2013 12th International Conference on Document Analysis and Recognition

Detecting Main Body Size in Document Images

Paraskevas Diamantatos Vasileios Verras Ergina Kavallieratou
dept. Information and Communication Systems Engineering
University of the Aegean
Samos, Greece
e-mail: kavallieratou@aegean.gr

***Abstract*—In this paper, two techniques are presented, appropriate to detect the text main body size in a document image. One measures it directly, while the other estimates the baselines first. Both are segmentation free. Experimental results are presented over a collection of handwritten text, as well as for a small collection of 10 printed document images, in order to give more objective results.**

Keywords-document image processing; word main body estimation; baseline detection; historical document images

I. INTRODUCTION

Main body or *core region* size is a characteristic that is used quite often in most document image processing systems. By this term, it is considered the central part of the text, excluding ascenders and descenders (Fig.1). Most of the times, it is referred to words.

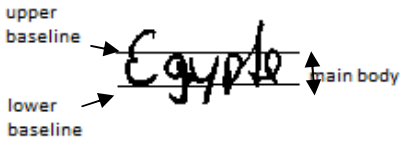

Figure 1. Word main body and baselines.

The last years this characteristic has been used in systems for OCR [1-3], segmentation [4-5], slant removal [6-7], dewarping [8-9], word matching [10], indexing [11], normalization [12], word spotting [13], etc. Its use gives a reference for thresholds and sizes as it is directly related to the size of the characters, the document image resolution and the text orientation.

In [14], they mention: *By mean width of character, we consider the width of characters such as a, b, c, d etc, excluding the characters i, l, j, m, w that are either too narrow (i, j, l), or too wide (m, w). ... Although the character width differs between characters and writers, a rough estimation of the mean width could be made by accepting that excluding the ascenders and descenders the characters with mean width (as defined above), present width equal to their height.*

Thus, main body is critical for document image processing systems. It is not a surprise that many techniques have been developed that detect the main body size directly or by localizing the text baselines and considering the distance between them. Most of these techniques are simple, since it is considered a trivial task of a document image processing system that shouldn't require a lot of the computational time, although some more complex techniques also exist. However, most of these techniques require an estimation of the words or text lines before they proceed to the main body size estimation.

In this paper two new techniques are presented, one of each categories mentioned above, one detects the main body size directly, while the other localizes the text baselines. Both are appropriate for cases that the word or text segmentation is not necessary and/or it would be difficult to be performed e.g. historical documents.

In the next section, a short description of the previous work is given, while the two proposed techniques are presented in sections III and IV. Finally, some results are described in section V and we conclude in section VI.

II. PREVIOUS WORK

To the best of our knowledge, there is no paper presenting main body detection methodology. However, many papers specialized on document image processing describe techniques simple or more sophisticated.

Many of them are based on pixel level processing. Lee and Verma [5] measure the distance from the upper-most pixel to the first foreground pixel, as well as the vertical transitions. Then a search algorithm for the best baselines was applied, exploiting the extracted information. Adamek, Connor and Smeaton [11] use the pixel density. Traversing the contours, he identifies the ascenders and descenders by comparing the distinct vertical limits of the lower case characters.

Others make use of histograms. Cheng and Blumenstein [4] calculate the average vertical value of the maxima and minima on the upper and lower contours respectively. Abnormal maxima and minima are removed based on this average value. Finally, the baselines are estimated by the average of the remaining maxima and minima. In [14] a tenth of the text line was used instead of words. The upper and lower parts of the line where the value of the histogram falls under the 1/3 of its peak value (threshold extracted experimentally) are excluded. Cote et al. [2] developed a method for baseline extraction that makes use of entropy. They compute histograms for different vertical projections and then, they calculate the entropy associated with them. In a more sophisticated technique, Marti and Bunke [1], Gatos,

1520-5363/13 $26.00 © 2013 IEEE
DOI 10.1109/ICDAR.2013.235

Pratikakis and Ntirogiannis [8] and Sharma and Shilpi [9] use linear regression to estimate the upper and lower baselines. They apply linear regression to the upper or lower black set of points. Finally, in order to locate more accurately the main body of each text line, Papavassiliou et al. [15] formulate an HMM for the text and gap stripes within the document image. The parameters are drawn from statistics of the initial set of text and gap areas.

Simple or more complex, these techniques require page segmentation up to word or text line level and they are not appropriate if this is not desirable. In this paper, we describe two techniques appropriate to be applied to document images instead of text line or word level.

III. FIRST TECHNIQUE

This technique, shown in Fig.2, detects the main body size of the text in a document page. Although, the initial idea has common points with the one presented in [6], it does not require text line segmentation.

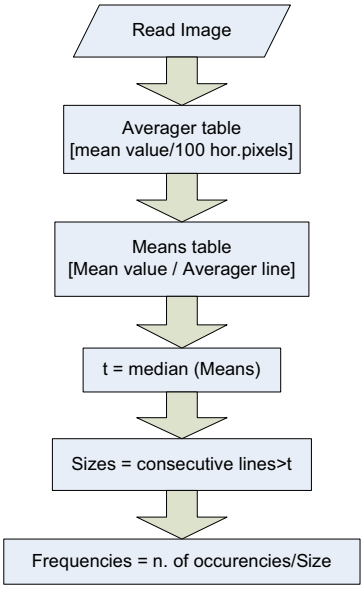

Figure 2. The first technique.

First, the average pixel value is calculated for every N pixels that exist in each pixel line of the image. N can be any value. It is only important for the skew angle of the page that it can handle. The smaller the N, the biggest the skew angle it handles. However, since this work does not emphasize on that, for the results presented here N=100 was chosen. The results are saved in the *Averager* table with size Hx⌈W/100⌉, where H and W are the height and the width of the image, respectively.

Next, the table *Means* is created of size Hx1, were its elements are the average values of the corresponding lines of *Averager* matrix. Then the threshold t is set as the median value of matrix *Means*. By this threshold, we set to zero the values of *Means* that are smaller than t, while we count the consecutive lines with value bigger than t. The set of the different amount of the consecutive lines are the *Sizes* of the various main bodies in the image. Then the occurrences for each size are also counted and saved in *Frequencies*. As main body size, the maximum in *Frequencies* is considered.

Figure 3. Schematique presentation of the first technique, through example.

In Fig. 3 the technique is presented through an example. This technique does not require binarization. Moreover, it can give more information if different main body sizes are present in the same page.

IV. SECOND TECHNIQUE

This technique estimates the baselines of the text in a document page.

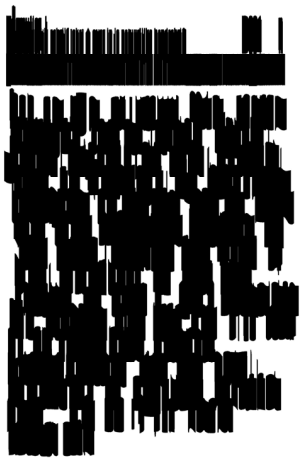

Figure 4. Vertical dilate.

Initially, the document is binarized. Then, the Connected Components (CCs) of the document, for 8-pixel neighborhood, are detected. All CCs bigger than 30000 pixels and smaller than 10 pixels are removed, that is very big area e.g scan noise or figures and very small noisy areas or accents, respectively.

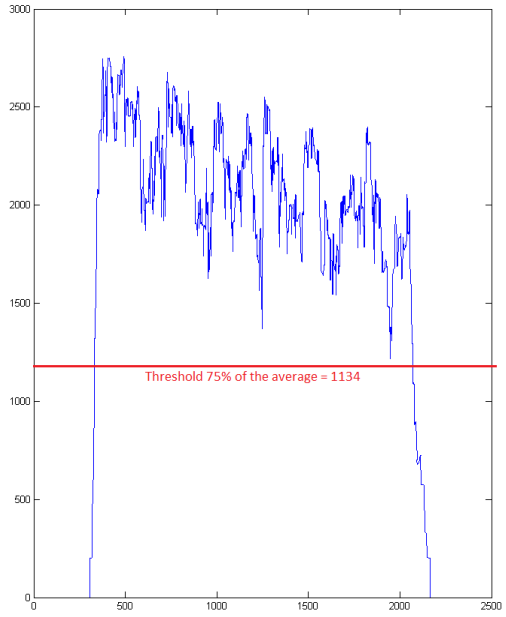

Figure 5. Vertical text localization.

Then vertical dilate is applied. The aim is to identify the horizontal borders of the text area (left – right) and if the text consists of text columns. This is necessary, since in the case of the text columns, each column is treated separately. After the vertical dilate the columns of text form a big connected area (Fig.4). Consequently, CCs are again detected and now only those bigger than 10000 pixels are kept. A vertical histogram is taken and those pixel columns with black pixels more than 75% of the average are marked as text and the others as background (Fig.5).

Then the document is scanned from left to right and the total number of text columns is identified. For each of the text columns a similar procedure is followed this time with a horizontal dilate (Fig.6). The text lines are detected with their respective start and end indexes in the document. To detect the main body of the text the pixel row must contain 170% of the average pixel rows (Fig.7). This ensures that the beginning and the end of the main body will be detected, without including the ascenders and the descenders. Finally the average baselines are calculated and returned as showed on the original document (Fig8).

The technique is presented in Fig.9.

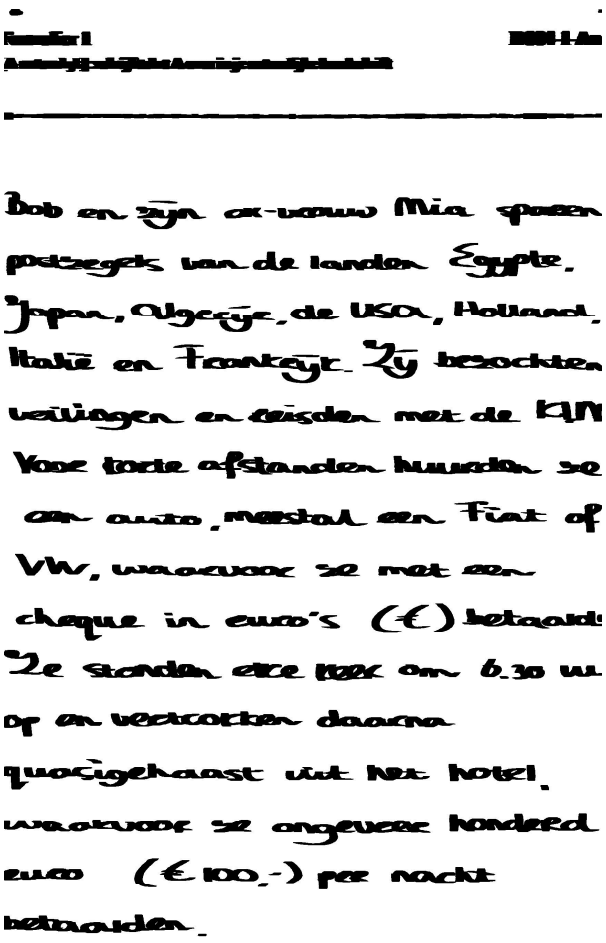

Figure 6. Horizontal dilate.

V. Experimental Results

The evaluation of a technique that estimates the text main body size is not easy, especially when we refer to handwritten text. Here, the TrigraphSlant data set [16] that contains images of handwriting, produced under conditions of natural and forced slant, were used. It includes 190 images from 47 persons. We used 30 images of natural writing by different writers.

In order to create ground truth data, the height of 10 ‘o’ of each image was measured and the mean value was considered. It took us by surprise that even on the same document image, written by the same person, differences of more than 10 pixels were found.

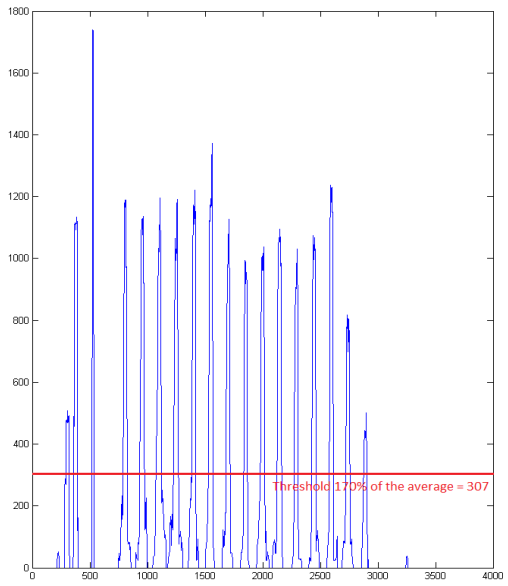

Figure 7. Horizontal text localization.

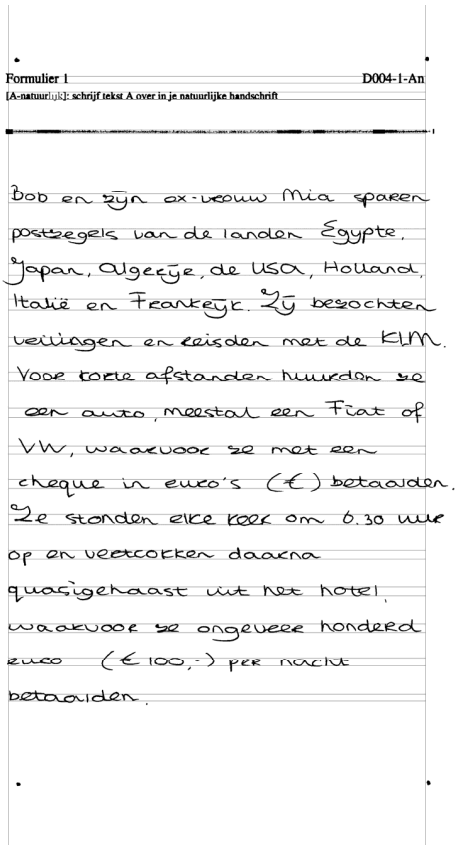

Figure 8. Final result.

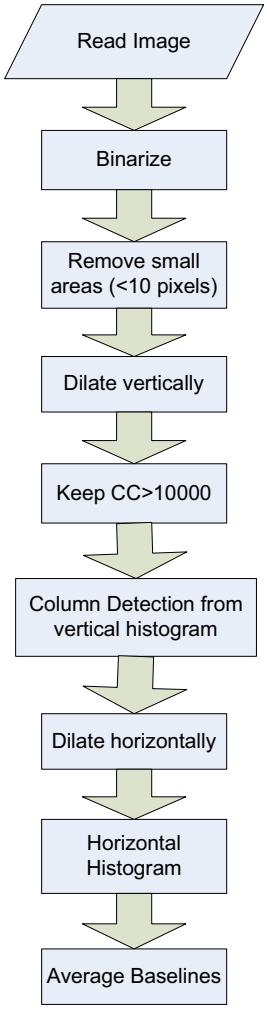

Figure 9. The second technique.

In table I, the mean estimated main body size for 5 writers (D00X) is shown, while you can see the results for 4 document images of the same writer (D00X-1, D00X-2, D00X-3, D00X-4). It is obvious how the size changes, even for the same writer. In our experiments, only the first document image (D00X-1) of each writer for X between 1 and 30 was used.

Since, as it was explained, it is difficult to have exact results, in table II, the average error deviation between the estimated values and the ones detected by the two techniques is given. Moreover, in order to give more objective results, in the same table, it is given the average error deviation between the real values and the ones detected by the two techniques over a collection of 10 printed images that includes font sizes between 8 and 24 pts.

TABLE I. EXAMPLES OF MAIN BODY ESTIMATION

Document Image Code	*Estimated Mean Main Body Size (pixels)*
D001-1-An	34,8
D001-2-Bn	35,2
D001-3-Bl	31,6
D001-4-Br	31,6
D002-1-An	33,6
D002-2-Bn	27,4
D002-3-Bl	37,2
D002-4-Br	27,2
D003-1-An	34,2
D003-2-Bn	33,4
D003-3-Br	30,8
D003-4-Bl	33,4
D004-1-An	29,2
D004-2-Bn	29,2
D004-3-Br	32,8
D004-4-Bl	34,2
D005-1-An	34,6
D005-2-Bn	34,2
D005-3-Br	28
D005-4-Bl	33,4

TABLE II. EXPERIMENTAL RESULTS

Technique	*Average error deviation (pixels) on Trigraph*	*Average error deviation (pixels) on printed DB*
first	2.17	0.67
second	4.96	1.05

VI. CONCLUSION

Two techniques were presented appropriate for the text main body size detection. The two techniques present different characteristics but both are appropriate for cases that the segmentation procedure would be difficult as e.g. historical documents. Experimental results are given on a known database and the error rate is less than the usual standard deviation among the same document or writer. However, although they look good, we keep in mind that it is difficult to give completely objective results. Thus, it is in our plans to evaluate it through indirect methodologies. References

[1] U.V. Marti, H. Bunke, "Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system", Int. Journal of Pattern Recognition and Artifical Intelligence, 15(1), pp. 65–90, 2001.

[2] M. Côté, E. Lecolinet, M. Cheriet, C.Y.Suen, " Automatic reading of cursive scripts using a reading model and perceptual concepts, The PERCEPTO system", IJDAR, vol 1, pp. 3-17, 1998.

[3] Kennard, D. J., Barrett, W. A. "Interactive Training for Handwriting Recognition in Historical Document Collections," DRR'07, Jan. 28 - Feb. 1, 2007.

[4] C. K. Cheng and M. Blumenstein, "The neural-based segmentation of cursive words using enhanced heuristics", Proc. of the 8th International Conference on Document Analysis and Recognition, pp.650-654, 2005.

[5] H. Lee and B. Verma, A novel multiple experts and fusion based segmentation algorithm for cursive handwriting recognition, Proc. of the International Joint Conference on Neural Networks, pp.2994-2999, 2008.

[6] A. Vinciarelli, J. Luettin, A new normalization technique for cursive handwritten words, Pattern Recognition Lett. 22 (9) (2001) 1043–1050.

[7] A. Papandreou, B. Gatos, "Word slant estimation using non-horizontal character parts and core-region information" In: 10th IAPR International Workshop on Document Analysis Systems (DAS 2012), pp. 307–311, 2012.

[8] B. Gatos, I. Pratikakis, and K. Ntirogiannis. Segmentation based recovery of arbitrarily warped document images. In Proc. Int. Conf. on Document Analysis and Recognition, Curitiba, Brazil, Sep. 2007.

[9] D. Sharma, W. Shilpi, "Dewarping Machine Printed Documents of Gurmukhi Script", Information Systems for Indian Languages, Communications in Computer and Information Science series , v.139, pp. 117-123, 2011.

[10] N. Doulgeri, E. Kavallieratou, Retrieval of Historical Documents by Word Spotting, IS&T/SPIE Electronic Imaging 2009.

[11] T. Adamek, N. E. Connor, A. F. Smeaton, Word matching using single closed contours for indexing handwritten historical documents, Int. J. Doc. Anal.Recognit. 9 (2) (2007) 153–165.

[12] J. Rodriguez and F. Perronnin, "Handwritten word-spotting using hidden Markov models and universal vocabularies," Pattern Recognition, vol. 42, no. 9, pp. 2106–2116, 2009.

[13] E. Kavallieratou, N. Dromazou, N. Fakotakis, G. Kokkinakis, An Integrated System for Handwritten Document Image Processing, *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 17, No. 4, pp. 101-120, 2003.

[14] Nikos Vasilopoulos, Ergina Kavallieratou, "A classification-free word-spotting system", Proc. SPIE 8658, Document Recognition and Retrieval XX, 2013.

[15] V. Papavassiliou, T. Stafylakis, V. Katsouros, G. Carayannis, "Handwritten Document Image Segmentation into Text Lines and Words, Pattern Recognition", Vol.43, No.1, pp. 369-377, 2010.

[16] http://www.unipen.org/trigraphslant.html