RECOGNITION DRIVEN PAGE ORIENTATION DETECTION

Yves Rangoni[1], *Faisal Shafait*[1], *Joost van Beusekom*[2], *Thomas M. Breuel*[1,2]

Image Understanding and Pattern Recognition (IUPR) Research Group
[1]German Research Center for Artificial Intelligence (DFKI) GmbH
[2]Technical University of Kaiserslautern
D-67663 Kaiserslautern, Germany
{rangoni,shafait,van_beusekom,breuel}@dfki.uni-kl.de

ABSTRACT

In document image recognition, orientation detection of the scanned page is necessary for the following procedures to work correctly as they assume that the text is well oriented. Several methods have been proposed, but most of them rely on heuristics of the script such as the graphical asymmetry between ascenders and descenders for Roman script. The literature shows that as soon as this assumption is not fulfilled, e.g. plain capital text, noisy or degraded characters, etc. they fail. For a large-scale digitalization process, a low error and rejection rate are expected in order to reduce the amount of human intervention. We propose a Recognition Driven Page Orientation Detection (RD-POD) which does not depend on external criteria or assumption on the shape of the script. It uses the OCR engine for estimating the right orientation with a few lines of the document image. The RD-POD is highly robust and accurate, and is able to detect multiple orientations. Experimental evaluation shows that our method outperforms the current state-of-the-art on UW-1 dataset with an accuracy of 99.7%. Further tests on other three large and public datasets (MARG, ICDAR07, Google 1000 books) show accuracies of above 99% on each of them.

Index Terms— Image orientation analysis, Document image processing, Optical character recognition

1. INTRODUCTION

The problem of page orientation detection (POD) deals with finding the correct orientation so that the characters are in an upright position. Four main orientations are considered: 0°, 90°, 180°, and 270°. For small office applications, the problem is being tackled directly by the user that tries to scan his pages correctly or manually rectifies the small amount of miss-oriented pages. With the growing number of large digitization initiatives, such as Google Books [1], with millions of paper-based materials, the human intervention is not desired for efficiency and also for financial cost reasons. We search for a method that outputs no errors to not corrupt the following recognition steps, since usually both layout analysis [2] and optical character recognition [3, 4] assume pages to be in the correct orientation. The method should also reject as few as possible, to keep the flow as fast and automatic as possible.

Most of the proposed approaches are only able to distinguish between portrait and landscape instead of the four possible orientations that can occur when batch digitizing a large corpus. Cattoni et al. [5] give an overview of the state of the art methods back in 1998. Their survey concludes by pointing two main drawbacks of the 24 observed techniques. First, they suffer from their inaccuracy and their bad computational efficiency. Secondly, they make use of strong assumptions about a class of document and do not provide sufficient generalization capacities. None of the presented methods is designed to determine if a page is top down or not (only $-90°$ and $+90°$). Comparisons are also difficult to establish since the results are given for small and synthetic/private corpus of documents; none of them makes use of a public dataset. For example, Akiyama et al. 1990 [6] report 3 errors among 33 documents with the use for example the vertical and horizontal variances of projection profiles. Hinds et al. 1990 [7] report 8 of the 13 documents with the use counts of short run-lengths in the vertical and horizontal histograms. Le et al. 1994 [8] obtain an error rate of 0.07% on a non-public dataset of 6,087 pages of medical journal with rules based on projection profiles and the Hough transform.

Several researchers [9, 10, 11, 12, 13] have presented page orientation detection techniques for Roman script based on the ascender-descender ratio of text. Since ascenders are more frequent than descenders, these techniques rely on counting the numbers of ascenders and descenders in the text to make a decision about its orientation. However, most of the authors [10, 11, 12] report results on private datasets so a direct comparison of these techniques is not possible. Only [9, 13] report results on the public UW-I [14] dataset and bring an accuracy of 95% and 99% respectively. One major drawback of these techniques is that they can not cope with the documents where all text is in upper-case letters.

The table 1 summarises some results reported by different authors during the last two decades. Only [9] and [13] used the public UW-I dataset. Others are not reproducible.

In this paper, we introduce a fully automatic and fast method for finding the main orientation of a document, based

978-1-4244-5654-3/09/$26.00 ©2009 IEEE

Authors	Year	Databases	Error	Reject
[6]	1990	journals (33 pages)	9.09 %	0.00 %
[7]	1990	forms, magazines (13 pages)	38.46 %	0.00 %
[8]	1994	medical journals (6,087 pages)	0.07 %	0.00 %
[8]	1994	medical journals (5,190 pages)	0.08 %	0.00 %
[10]	2000	fax in Times 12pt (226 pages)	0.00 %	0.00 %
[11]	2005	scientific articles (22,140 pages)	0.95 %	0.00 %
[19]	2006	printed documents (492 pages)	2.44 %	0.00 %
[12]	2007	digital library (52 pages)	5.77 %	0.00 %
[9]	1995	UW-I (979 pages)	0.11 %	4.23 %
[13]	2009	UW-I (979 pages)	0.92 %	0.00 %

Table 1: Overview of the results obtained by different authors

on OCR text recognition performance. The quality of the transcription in the four orientations is used as a score for deducing the main orientation and eventually pages containing text in several orientations. The *recognition driven page orientation detection* (RD-POD) method has major key points. It is highly robust even on degraded documents and can be efficiently integrated in the recognition process flow. It does not rely on heuristics of the script, and it can work in difficult cases even if the page contains only one line of text. The next section describes in detail the steps of the RD-POD. Then, experimentations will show how it works well on different public datasets and outperforms the current state of the art. Finally, conclusions and perspectives will be discussed.

2. RECOGNITION DRIVEN DETECTION

The recognition driven page orientation detection (RD-POD) consists in few steps: binarizing, extracting some lines in the four possible orientations, performing a line recognizer on the lines, and evaluating if the obtained transcriptions match or not a language modelling to deduce the orientation. The RD-POD easily fits into a natural pipe of tasks that most of the OCR systems follow [2, 3, 4]. RD-POD just needs inputs that must be performed anyway during the recognition flow.

The two main steps, fast line extraction and orientation evaluation based on line transcription, require a binarization since they assume that the page is black and white. We chose Sauvola's method [15, 16] as several comparison surveys, like [17, 18], suggested that it outperforms all the other approaches when the target is document recognition.

2.1. Fast line extraction

In order to evaluate the page orientation, the quality of the recognition will tested on a few lines of the document in the four possible orientations. Contrary to previous methods, the focus is put on the lines of the text, and no time is wasted in optimizing a criterion on needless part of the document like pictures or drawings for example. The goal is to quickly find a representative subset of line images, and only evaluate orientation on this small part of the full document.

A fast and robust method for having such a subset of text lines is the geometric text-line model proposed by [20]. In one pass, it can find skew using the RAST algorithm [13, 21], and then a geometric matching is applied to extract text-lines from the binarized image. The interesting property lies in its capacity of working with a targeted number of lines and the results are returned in decreasing order of quality. More importantly, it successfully ignores lines originating from marginal noise [22] since they are short in length and hence have a low quality. The amount of computation to extract only the n-best lines, is small compared to the full processing of the page. On top of that, most of the pre-computations are still valid to obtain the other part of the non-extracted lines for the further step of the document recognition.

2.2. Orientation evaluation based on line transcription

After obtaining four subsets of line images S_{rot} in the four possible rotations $rot \in \{0, 90, 180, 270\}$, the next step is to find the right one. Each rotation has a score, estimating if there is real text in that rotation. The evaluation is goal-directed: transcriptions are obtained with a line recognizer. In this paper, we are dealing with Tesseract [3], a raw OCR engine. It performs efficient shape matching and uses heuristic search to cope with touching and broken characters. If the lines are in correct orientation, the OCR output fits, or is close to fit, a language modelling. Otherwise, the output is only garbage, with "random" and non-common characters, producing unlikely transcriptions.

We have tested several methods to make the distinction between a text and a line of garbage characters. Language models based on trigrams have been thought to be good candidates and able to deal with several languages [23]. Edit distance based matching with dictionary can also work. Some experimentation shows that strict comparison with a dictionary is the best choice. Indeed, even if it gives sometimes slightly worse score for the correct orientation, on the other hand, it strongly penalises the incorrect orientations, such as the gap between text and non-text is larger. Additionally, with a binary search, computing the cost function is negligible even for a large dictionary. Let D a set of words representing the dictionary, we propose to cost a line with:

$$cost(line) = length(line)^{-1} \sum_{\substack{word \in line \\ word \in D}} length(word)$$

When a line really contains text, then $cost(line)$ is close to 1. Otherwise, the score is close to zero (and most of the time equals to zero), when the line is just garbage characters. The main orientation MO over four angles is simply the one giving the higest score for all the lines in the set:

$$MO = \underset{rot}{\operatorname{argmax}} \left(\sum_{line \in S_{rot}} cost(line) \right)$$

Let define SO_{rot} the maximum score for one orientation:

$$SO_{rot} = \max_{line \in S_{rot}} cost(line)$$

The ambiguity detection AD, indicating if a page contains multiple orientations, is defined as:

$$AD = min(\{|SO_{r_1} - SO_{r_2}|, r_1 \neq r_2\})$$

Method	Dataset	Pages	Error	Rejection
RD-POD	Alice	44	0.000%	0.000%
	ICDAR07	40	0.000%	0.000%
	MARG	1553	0.064%	0.064%
	G1000	740	0.000%	0.676%
	UW3	1600	0.000%	0.313%
[9]	UW1	979	0.110 %	4.227 %
[13]	UW1	979	0.928 %	0.000 %

Table 2: Error and rejection rates for 5 datasets. RD-POD outperforms the state of the art on UW dataset

If AD is close to 0, then there is an ambiguity between at least two orientations. If r_1 and r_2 are respectively the best and the second best orientations, then a small AD indicates that the page is probably multi-oriented, or at least, contains some lines in a different orientation than the main one.

The values SO_{MO} and AD can be checked to accept or reject a document (eg. SO_{MO}=0 suggest that the page contains no text at all). Due to the robustness of the *cost* function, small values of SO_{MO} have never been seen, only $AD = 0$ can occurs when one or several lines are not following the main orientation (eg. lot of figure captions in 90°). AD can be used as a detector for multi-oriented pages.

3. EVALUATION AND RESULTS

3.1. Description of the datasets

In order to evaluate the RD-POD, we used 4 public datasets:

- UW3, University of Washington III Dataset [14], contains 1600 binarized images of scientific journals
- MARG, Medical Article Records Groundtruth [24], 1553 binarized images of the first page of journals
- ICDAR07, 40 training and test pages from ICDAR 2007 page segmentation competition [25]
- G1000, 740 pages from the inner sections of each English volume have been picked form the Google Book Search Dataset [1]. It is composed of scans of old books for which copyrights have expired.

One other dataset has been generated: "Alice's Adventures in Wonderland" book, rendered with LaTeX, in Times 12pts, printed and scanned at 300 dpi for a total of 44 pages. A subset of 3 lines and an English dictionary have been chosen to evaluate the orientation for all the datasets. Table 2 gives the results obtained on those 5 datasets. The open source project OCRopus [4] 0.3.1 has been used to run the experiments using Tesseract [3] 2.0.3 as the line recognition engine.

3.2. Results

There is no error and no rejection at all for Alice and ICDAR07. The scores characterizing the right orientation are around 90% and 0% for the 3 others; no ambiguity is possible. One line would have been sufficient to obtain these results.

For MARG, only one page is not correctly oriented, and another page is rejected (no orientation found). They are mainly due to an incorrect extraction of the 3 best lines, which causes a bad recognition in the right orientation. Note that the error comes from the line extractor and not the RD-POD method. By choosing 4 lines instead of 3, both error and rejection disappear. Less than a dozen differences are quite small (ambiguities) due to a subset of lines picked in the affiliations which are not in English and contain lot of proper names.

For G1000, although this kind of document is really hard to recognize with an OCR, the extraction of 3 lines with a words-in-dictionary ratio produces no errors for orientation detection. The figure 1 presents and explains the 5 rejections.

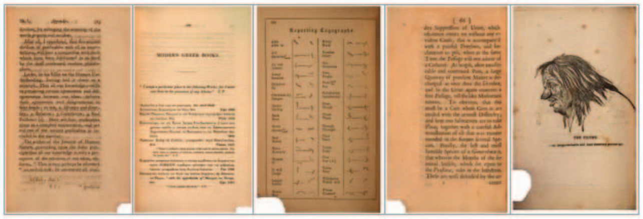
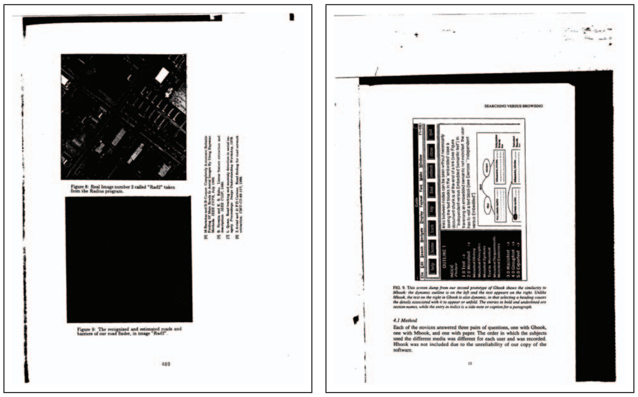


Fig. 1: Rejections for G1000: first image contains short lines, with non-Latin symbols and transparency of the verso, image 2 is mainly in Greek, image 3 is a table with few words, image 4 is similar to image 1 with a lot of "medial S", image 5 contains only 2 lines of highly degraded text

For the UW3, there are no errors, and 5 pages are rejected as they contain 2 main orientations like in Fig. 2a. Some pages are also well detected but with low scores, small differences of scores are also due to a second orientation in the page e.g. some text is 90° oriented in figures or in tables (Fig. 2b). The UW-I (979 pages) is a subset of the UW-III (1600 pages), we can compare with [9] that produces low error (1 page) and reject a lot (41 pages), and with [13] that does not reject at all but produces more errors (9 pages) (Tab. 1). On a wider UW, we reject only 5 pages and produce no errors at all. The rates of correct orientation [13] (correct orientations vs. dataset size) are respectively 95.71%, 99.08%, and 99.69% for [9], [13], and us.

A final test has been performed for the UW-III. It consists in randomly changing the orientation of the page and applying a random skew with angles in a range of $[-1, 1]°$. The same procedure has been applied. We obtained one error and 7 pages were rejected. In all cases, the rejection is due to two orientations in the same page. For the unique error, (HO4EBIN), RD-POD was not able to find the main orientation and answer 90° because of the y-axis captions of the figures where the transcriptions have been perfectly recognized.

The RD-POD requires only few extra computations since the final aim is the page recognition and RD-POD is completely integrated in the work-flow. Rotating the image and evaluating transcriptions cost nothing. Finding and extract-



(a) W1H4BIN (b) D05MBIN

Fig. 2: Two samples where the RD-POD founds ambiguities

ing the lines with RAST must be performed four times, but as only three lines are requested, the total time is equivalent to one pass on the full document. The recognition of the three lines takes less than 10% of the global recognition.

4. CONCLUSIONS

In this paper, we presented a recognition driven method for page orientation detection. A small subset of text lines are extracted in the four orientations, and then the output of the OCR is evaluated to deduce the right orientation. The method can handle documents containing few and also small lines of text and it is able to detect multiple orientation is the same image. When the method is integrated in a full document image recognition work-flow, it benefits from the outputs of other sub-tasks so that it performs fast and helps the further process. For easy cases, such as a well scanned book, only one line is enough to detect perfectly the orientation. The technique has been evaluated on four public datasets and outperforms the current state-of-the-art methods on the large UW-III dataset. With few and comprehensible rejections, the method easily reach 0% for error rate, even on degraded images like the Google 1000 books dataset.

References

[1] L. Vincent, “Google book search: document understanding on a massive scale,” *Int. Conf. on Document Analysis and Recognition*, pp. 819–823, 2007.

[2] F. Shafait, D. Keysers, and T.M. Breuel, “Performance evaluation and benchmarking of six page segmentation algorithms,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 941–954, 2008.

[3] R. Smith, “An overview of the Tesseract OCR engine,” *I. C. on Document Analysis and Recognition*, pp. 629–633, 2007.

[4] T.M. Breuel, “The OCRopus open source OCR system,” *SPIE Document Recognition and Retrieval XV*, pp. 0F1–0F15, 2008.

[5] R. Cattoni, T. Coianiz, S. Messelodi, and C.M. Modena, “Geometric layout analysis techniques for document image understanding: a review,” *Technical Report*, pp. 9703–09, 1998.

[6] T. Akiyama and N. Hagita, “Automated entry system for printed documents,” *Pattern Recognition archive*, vol. 23, pp. 1141–1154, 1990.

[7] S.C. Hinds, J.L. Fisher, and D.P. D’Amato, “A document skew detection method using run-length encoding and the hough transform,” *10th Int. Conf. on Pattern Recognition*, pp. 464–468, 1990.

[8] D.S. Le, G.R. Thoma, and H. Wechsler, “Automated page orientation and skew angle detection for binary document images,” *Pattern Recognition*, vol. 27, pp. 1325–1344, 1994.

[9] D.S. Bloomberg, G.E. Kopec, and L. Dasari, “Measuring document image skew and orientation,” *Proc. SPIE Document Recognition II*, pp. 302–316, 1995.

[10] R.S. Caprari, “Algorithm for text page up/down orientation determination,” *Pattern Recognition Letters*, vol. 21, pp. 311–317, 2000.

[11] B.T. Avila and R.D. Lins, “A fast orientation and skew detection algorithm for monochromatic document images,” *ACM symposium on document engineering*, pp. 118–126, 2005.

[12] S.J. Lu, J. Wang, and C.L. Tan, “Fast and accurate detection of document skew and orientation,” *Int. Conf. on Document Analysis and Recognition*, vol. 2, pp. 684–688, 2007.

[13] J. van Beusekom, F. Shafait, and T.M. Breuel, “Resolution independent skew and orientation detection for document images,” *Proc, of SPIE Electronic Imaging: Document Recognition and Retrieval*, vol. 7247, pp. 72470K–72470K, 2009.

[14] I.T. Phillips, “User’s reference manual for the UW english/technical document image database III,” *Seattle University, Washington*, 1996.

[15] J. Sauvola and M. Pietikinen, “Adaptive document image binarization,” *Pattern Recognition*, vol. 33, pp. 225–236, 2000.

[16] F. Shafait, D. Keysers, and T.M. Breuel, “Efficient implementation of local adaptive thresholding techniques using integral images,” *Proc, of SPIE Electronic Imaging: Document Recognition and Retrieval*, vol. 6815, pp. 81510–81510, 2008.

[17] M. Sezgin and B. Sankur, “Survey over image thresholding techniques and quantitative performance evaluation,” *Journal of Electronic Imaging*, vol. 13, pp. 146–165, 2004.

[18] E. Badekas and N. Papamarkos, “Estimation of proper parameter values for document binarization,” *Int. Conf. on Computer Graphics and Imaging*, vol. 10, pp. 600–037, 2008.

[19] S. Lu and C.L. Tan, “Automatic document orientation detection and categorization through document vectorization,” *ACM international conference on multimedia*, pp. 113–116, 2006.

[20] T.M. Breuel, “Robust least square baseline finding using a branch and bound algorithm,” *Proc. SPIE Document Recognition and Retrieval IX*, pp. 20–27, 2002.

[21] T.M. Breuel, “Implementation techniques for geometric branch-and-bound matching methods,” *Computer Vision and Image Understanding*, vol. 90, pp. 258–294, 2003.

[22] F. Shafait, J. van Beusekom, D. Keysers, and T.M. Breuel, “Document cleanup using page frame detection,” *Int Journal on Document Analyis and Recognition*, vol. 11, no. 2, pp. 81–96, 2008.

[23] S.F. Chen and J.Goodman, “An empirical study of smoothing techniques for language modeling,” *Meeting on Association for Computational Linguistics*, pp. 310–318, 1996.

[24] G. Ford and G.R. Thoma, “Ground truth data for document image analysis,” *Symposium on Document Image Understanding and Technology*, pp. 199–205, 2003.

[25] A. Antonacopoulos, B. Gatos, and D. Bridson, “ICDAR 2007 page segmentation competition,” *9th Int. Conf. on Document Analysis and Recognition*, pp. 1279–1283, 2007.