. Sarkar, P., Lopresti, D., Zhou, J., Nagy, G.: Spatial Sampling of Printed Patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(3), 344–351 (1998)
8. Baird, H.S.: The State of the Art in Document Image Degradation Modeling. In: Chaudhuri, B.B. (ed.) Digital Document Processing, pp. 261–279. Springer (2007)
9. Barney Smith, E.: Characterization of image degradation caused by scanning. Pattern Recognition Letters 19(13), 1191–1197 (1998)
10. Levenshtein, V.I.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Doklady Akademii Nauk SSSR 163(4), 845–848 (1965)
11. Wagner, R.A., Fischer, M.J.: The String-to-String Correction Problem. Journal of the ACM 21(1), 168–173 (1974)
12. Hall, P.A.V., Dowling, G.R.: Approximate String Matching. ACM Computing Surveys 2(4), 381–402 (1980)
13. Sankoff, D., Kruskal, J.B.: Time warps, string edits, and macromolecules: The theory and practice of sequence comparison. Addison Wesley (1983)
14. Navarro, G.: A guided tour to approximate string matching. ACM Computing Surveys 33(1), 31–88 (2001)
15. Marzal, A., Vidal, E.: Computation of Normalized Edit Distance and Applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(9) (September 1993)
16. Neuhaus, M., Bunke, H.: Edit distance-based kernel functions for structural pattern classification. Pattern Recognition 39, 1852–1863 (2006)
17. Schauerte, B., Fink, G.A.: Focusing Computational Visual Attention in Multi-Modal Human-Robot Interaction. In: Proc. ICMI (2010)
18. Multilingual Automatic Document Classification and Translation Evaluation (MADCAT), `http://www.nist.gov/itl/iad/mig/madcat.cfm` (accessed August 1, 2014)
19. Montgomery, B.P.: The Iraqi Secret Police Files: A Documentary Record of the Anfal Genocide. Archivaria 52, 81–82 (2001)
20. Montgomery, B.P.: Returning Evidence to the Scene of the Crime: Why the Anfal Files Should be Repatriated to Iraqi Kurdistan. Archivaria 69, 143–171 (2010)
21. Nagy, G., Lopresti, D.: Form similarity via Levenshtein distance between ortho-filtered logarithmic ruling-gap ratios. In: SPIE/IST Document Recognition and Retrieval (February 2014)
22. Pickover, C.A.: The Problem of the Bones. In: The Mathematics of Oz: Mental Gymnastics from Beyond the Edge, ch. 8. Cambridge University Press, New York (2002)
23. Samet, H.: The Design and Analysis of Spatial Data Structures. Addison-Wesley, Reading (1990)