INTER-LINE DISTANCE ESTIMATION AND TEXT LINE EXTRACTION FOR UNCONSTRAINED ONLINE HANDWRITING

EUGENE H. RATZLAFF

IBM T. J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, USA
E-mail: ratzlaff@us.ibm.com

Methods for detecting and extracting whole text lines from unconstrained online handwritten text are described. The general approach is a "bottom-up" clustering of discrete strokes into small groups that are then merged into isolated lines of text. Initial clustering of strokes into groups is based on combined temporal and spatial stroke proximity. Spatial stroke proximity is gauged relative to estimated inter-line distance and mean character height. Two methods applicable to off-line or on-line data are described for estimating the inter-line distance: autocorrelation (self-convolution) of the Y-axis projection histogram, and a fitting function. Inter-line distance is accurately determined for 99% of all text pages. Text line extraction accuracy on letters (correspondence) is 98.7% and on tables is 94.9%.

1 Introduction

One of the first steps often required in applying machine recognition to unconstrained handwriting is the identification and extraction of each line of text from among other lines. The purpose of text line extraction is to prepare data to meet the requirements of succeeding processing steps such as size normalization, word segmentation, or feature extraction. These steps typically require the data to be no more than a single row of characters. The goal of text line extraction is to assign correctly each stroke or component to its appropriate text line so that each isolated line may be passed in turn to the following analysis stage. The task is made difficult by the fact that data frequently contain undulations and shifts in the baseline, baseline skew, baseline-skew variability, character-size variability, sparse data, skipped lines and inter-line distance variability. Several of these issues have been addressed in much of the recent work focused on text line extraction for offline data [2,4,8].

Our interest has been in the area of recognition of unconstrained, online handwriting, particularly in support of personal digital notepad (PDN) devices, the IBM Ink Manager™ application, and the IBM Ink Software Development Kit (SDK) [11]. A PDN is a portable digitizer-and-pen device that electronically records the pen strokes while the user writes on a standard paper notepad. Each ink stroke is recorded as a sequence of (X, Y) points indexed in time. The primary role for the IBM unconstrained, online handwriting recognizer [5,9,10] in this context is delayed recognition of this dynamically recorded data for keyword and note-taking

In: L.R.B. Schomaker and L.G. Vuurpijl (Eds.), Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, September 11-13 2000, Amsterdam, ISBN 90-76942-01-3, Nijmegen: International Unipen Foundation, pp 33-42

the result using an error count of 1 for each of the following operations: word deletion, word replacement, word insertion, text line deletion.

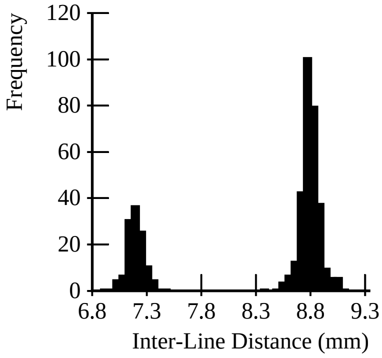


Figure 5. Histogram distribution of 436 correctly estimated inter-line distance estimates (one correct estimate of 17.4mm for a completely double-spaced Table page is excluded).

Table 1. Breakdown of data and errors attributable to text line clustering on Letter and Table data sets at the page, text line, and word level. Error rates are given as percentage of total.

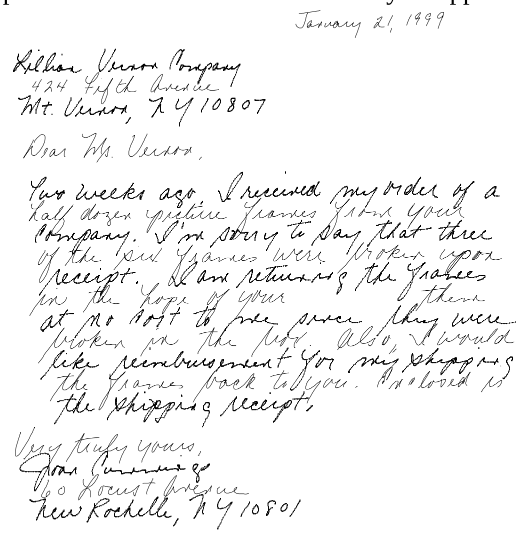
	Pages		Text Lines			Words	
	Total	Flawed	Total	Lines Merged	Lines Split	Total	Error
Letters	114	16 14%	1608	6 0.4%	15 0.9%	10090	42 0.4%
Tables	327	118 36%	5291	39 0.7%	231 4.4%	16121	373 2.3%

4 Discussion and Future Developments

The overall performance of the system meets most expectations (Table 1; Figure 6), however, several problems were observed. Significant skew was found in several instances. As seen in Figure 7b, a mere 2° line skew can shift words at either end of a text line by one inter-line space, resulting in significant text line splitting. Despite the use of lined paper, undulate text and somewhat irregular line spacing is observed (Figure 7c); some of these irregularities apparently occurred when the paper occasionally shifted on the PDN. This was especially true when writers entered lists into tables or appended dates to the top of a block of text. Reliable inter-line distance estimates were made despite skew and shift problems; however, text line extraction performance occasionally degraded.

The application of spatiotemporal information is clearly advantageous for the accurate partitioning of dots, short strokes, delayed strokes and diacritical marks that are temporally bounded by other strokes in their own text line (Figure 6). Many of the errors that were observed occurred when the misassigned stroke was temporally at the beginning or end of a text line and the spatiotemporal information was ambiguous. This frequently occurred with short strokes landing high above or well below the baseline, especially at the start or end of a text line, as with the comma in phrases such as *Dear Tom,* or *Best regards,*. In Figure 7d, for example, four commas were inserted after four city names were written, resulting in a line split. Similar errors were more common in the Tables, where spatiotemporally adjacent strokes were sparse, as with the numbers in the right-most column of Figure 4, or where there were significant gaps between table entries in different columns.

Although several of the errors occurred with visually ambiguous data, there appears to be room for significant improvement. Skew detection and correction should be considered. Reliable inter-line distance estimates can be used to place upper bounds on the number of lines possible, enabling the detection and correction of gross splitting errors (Figures 7b, 7c). In addition, by replacing the many programmatically-tuned heuristics involved in the clustering process with a more probabilistic approach based on training data, the bottom-up clustering assignments should improve. The application of generalized projections [6] and local extrema points [4,8] hold promise for improved character height estimation and baseline estimation; these improved or added features could likely be applied to advantage.



January 21, 1999

Lillian Vernon Company
424 Fifth Avenue
Mt. Vernon, NY 10807

Dear Ms. Vernon,

Two weeks ago, I received my order of a
half dozen picture frames from your
Company. I'm sorry to say that three
of the six frames were broken upon
receipt. I am returning the frames
in the hope of your there
at no cost to me since they were
broken in the box. Also, I would
like reimbursement for my shipping
the frames back to you. Enclosed is
the shipping receipt.

Very truly yours,
Joan Cummings
60 Locust Avenue
New Rochelle, NY 10801

Figure 6. Extraction results for page with appreciably overlapping text lines. Extracted text lines are shown alternately with normal and bold points from top; each of 20 lines was correctly extracted.

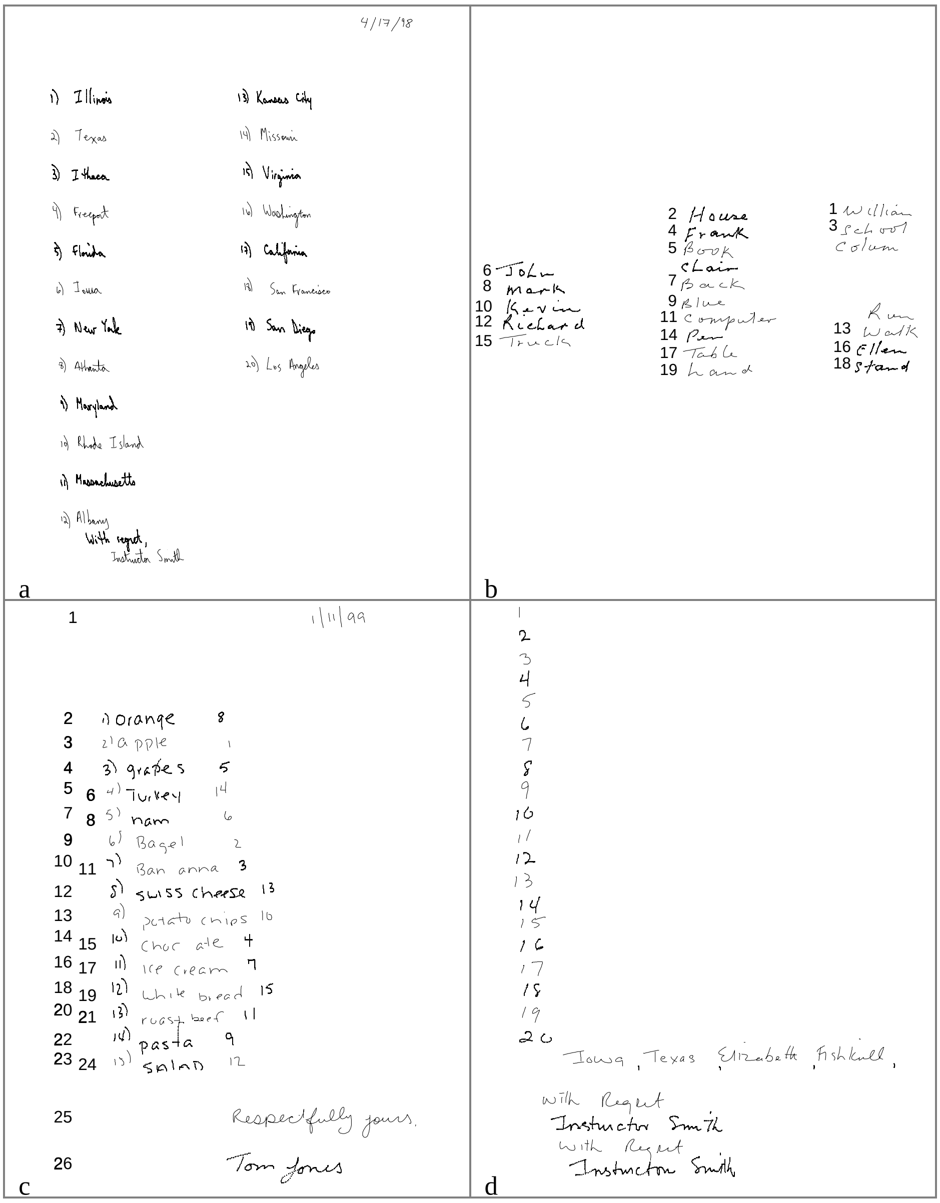


Figure 7. Four results with extracted text lines denoted by alternating normal and bold points (with numbers in **b**, **c**) from top: (**a**) mainly double-spaced, (**b**) 2° baseline skew, (**c**) undulating baseline, and (**d**) delayed comma insertion. Lines found / actual lines: (**a**) 15/15 (**b**) 19/10 (**c**) 26/18 (**d**) 26/25.

5 Acknowledgements

The author thanks Millie Miladinov, Michael Perrone, John Pitrelli, and Jay Subrahmonia of the IBM Pen Technologies Group for their contributions.

References

1. Bozinovic R. and Srihari S., Off-line cursive script word recognition, *IEEE Trans on PAMI*, **11** (10), January 1989, pp. 68-83.
2. Bruzzone E. and Coffetti M. C., An algorithm for extracting cursive text lines, *Proceedings of the Fifth International Conference on Document Analysis and Recognition* ICDAR '99 Bangalore, India 20-22 Sept. 1999, pp. 749-752.
3. Ernest L. D., Machine recognition of cursive script, *Proc. IFIP Congress*, **62**, 1992, 462-466.
4. Kim G., Govindaraju V. and Srihari S. N., Architecture for handwritten text recognition system, *Proceedings 6th IWFHR*, Taejon, Korea, August 1998, pp. 113-122.
5. Nathan K. S., Beigi H. S. M., Subrahmonia J., Clary G. J. and Maruyama H., Real-time on-line unconstrained handwriting recognition using statistical methods, *Proceedings of ICASSP 95: IEEE International Conference on Acoustics, Speech, and Signal Processing*, Detroit, Michigan, May 8-12, 1995, **4**, pp. 2619-2622.
6. Nicchiotti G. and Scagliola C., Generalised projections: a tool for cursive handwriting normalization, *Proceedings of the Fifth International Conference on Document Analysis and Recognition* ICDAR '99 Bangalore, India 20-22 Sept. 1999, pp. 729-732.
7. Press W. H., Teukolsky S. A., Vetterling W. T., and Flannery B. P., *Numerical recipes in C*, 2nd Ed. (Cambridge University Press, New York, 1992).
8. Pu Y. and Shi Z., A natural learning algorithm based on Hough transform for text lines extraction in handwritten documents, *Proceedings 6th IWFHR*, Taejon, Korea, August 1998, pp. 637-646.
9. Ratzlaff E. H., Nathan K. S. and Maruyama H., Search Issues in the IBM Large Vocabulary Unconstrained Handwriting Recognizer, *Proceedings 5th IWFHR*, Colchester, England, September 1996, pp. 177-182.
10. Subrahmonia J., Nathan K. S. and Perrone M., Writer dependent recognition of on-line unconstrained handwriting, *Proceedings of ICASSP 96: IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, May 7-11, 1996, **6**, pp. 3478-3481.
11. http://www.research.ibm.com/electricInk