A Histogram-Based Approach to Mathematical Line Segmentation

Mohamed Alkalai and Volker Sorge

School of Computer Science, University of Birmingham
{M.A.Alkalai,V.Sorge}@cs.bham.ac.uk

Abstract. In document analysis line segmentation is a necessary prerequisite step for further analysing of textual components. While much work has been devoted to line segmentation of regular text documents, this work can not be easily adopted to documents that contain specialist components such as tables or mathematical expressions. In this paper we concentrate on a line segmentation technique for documents containing mathematical expressions, which, due to their two dimensional structure are often comprised of multiple distinct lines. We present an approach to line segmentation in the presence of mathematics that is based on a set of histogram measures and heuristics considering vertical and horizontal distances of characters only. The method also provides a technique to distinguish consecutive lines that are vertically overlapped but belong to different mathematical expressions. Experiments on data sets of 200 and 1000 maths pages, respectively, show a high rate of accuracy.

1 Introduction

Line segmentation is a prerequisite step for structural analysis of both printed and handwritten documents. While much work has been done for text line segmentation of documents containing primarily text only. The developed techniques such as projection profile cutting [3,2], smearing [6], grouping [4] or seam carving [5], rely to some extent on the fact that in regular text documents generally lines can be clearly separated by detecting consecutive whitespace between them.

For documents containing mathematical expressions, however, these techniques do not suffice due to the occurrence of particular artifacts of mathematical notations such as math accents, the limits of sum symbols, etc. that, while actually constituting a single line, can appear spatially lay out over more than one separable line. And while there exists quite a body of work on the segmentation of mathematical documents, this work is generally more concerned with the identification and separation of mathematical structures from surrounding text and their subsequent layout analysis [8].

In this paper we present a math line recognition algorithm that is reliable independent of knowledge on any peculiarities of mathematical expressions (Sec. 2). It is based on spatial considerations only, thus avoiding committing to premature errors, that stem from considering actual content such as symbols or fonts. In particular, we use a histogram-based approach, considering horizontal spaces

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part I, LNCS 8258, pp. 447–455, 2013.

between glyphs in lines of a page, in order to classify lines into two types: principal and non-principal, where the former are lines in their own right, while the latter are only parts of mathematical expressions and should be merged with neighbouring lines. In addition to this technique we have developed a set of heuristics using simple yet effective measures for correction of classification errors as well as to separate lines that share vertically overlapping characters but that belong to distinct mathematical expressions. We demonstrate the effectiveness of our approach by presenting experiments on two distinct data sets containing 200 and 1000 pages from mathematical documents, where we achieve an accuracy rate of 96.9% and 98.6%, respectively for line detection (Sec. 3). A previous version of the algorithm, that in particularly did not allow for splitting lines with vertical overlap, has been successfully applied in experiments to improve the identification rate of mathematical expressions was presented in [1].

2 A Histogrammatic Approach to Line Segmentation

The basic idea of our approach is to detect all possible individual lines first and then merge neighbouring lines into single lines likely to contain mathematical expressions. Thereby we rely neither on knowledge of the content of lines, font information nor vertical distance. Instead we use a histogrammatic measure on space within a single line. We then employ simple height considerations to detect lines that have not been correctly classified to be merged or not merged. In a final step, each line that is classified to be merged is clustered with its closest line as long as they are horizontally overlapped. In summary our procedure consists of the following steps:

1. *Initial line separation* by vertical cuts (cf. [7]).
2. *Detect and split lines* with vertically overlapping characters.
3. *Initial classification of lines* into principal and non-principal, where the latter should be merged with the former.
4. *Improvement of classification* using two measures based on character height.
5. *Merge* non-principal with neighbouring principal lines to obtain final lines.

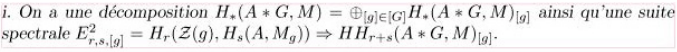
We now define the concepts of our procedure more formally. Step 1 is given by the following three definitions.

Definition 1 (Bounding Box). *Let g be a glyph, then the limits of its* bounding box *are defined by $l(g), r(g), t(g), b(g)$ representing left, right, top and bottom limit respectively. We also have $l < r$ and $t < b$.*

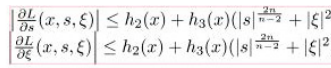
Definition 2 (Vertical and Horizontal Overlap). *Let g_1,g_2 be two glyphs. We say g_1* overlaps vertically *with g_2 if we have $[t(g_1), b(g_1)] \cap [t(g_2), b(g_2)] \neq \emptyset$, where $[t(g), b(g)]$ is the interval defined by the top and bottom limit of glyph g.*

Similarly we define horizontal overlap *of two glyphs g_1,g_2 by $[l(g_1), r(g_1)] \cap [l(g_2), r(g_2)] \neq \emptyset$.*

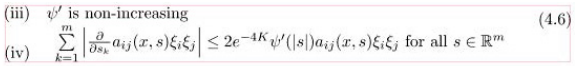
We can now define a line using the vertical overlap on a set of glyphs.



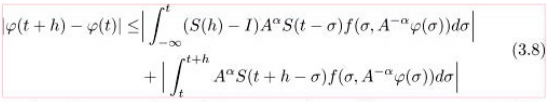
(a) Lines contain embedded math expressions that share vertically overlapping characters.



(b) Lines with displayed math expressions that share vertically overlapping characters.



(c) Lines contain embedded math expressions that are overlapped because a part of these lines is misaligned



(d) Lines with displayed math expressions that are overlapped because a part of these lines is misaligned

Fig. 1. Examples of different types of lines overlapping

Definition 3 (Line). *Let $G = \{g_1....g_n\}$ be a set of glyphs. We call a $L \subseteq G$ a line if for every $g \in L$ there is a $h \in L$ such that g and h overlap vertically and there is no $g \in G \setminus L$ that overlaps vertically with any element in L.*

Since this initial step separates lines that share vertically overlapping characters as one line. For examples of different types of overlaps see Fig. 1. Therefore, we perform a post-processing step to detect and split those lines, which is formalised in the next three definitions:

Definition 4 (Detect Overlapping Line). *Let $L = \{g_1....g_m\}$ be a line where the glyphs are sorted in ascending order according to $l(g)$ and m. We split L if the following conditions are satisfied:*

(i) Neighbouring glyphs $g_1, g_2 \in L$ horizontally overlap such that $[l(g_1), r(g_1)] \cap [l(g_2), r(g_2)] \neq \emptyset$.
(ii) The same two neighbouring glyphs $g_1, g_2 \in L$ not both vertically overlap with any $g \in L$ where $g \neq g_1$ and $g \neq g_2$ such that $[t(g_1), b(g_1)] \cap [t(g_2), b(g_2)] \cap [t(g), b(g)] \neq \{g_1, g_2, g\}$.
(iii) $h(g_1) < (t(g_2) - b(g_1))$ and $h(g_2) < (t(g_2) - b(g_1))$ where h is the height of glyphs such that $h = b(g) - t(g)$.
(iv) $h(g_1) > (w(g_1)/2)$ and $h(g_2) > (w(g_2)/2)$ where w is the width of glyphs such that $w = r(g) - l(g)$.

We then split the line into two lines by using a threshold that is determined by horizontally projecting lines across the whole vertical distance between the two overlapping glyphs, using the y-coordinate value of the line that crosses the least number of glyphs.

Definition 5 (Separator Value). *Let $g_1, g_2 \in L$ overlapping glyphs, and $Y = \{y_1....y_n\}$ be y-coordinate values between $b(g_1)$ and $t(g_2)$. Then let the* separator value *S is defined as the $y \in Y$ that minimises the number of vertically overlapping glyphs.*

We then cluster glyphs into two lines using the separator value S as a threshold.

Definition 6 (Split Overlapping Lines). *Let $L = \{g_1....g_m\}$ be the line to be split. Then we define $L_{above} = \{g \in L | h(g)/2 < S\}$ and $L_{below} = L \setminus L_{above}$.*

We can now define the distance measure with respect to which we will consider histograms.

Definition 7 (Horizontal Distance). *Let $L = \{g_1....g_n\}$ be a line. We call two glyphs $g, g' \in L$* neighbours *if $r(g) < l(g')$ and there does not exist a $g'' \in L$ with $g'' \neq g$ and $g'' \neq g'$ such that $[l(g''), r(g'')] \cap [r(g), l(g')] = \emptyset$. We then define the* horizontal distance *d between two neighbouring glyphs g, g' as $d(g, g') = l(g') - r(g)$.*

Observe that in the above definition we define distances only for elements in the line that do not overlap horizontally. Thus the distances represent the actual whitespace in lines.

The distance measure from the previous definition allows us now to compute a histogram that captures the horizontal distances between glyphs in lines for the entire page. Figure 2 shows two examples for the histograms, where the x-axis denotes the values for the distance measure d in pixels and the y-axis the number of occurrences of a particular distance. Note, that when building the histogram, we deliberately omit all the values where no distance occurs or in other words, where the y value is equal to 0.

We can observe a general pattern in these histograms: They can be split into two parts by a global minimum that is roughly in the middle of the x-axis. This leaves two parts, each with a global maximum. Furthermore in the right part one can identify a further global minimum. While this can be at the very end of the x-axis it usually is not. We call these two minimal points v_1 and v_2, respectively, and use them to define classification of lines as follows:

Definition 8 (Principal Lines). *Let L be a line. We call L a* principal line *if there exists two neighbouring glyphs $g, h \in L$ with $v_1 \leq d(g, h) \leq v_2$. Otherwise L is a* non-principal line.

The intuition behind this definition is that the values in the histogram less than v_1 represent distances between single characters in a word or a mathematical expression, whereas the area between v_1 and v_2 represents the distance between single words, which generally do not occur in lines that only constitute part of a mathematical formula, for example, those consisting of limit expression of a sum.

While the measure alone already yields good results, it can be improved upon by considering a simple ratio between glyph heights of principal and non-principal lines.

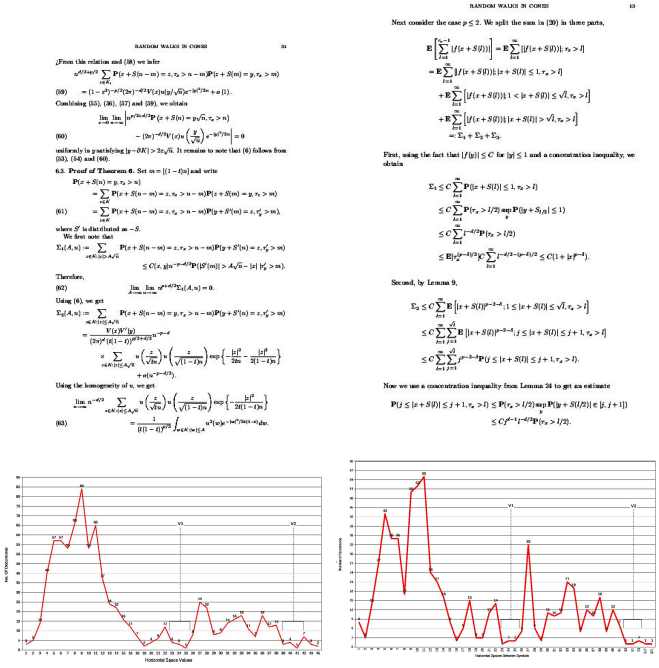


Fig. 2. Examples of pages and their histogram of the gap between glyphs

Definition 9 (Height Ratio). *Let L_1 and L_2 be two consecutive lines, such that L_1 is a non-principle line and L_2 is the nearest principle line to L_1. If $\max_{g\in L_1}[b(g) - t(g)] > \frac{1}{T}\max_{g\in L_2}[b(g) - t(g)]$, where $1 \leq T \leq 2$ then L_1 is converted into a principal line.*

Observe that the value for the parameter T is fixed and determined empirically by experiments in on a small sample set.

Since the previous step tackles only the problem of wrongly classified principal lines, we also need to define a corrective instrument to detect non-principal lines that have wrongly been classified as principal lines. This is achieved as follows:

Definition 10 (Non-principal Height Bound). *Let $L_n = \{n_1, n_2, ...n_l\}$ be the set of non-principal lines of a page and $L_p = \{p_1, p_2, ...p_k\}$ be the principal lines of the same page.*

Then we define the non-principal height bound *as the maximum height of all non-principal lines M as*

$$M = \max_{n\in L_n} |b'(n) - t'(n)|,$$

where t' and b' are the top and bottom limits of L respectively, such that $t'(n) = \min_{g\in n} t(g)$ and $b'(n) = \max_{g\in n} b(g)$.

Any $p \in L_p$ is converted to a non-principal line, if and only if, $[b'(p) - t'(p)] \leq M$.

Table 1. Results for vertical splitting

Overlap Type		Correct Split	Incorrect Split
embedded math expression	direct overlap	36	2
	misaligned	1	0
display math expression	direct overlap	2	2
	misaligned	2	0

Table 2. Experimental results for line recognition

Method	Pages.	Total line	Lines found	Correct lines	Accuracy
Vert. Cuts	200	5801	6987	5015	86.4%
Hori. Dist.	200	5801	5727	5265	90.7%
Height Ratio	200	5801	5910	5587	96.3%
Height Bound	200	5801	5863	5625	96.9%

Table 3. Experimental results of 1000 pages

Pages.	Total line	Lines found	Correct lines	Accuracy
1000	34146	34526	33678	98.6%

Table 4. Evaluation results of 1000 pages

Line Type	Precision(P)	Recall(R)
Principal Line	99.39%	99.15%
Non-principal Line	93.87%	81.49%

Once the classification of lines is finished, in the final step we merge non-principal lines with their horizontally closest neighbouring principal line, but only if there exists horizontal overlapping between them. If not, the non-principal line is converted to a principle line.

Definition 11 (Merging Lines). *Let N and P be non-principal and principal lines respectively, such that P is the nearest neighbour of N. Let l' and r' be the left and right limits of L respectively, such that, $l' = \min_{g \in L} l(g)$ and $r' = \max_{g \in L} r(g)$. If $l'(P) < r'(N)$ and $r'(P) > l'(N)$ then N and P are merged. Otherwise, N is converted to P.*

3 Experimental Results and Discussion

We have run experiments on two datasets of 200 and 1000 pages, respectively, taken from a wide variety of mathematical documents. Before discussing the results of our overall procedure we first present the results of the line separation step alone. Our dataset contained 36 pages with lines that share vertically overlapping characters. These lines were effectively of two types: text lines with embedded math expressions and lines with display math expressions. Each of the two categories are further divided into two sub-categories: lines that overlap

because of at least one overlapping glyph and lines that overlap because a part of these lines is misaligned. (See Figure 1 for examples of the four types.)

Table 1 shows the results of the splitting step. Observe that the two lines that are incorrectly split fail due to characters being wrongly clustered to either of the two result lines, which suggests that some improvement on our separator threshold method should be investigated in the future.

In terms of experiments of the overall procedure we have carried out initial experiments on 200 pages. These pages are taken from 12 documents comprising a mixture of books and journal articles. Table 2 presents the experimental results for this dataset. We have compared using simple vertical cuts, with our techniques of using the horizontal distance measure introduced in Def. 7 as well as using additionally the height ratio defined in Def. 9 and the non-principal height bound defined in Def. 10. As height ratio parameter we have set $T = 1.7$, a value that was experimentally determined on a small independent sample set.

Altogether we manually identified 5801 lines in the 200 pages of the dataset. We compare this number with the number of lines found altogether and the number of lines identified correctly, that is, those lines corresponding to the actual line as manually identified in the dataset.

Not surprisingly simple vertical cuts results in a larger number of lines and, as there is no subsequent merging of lines, in a relatively low accuracy of 86.4%. Using the horizontal distance measure improves this accuracy, however, in general merging too many lines. This is corrected by the addition of the height ratio that re-classifies some of the lines incorrectly assumed to be non-principal as principal lines. As a consequence we get a slightly higher number of lines but also a higher accuracy of 96.3%. A further slight improvement in this accuracy to 96.9% is obtained using the height bound.

To further examine the robustness of our technique and in particular to rule out that there was overfitted to our original data set we have experimented with a second independent and larger data set. The data set contains 1000 pages composed from more than 60 mathematical papers different of our original set.

We ran our technique on this second larger data set and then manually checked the results by painstakingly going through every page line by line. Consequently we have done this comparison only for the full classification including both height ratio and height bound correction. And while we can not rule out some classification mistakes due to human error we are very confident that the experimental results given in Table 3 are accurate.

Table 3 demonstrates that although, the data set is five times the size of the previous one our classification results remain stable. In fact, one can see that in comparison with table 2 we have even a increase of recognition rate by approximately 2%. This result gives us confidence about the effectiveness of our technique even on large datasets and documents.

Further evaluation is shown in table 4. Precision (P) and recall (R) measurements are used. As can be seen, the (P) and (R) percentages for principal line are close and high since there are 33186 correct principal lines and a very small number of incorrect ones. For non-principal lines, the percentages are not as

high as the fronter lines. However, one can still claim that these results are very promising in comparison with the Vertical Cuts results where all non-principal lines are not recognized.

For the lines that were not identified, it is possible to categorise the recognition error into two types.

Incorrect Non-principal Lines: The most common error stems from classifying a line with respect to the horizontal distance measure as a principal line that should actually be non-principal. This is the case when there is a gap between two neighbouring glyphs that satisfies the horizontal distance condition. Below are some examples show several cases of errors taken directly from our dataset.

$$\leq \mathbf{C} \sum_{\mathbf{i=1}}^{\infty} \sum_{\mathbf{j=1}}^{\sqrt{\mathbf{l}}} \mathbf{E}[|\mathbf{x} + \mathbf{S(l)} \dots \qquad \widetilde{\mathbf{B}_i}\widetilde{\mathbf{B}_{i+1}}\widetilde{\mathbf{B}_i} = \mathbf{2}\widetilde{\mathbf{M}_i} + \dots$$

Although, the first expression should be detected as a single line, the limits under the two summation symbols are at a distance that coincides with the distance identified by the histogram for the entire page. Likewise, in the second expression, also taken from our dataset, the tilde accents have a similar distance.

Incorrect Principal Lines: This error occurs when a line is initially classified as non-principal line as it does not contain any glyph gaps that coincide with the distance measure derived from the histogram. Examples of these lines are those with single words, page numbers, single expressions etc. While these can be corrected by the height ratio, sometimes they are not as they do not satisfy the ratio condition. Below is an example taken from our dataset.

$$+\frac{\mathbf{3}}{\mathbf{5}}\left(\mathbf{V_1^{k-1,k,2}(n;(1))}.(\mathbf{L_{n-3}^{k-1,k,2}} - \mathbf{L_3^{k-1.k.2}}\right)$$
$$12$$

Here the page number 12 is merged as a non-principal line to the expression above, as firstly it does not exhibit a glyph gap satisfying the distance measure and secondly its height is significantly smaller as the height of the open parenthesis in the mathematical expression.

4 Conclusions

In this paper, we presented a line detection technique that is geared towards documents that contain a large number of complex mathematical expressions. Our approach can not only deal with detecting compound lines that consist of combination of several independent lines separated by vertical whitespace, but we also have most recently added a method to detect and split math lines that share vertically overlapping characters. The procedure exploits only simple spatial features in a histogrammatic approach, avoiding the use of many parameters that need to be fine tuned or relying on statistical data from large sample sets. Our experiments show that we nevertheless get a high rate of accuracy in detecting correct lines. The algorithm currently serves as a basis for our work on layout analysis of tabular mathematical expressions.

References

1. Alkalai, M., Baker, J., Sorge, V., Lin, X.: Improving formula analysis with line and mathematics identification. In: Proc. of ICDAR (to appear, 2013)
2. Boussellaa, W., Zahour, A., El Abed, H., BenAbdelhafid, A., Alimi, A.: Unsupervised block covering analysis for text-line segmentation of arabic ancient handwritten document images. In: ICPR, pp. 1929–1932 (2010)
3. Marti, U., Bunke, H.: On the influence of vocabulary size and language models in unconstrained handwritten text recognition. In: Proc. of ICDAR 2001, pp. 260–265. IEEE Computer Society (2001)
4. O'Gorman, L.: The document spectrum for page layout analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(11), 1162–1173 (1993)
5. Saabni, R., El-Sana, J.: Language-independent text lines extraction using seam carving. In: Document Analysis and Recognition, pp. 563–568. IEEE Computer Society (2011)
6. Wong, K., Casey, R., Wahl, F.: Document analysis system. IBM Journal of Research and Development 26(6), 647–656 (1982)
7. Zanibbi, R.: Recognition of mathematics notation via computer using baseline structure. Technical report, Queen's University, Kingston, Canada (2000)
8. Zanibbi, R., Blostein, D.: Recognition and retrieval of mathematical expressions. IJDAR 15(4), 331–357 (2012)