Proc 13th ICPR, Vienna (1996)
Vol III, pp 818-822

Statistical Zone Finding

András Kornai
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120
kornai@almaden.ibm.com

Scott D. Connell
Department of Computer Science
Michigan State University
East Lansing, MI 48824
connell@cps.msu.edu

Abstract

We propose a statistical technique of zone finding for the class of documents that are neither rigidly structured like tax forms nor very unstructured like magazine pages or engineering drawings. Given an initial window assumed to contain the final zone (bounding box) of interest, and a 'signature' of the target, we propose to locate the final zone by a combination of simple outside in and inside out searches based on the assumption that the coordinates of the target have unimodal distribution. Results are presented in the bank check domain, and the applicability of the technique to other domains is discussed.

0. Introduction

Real world Optical Character Recognition (OCR) systems rarely enjoy the luxury, often taken for granted in more academic systems, of working with clearly delineated text zones. In fact, the task of *zoning*, or *region extraction*, i.e. identifying and precisely demarcating the zone(s) containing the text to be recognized, usually requires a dedicated module in commercial OCR systems. There are, to be sure, cases where zoning is nearly trivial, most importantly in forms processing systems dealing with a small variety of predefined structures, as is the case with tax and credit application OCR (see [1]). Such forms, with their fixed layout, are at the high end of a continuum of decreasingly rigid layout structures. At the low end of this continuum we find newspaper articles, engineering drawings, and other relatively free-form structures where the material that is of interest for OCR is distributed with nearly uniform probability (after cropping the white margins) over the whole page. When page layout is very rigid, probabilistic techniques are of no interest, since after registration of the page the zones can be found deterministically. When page layout is very flexible, probabilistic techniques can not gain sufficient purchase in the near-uniform priors, and region extraction is replaced by various top-down and bottom-up page decomposition techniques (for a recent survey see Chapter 4.3 of [3]). The statistical technique described in this paper comes into play when the distribution of zones to be OCR-ed is neither uniform nor fully predictable.

Section 1 of the paper presents the key ideas of the algorithm in the bank check domain, where the zones of interest are the *courtesy* zone containing the handwritten dollar amount written in digits and the *legal* zone containing the same amount spelled out in words. Section 2 presents some experimental results and discusses their significance. The concluding Section 3 is devoted to the larger question of how to extend the domain of the algorithm from bank checks to other problems of great practical importance such as finding the *address block* on mail pieces ([7], [2]).

1. Bank check zone finding

In this section we present a bird's-eye view of our zone finding algorithms, concentrating on their abstract logical structure at the expense of implementation details. In 1.1 we describe the main characteristics input data, in 1.2 we sketch the main logical steps of the algorithm, and in 1.3 we describe the preprocessing (deskewing) and postprocessing (sanity checks). The methods used in developing and fine-tuning the algorithm are described in 1.4.

1.1. The input data

To the casual observer the personal checks used in the US appear highly consistent: they come in a single size (6 by 2.7 inches), the courtesy amount box has fixed dimen-

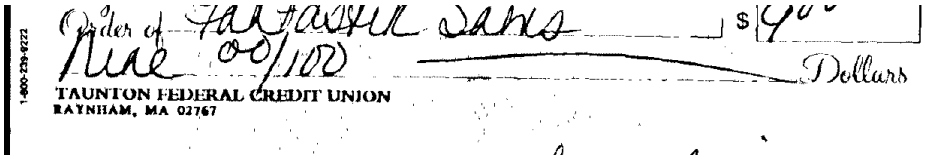

Figure 1. Input window for legal zone

sions (1 by .2 inches) and always appears at the same location (centered at 5.25,1.1), and the preprinted date, pay_to, legal, memo, and signature baselines look as if they always appear at predefined distances from the top. Indeed, direct superposition of paper checks from different banks and printers rarely reveals discrepancies larger than 1 mm vertically or 10 mm horizontally, equivalent to 10 or 100 pixels at the 240 dpi resolution our images are generated. To deal with this amount of variability, a simple heuristic registration process involving rotation (deskewing), vertical and horizontal translation, and perhaps magnification, would suffice. But our input data actually shows far greater variability.

First, the high speed/high volume commercial scanner generating the images does not capture the check at the center (or some other consistent location) of the imaging zone: vertical displacements over 150 pixels are quite common (independent of skew, which is in the $\pm 3\%$ range). Second, the background images are rarely dropped out completely, resulting in an uneven and largely unpredictable background noise. Third, the position of the writing is not fully determined by the position of the preprinted baselines and the courtesy box. Alphabetic writing routinely descends below the baseline and ascends above the pay_to line, and digits will often extend beyond the courtesy box. Finally, as if to make the task artificially harder, data is presented to the system in two highly cropped (and uncorrelated) windows (Figs. 1-2).

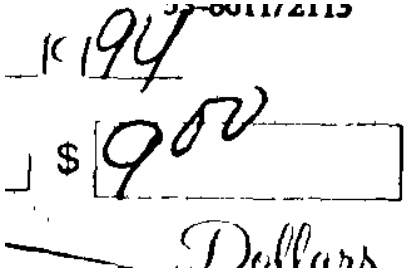

Figure 2. Input window for courtesy zone

Being restricted to limited windows means that we can not take advantage of the full pattern of preprinted lines and boxes, because one window contains only the pay_to and legal baselines (but not the full courtesy box or the signature or memo lines) and the other window only contains the courtesy box and a short segment of the pay_to line. In the check domain such restrictions are accidental and in fact could be entirely avoided simply by permitting a full view rather than these restricted windows. But, as we shall see in Section 3, the algorithm gains a great deal in generality from treating these restrictions as unavoidable.

1.2. The main algorithm

We choose registration *pivots* based on their ease of detection and their consistent location in relation to the desired zone boundaries: in the legal window, the pay_to and legal baselines, and in the courtesy window, the white space surrounding the preprinted courtesy box. The key aspect of the input data that we take advantage of is that *the coordinates of pivots have unimodal distributions*. We do not assume these distributions to be normal (because the overall density plot has a flatter peak) but in fact they come reasonably close (Figs. 3-4).

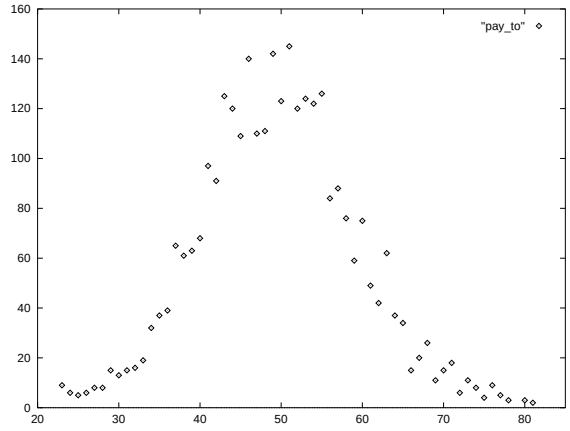

Figure 3. Distance of pay_to line to top of legal window

We use nonparametric estimates first to define upper and lower bounds for each characteristic point: these are analogous to the standard confidence intervals associated to normal distribution. For example, the y coordinate of the

(unskewed) pay_to baseline is assumed to be between two empirically defined constants PAYTOMIN and PAYTOMAX which define the minimum and maximum permissible values.

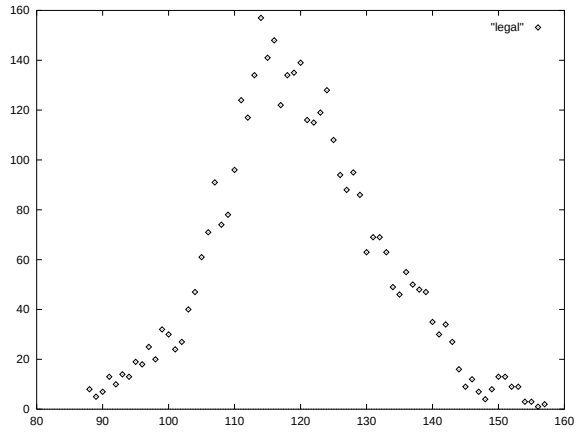

Figure 4. Distance of legal line to top of legal window

Within this range, we employ two search techniques: outside in and inside out, depending on the nature of the pivot that we search for. The characteristic 'signature' associated with horizontal (vertical) lines is a sharp peak in the horizontal (vertical) blackness counts, the characteristic associated with white zones or margins is a flat valley. Outside in search starts from the extremes and progresses toward the more likely values, while inside out search starts from the most likely value and progresses outward. The former is better suited to peaks, the latter to valleys. Therefore, for black line pivots we use use outside in and for white zone pivots we use inside out search. In an ideal case, the (outside in) search starting from the top of the pay_to baseline range would find the same peak as the (outside in) search starting from the bottom, and this peak would be at or near the midpoint of this interval. But the data is noisy enough to produce a significant number of false positives for any peak-picking algorithm, and this makes both pre- and postprocessing necessary.

1.3. Pre- and postprocessing

The most critical step in finding the legal zone is finding the legal and pay_to baselines. While the cursive writing in the legal zone will generally extend below the legal baseline (descenders) and will often be contaminated by descenders from the pay_to zone, taking care of these matters is of secondary importance compared to the primary task of finding the pivots. Since projecting along the skew direction (rather than in the nominally horizontal direction) of the document will produce stronger, more well-formed peaks, we first need to estimate the skew. We analyze the horizontal displacement between vertically close runs and compute tangents at both ends. After rejecting the smallest and largest quartile to avoid outliers, global skew is calculated as the average of the remaining tangents.

If the result is unrealistic (falls outside a predefined range) or was computed on the basis of too few tangents, we assume skew to be zero. Throughout the zoning process, a large number of similar sanity checks are performed to reduce the chances of selecting false positives as the pivots. For example, if the vertical distance between the estimated legal and pay_to baselines falls outside of a predetermined confidence range we search for different baselines. At the end of the process, if the results do not meet all sanity criteria, the whole check is rejected, because the cost of correcting recognition errors would be considerably larger than the cost of not doing OCR.

1.4. Algorithm development

At the very first (preprocessing) step of searching for horizontal black pixel runs in the image we are already confronted with a definitional problem: what constitutes a run of the kind found in the baselines? Clearly, those runs formed by a few horizontally adjacent black pixels e.g. in the vertical strokes of letters are not to be counted here – we need some MINRUNWIDTH parameter to filter these out. Similarly, runs broken up by salt-noise must be mended, which requires setting a MAXRUNGAP or MINRUNDENSITY parameter analogous to the *smearing* parameter C used in constrained run length algorithms (CRLA, see [9] and the references cited therein). Altogether there are some two dozen #define-s in the algorithm and setting them appropriately for the task is a major issue. For example, we take MINRUNWIDTH to be 128 and MINRUNDENSITY to be .75 (every three out of four pixels must be black), but why this and not some other values?

To gain detailed data sets for statistical inferencing we used the bootstrapping methodology described in [6] whereby we iteratively gained more and more reliable knowledge of many empirically defined parameters related to the size, location, and density of the horizontal runs and projection profiles formed by these two lines. Visual inspection of graphs such as those given in Figs. 3-4, combined with hand-verification of the rejected check images, is sufficient to determine the confidence intervals for the location

of the pivots.

For other parameters it was necessary to perform the kind of gradient descent search described in [4] whereby the same algorithm is run with different settings until a performance optimum is found. For example, in the runs profile peaks are formed by merging those runs that are within a close proximity of each other. The maximum vertical thickness of baselines, a parameter we called `PEAKWIDTH`, is used to limit this merging process by imposing a maximum width. This parameter is closely linked to `TANSTEP` which defines the resolution of the deskewing: the more skew we tolerate the wider the peaks become. Another parameter affecting the final number and quality of peaks found by the merging process is the number of seeds it starts with, `NUMOFPEAKS`. The best values of `TANSTEP`, `PEAKWIDTH` and `NUMOFPEAKS` can only be found in combination.

2. Results

After guesstimating the parameters of the extraction algorithm, the legal region extraction algorithm was first tested by collecting the results of a more detailed hand-verification/data entry phase involving over 17,000 legal windows. Of these, the algorithm has selected a horizontal coverage which spanned too high or too low in 994 cases (5.82% error rate) at a rejection rate slightly below 9%. With gradient descent fine-tuning, the performance of the algorithm improves considerably, to 1.44% error at 9% rejection on the over 8,600 courtesy windows we used for testing (a set fully independent of the legal set used for fine-tining). Perhaps some part of this improvement can be attributed to the courtesy zone being inherently easier to extract than the legal zone, but if so, it is not obvious why. The pivots of the legal zone, the two parallel baselines, are hardly ever dropped out entirely, while the pivot of the courtesy zone, the preprinted courtesy box, is missing in over half of the images (because of dropout ink or too high thresholding) which makes finding the courtesy signature intrinsically harder inasmuch as valleys, i.e. negative information, must be located.

One aspect of the results that is of particular interest to those working on similarly degraded data is a brief overview of the main modes of failure. In one experiment we tested 8652 legal windows, of which 757 (8.75%) were rejected. Table 1 shows a summary of each error category reported by the algorithm. The category *Oversize* contains a certain type of machine printed (e.g. payroll) check that is generally larger than the standard personal check. Since our recognition system ([5],[6]) focuses on handwritten checks, rejecting such oversize checks is actually a positive result. The categories *Few Runs* and *Few Peaks* contain those images in which less than a minimum number number of runs or horizontal projection peaks above a blackness threshold are found. Hand-inspection reveals that for such checks the baselines are almost always missing, because of thresholding problems, dropout ink, or the regrettable fact that certain 'designer' check styles simply do not have baselines.

Error Category	Absolute # of errors	Percentage
Oversize	71	0.82%
Few Runs	438	5.06%
Few Peaks	412	4.76%
No Bottom	274	3.17%
No Top	108	1.25%
True Reject	757	8.75%

Table 1: Error Category Breakdown

In the current system, *Few Runs* does not signal complete failure, only failure of the first pass which is based entirely on the analysis of runs. We found that our method of runs analysis is very fast and can be made highly reliable with the aid of simple sanity checks. Therefore, this is the first pass of the analysis, disposing of over 80% of the cases. But when the baselines are not prominent enough, runs analysis does not have enough to go on, and we employ a second pass based entirely on the analysis of peaks in the horizontal blackness profile. *Few Peaks* signals the failure of this second pass, and the check is rejected.

No Bottom and *No Top* refer to the cases in which no legal or pay_to baseline could be found in the first three horizontal projection peaks after the run based method has already failed. These cases typically come from confusion between the baselines and the prominent horizontal line that the check writer often puts in as a space-filler after the number. As can be seen on Fig. 1, this line is often much better recognizable than either the legal or the pay_to line. On this particular example our algorithm performs correctly because the handwritten line is sufficiently curved and not very long, but straight and long handwritten lines can be taken for a baseline. *No Bottom* images are rejected, but *No Top* is passed on to the recognizer using the top of the window in place of the pay_to line.

3. Conclusions

By defining the problem as one of zone finding rather than as one of registration our algorithm gains a great deal of generality. There are only three assumptions that the input must meet for the technique to be applicable. First, one must be able to specify a fixed window which has the property that the target will either appear inside this window or the document can be assumed not to contain the target at all. This assumption is met by a large variety of targets, such as headers/footers in books, journals, and newspapers, intercolumnar line numbering (e.g. in patent documents), Library of Congress Cataloging-in-Publication Data, etc. Sometimes, it can be met only at a price of rejecting a small fraction of items, as is the case with postal address blocks if we define the window to be the bottom two-thirds of the piece (see [8]).

Second, the distribution of the target within the initial window must be unimodal. This assumption is met by any target of which we *approximately* know where it will fall. In fact, once this second assumption is met, the first assumption in a sense follows by taking the initial window to be a large enough *confidence rectangle* around the mode. Nevertheless, we keep these two assumptions separate as their practical import are quite different: the second assumption is an abstract assumption about the spatial distribution of the data, while the first assumption is really an assumption about the overall system's ability to deal with missing data. As long as a clear decision can be made on the basis of the window the first assumption is met, independent of whether the decision is to proceed or to abort. For example, if the date zone of a bank check contains no date we can assume it has not been dated and proceed, but if the legal zone contains no legal amount we must reject the check entirely.

Third, the target must have a characteristic *signature* that distinguishes it from the background. The actual nature of the signature affects only the inner loop of the algorithm presented here: for example, we used both long runs and horizontal projection peaks for baseline detection with about the same success (and a two-pass combination with slightly improved results). For courtesy amount detection we used the white margins around the numbers as the signature, a choice that would be appropriate for many relatively isolated targets such as page numbers in books and articles. Clearly, an appropriate signature is a precondition for the success of the whole enterprise, and in this paper we make no claims to having found some universal signature equally appropriate for all targets. Rather, we proposed to decompose the zoning task into two steps: first, designating a window where the target must appear and second, searching for the characteristic signature of the target within that window.

References

[1] Casey, R.G., D. Ferguson, K. Mohiuddin, and E. Walach "Intelligent Forms Processing System" *Machine Vision and Applications* **5** (1992) 143-155

[2] Downton, A.C. and C.G. Leedham, "Preprocessing and presorting of envelope images for automatic sorting using OCR," *Pattern Recognition* **23** (1990) 347-362

[3] O'Gorman, L. and R. Kasturi (eds): *Document Image Analysis*. IEEE Computer Society Press, Los Alamitos, CA, 1995.

[4] Kohavi, R. and G.H. John, "Automatic Parameter Selection by Minimizing Estimated Error" In: Prieditis & Russell, eds., *Machine Learning: Proceedings of the Twelfth International Conference* Morgan Kaufmann Publishers, San Francisco, CA, 304-312

[5] Kornai, A., K.M. Mohiuddin, and S.D. Connell, "An HMM-Based Legal Amount Field OCR System for Checks," *1995 IEEE International Conference on Systems, Man and Cybernetics*, Vancouver BC, October 1995, 2800-2805

[6] Kornai, A., K.M. Mohiuddin, and S.D. Connell, "Recognition of Cursive Writing on Personal Checks," to appear in Proc. 5th International Workshop on Frontiers in Handwriting Recognition, Essex, 1996

[7] Lii, J., P.W. Palumbo, and S.N. Srihari, "Address block location using character recognition and address syntax," *Proc. 2nd International Conference on Document Analysis and Recognition* Tsukuba Science City, Japan, Oct. 1993, 330-335

[8] Srihari, S.N., C.-H. Wang, P.W. Palumbo, and J.J. Hull, "Recognizing address blocks on mail pieces: specialized tools and problem-solving architecture," *AI Magazine* **8** (Winter 1987) 25-40

[9] Wahl, F.M., K.Y. Wong, and R.G. Casey, "Block segmentation and text extraction un mixed text/image documents," *Computer Graphics and Image Processing* **20** (1982) 375-390