Document Inspection Using Text-Line Alignment

Joost van Beusekom
University of Kaiserslautern
German Research Center for Artificial Intelligence (DFKI)
Trippstadterstraße 122
67663 Kaiserslautern, Germany
joost.van_beusekom@dfki.de

Faisal Shafait
German Research Center for Artificial Intelligence (DFKI)
Trippstadterstraße 122
67663 Kaiserslautern, Germany
faisal.shafait@dfki.de

Thomas M. Breuel
University of Kaiserslautern
Trippstadterstraße 122
67663 Kaiserslautern, Germany
tmb@cs.uni-kl.de

ABSTRACT

Passports, ID cards, banknotes, and degrees are considered as valuable documents that need to be secured against forgery. Apart from those, there are many other document types that are valuable, too, but that do not have any security features, as e.g. bills and vouchers. These may be used by fraudulent people to defraud money from e.g. a car insurance company. The wide availability of scanning and printing hardware allows even non-experts to easily forge a document. We therefore present a new aspect in the examination of intrinsic document features for optical document security: the goal is to automatically detect text-lines that have been manipulated or additionally inserted in a document by inspecting their alignment (left, right or center) with respect to the other text-lines in the document. This constitutes an additional feature in the goal of developing a powerful toolbox for automatic document inspection. Using the extracted text-lines, the alignment margins are extracted. Statistics on the distances of the text-lines to the alignment margins are used to identify lines that might have been forged. Such documents can then be presented to a human operator for further inspection. Due to lack of public datasets containing forged documents, a new dataset had to be created. Evaluation showed a classification accuracy of 90.5%.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

1. INTRODUCTION & PREVIOUS WORK

When talking about valuable documents, most people think of passports, ID cards, banknotes, or diplomas. However, it is widely ignored that even every-day's documents like bills and vouchers may be forged to gain financial advantages. Van Renesse [13] distinguishes five different document value types:

- direct value: documents giving access to unconditional and immediate value, e.g. banknotes.
- indirect value: documents that support a transaction or a right, e.g. diplomas and passports
- conditional value: documents that give access to a value after the document has been inspected, e.g. admission tickets or cheques
- informative value: documents that have no immediate value apart from the informations that it contains, e.g. confidential reports
- fictious value: documents that have no immediate value and that do not contain any valuable information, e.g. stationeries of institutions or companies

Among the class of documents with conditional values, documents of every-day's life can be found, as e.g. bills and vouchers. These documents may be exchanged to get access some services or money, as in the case of a bill of a garage that is sent to the insurance company.

There is thus a motive to forge documents to defraud a value. This motive in combination with the fact that nowadays a broad public has access to scanning and printing techniques that allow even untrained users to easily modify or generate forged documents, is likely to lower the inhibition threshold to actually do so. Unfortunately the authors did not find any references in literature that state or estimate how many documents are being forged in every-day's life. However, due to the previously given reasons, we assume that there is an important number of document forgeries in every-day's life.

Due to the wide diversity of documents and documents generating sources, it is not feasible to add a set of security features to each of these documents. Neither is it possible to perform a *third line inspection* [13] where specialists check each and every document. In the ideal case, a *first line inspection*, consisting of visual inspection without any additional means should be enough. However even simple document forgeries are hard to detect by bare eye. Therefore, so called *second line inspection*, using machines to testify a documents originality could be suitable. However, van Renesse [13] argues that due to missing standardization it is nearly impossible to have an automatic tool to verify all

DAS '10, June 9-11, 2010, Boston, MA, USA

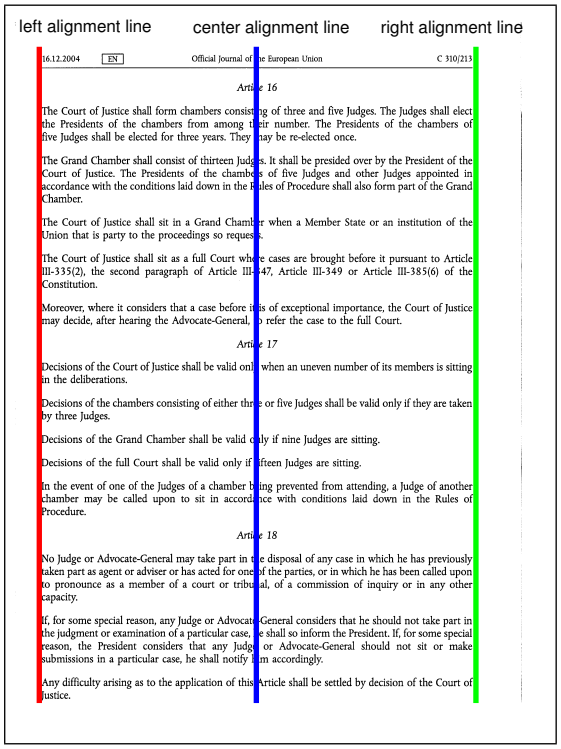


Figure 1: Visualization of the left, center and right alignment lines.

sorts of documents. Therefore, we plan to develop a toolbox that allows the user to check a document for several intrinsic document features that are likely to show significant changes when examining a document that was forged using standard PC hardware. This will then be used to make a plausibility check for a document so that anomalies are detected. The advantage is that we use features that are intrinsic to the document and are present for every document type, so no special watermarks have to be added to the document.

In this paper we propose a method to detect the alignment of text-lines and whether the distance to the alignment boundary is *normal* or *suspicious*. We therefore developed a method to detect the so called left, right and center alignment line. A visualization of the alignment lines can be found in Figure 1. Using these lines, the distances of the starting, middle and end points of the text-lines to the respective alignment lines are computed. Statistical measures are then used to decide if a given text-line is suspicious or not.

Much work has been done in the domain of optical document security. Most approaches however try to add some external security features that should be hard to forge. Examples are specialized paper, holographic images [8], specialized printing techniques [1] and other physical and chemical signatures [4]. To our best knowledge, apart from our previous contributions [9, 10, 6] no previous work has been published that uses intrinsic document layout features for automated document security purposes.

The paper is organized as follows: Section 2 give a short description over the proposed approach. In Section 3, the text-line extraction method used to detect text-lines in the image is shortly explained. Section 4 explains the approach to detect the alignment lines. Section 5 describes how the alignment information can be used to detect forged documents. Evaluation and results are to be found in Section 6. Section 7 concludes the paper.

2. DESCRIPTION OF THE APPROACH

The scenario of application of our approach is the following: assume that someone wants to forge a bill or a contract by adding a paragraph or a single line of text. One possibility to do so would be to scan the document, edit the image using some image manipulation software, add the new line to the document and print it out again. The other possibility would be to take the original document and print the line or the paragraph in a second pass on the empty space of the original document. The third possibility would be to print the modified text on an empty sheet, cut it out and paste it over the original document. Then a copy is made so that the pasted piece of paper is not visible anymore. In all cases the fraudulent person will try to put the text-line in such a way that it looks like original, thus also conserving the alignment. Although the fraudulent person will try to get the alignment as exact as possible, technical and manual constraints make it unlikely that he achieves an accurate alignment.

This is where our method approaches: it tries to identify lines that are slightly misaligned to the rest of the lines. The hypothesis is that for original documents, the distances of the start, center and end points of the text-lines to the respective alignment lines follow some distribution that has a peak for very short distances (approximately zero pixels), followed by a valley and and then some sporadic peaks that represent indented lines. The peak around zero pixels originates from the aligned text-lines. The slight variations might be due to pixel noise, binarization noise or some characters having some parts that are not exactly lying on the alignment line.

As the variations produced by forging will likely be higher than the document intrinsic variations, it is expected, that this will result in a wider variability of distances around the zero pixel distance. This can be used to detect suspicious lines.

The proposed approach works as follows: first, the text-lines are extracted. Then the alignment lines are computed. Finally the distances of the text-lines to the alignment lines are computed and based on these distances a decision is made whether a text-line is normally aligned or not. A visualization of the approach can be found in Figure 2.

It should be noted that not all forgeries can be detected using this method. If the fraudulent person is an expert or if by chance the alignment is exactly the same as for the other text-lines, than the proposed approach will not be able to detect the forged text-lines. In the end, many different features will have to be combined, as e.g. the alignment feature, the text-line orientation feature [10] and the text-line spacing.

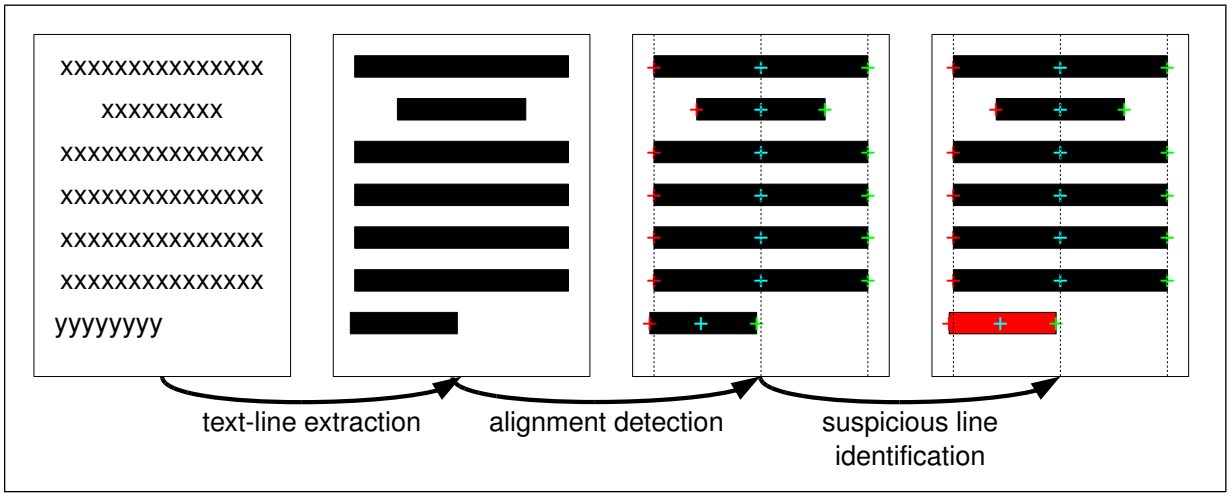


Figure 2: Visualization of the approach for detecting suspicious non-aligned text-lines.

Figure 3: The anatomy of a text-line showing positions where characters align from top and bottom.

3. TEXT-LINE FINDING

Although several text-line finding algorithms have been reported in literature [7], we use the text-line extraction presented by Breuel [2] due to its ability to give geometric parameters of detected text-lines. For the sake of completeness we will shortly explain the text-line finding method, as a similar approach with a different line model is used to solve the alignment line detection.

The overall idea is to search the parameter space given by the text-line parameterization for parameters that best describe the text-lines given a quality function.

The text-line model is the same as used in [2]. An illustration can be found in Figure 4. The following three parameters are used: (r, θ, d), where r denotes the y-intercept, θ the angle of rotation of the line and d the distance of the base line to the line of descenders. Figure 3 illustrates the base line and the line of descenders.

The goal of the quality function is to guide the parameter search to find good solutions that represent text-lines. As feature points, the centers of the lower line of the bounding box of the connected components are used (see Figure 5). The quality function should give a good quality for parameters that match many feature points to the line, either to the base line or to the descender line.

The search itself follows a branch-and-bound approach. The initial parameter space (search space) is recursively subdivided into subspaces, for which upper bounds for the quality function are computed. Using this upper bound, the most promising subspaces are subdivided first. Finally, when the subspace to be analyzed is small enough, it is returned as a text-line.

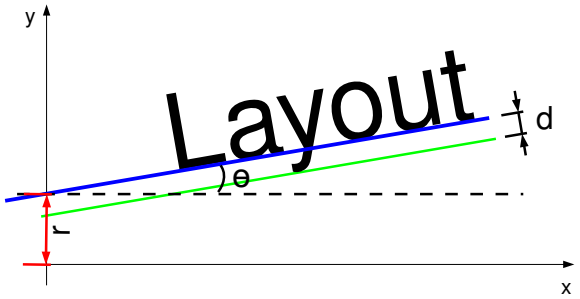


Figure 4: An illustration of Roman script text-line model proposed by Breuel [2]. The baseline is modeled as a straight line with parameters (r, θ), and the descender line is modeled as a line parallel to the baseline at a distance d below the baseline.

4. ALIGNMENT DETECTION

After the text-line extraction, the alignment lines have to be detected. Four different alignment types are commonly distinguished in typesetting:

- left aligned: text-lines start at the left margin
- right aligned: text-lines end at the right margin
- justified: text-lines start at the left margin and end at the right margin
- centered: text-lines do neither touch the right nor the left margin. The gaps between both margins are equal in size.

Thus, in order to find the alignment lines, we have to analyze the starting, the end and the center points of the text-lines.

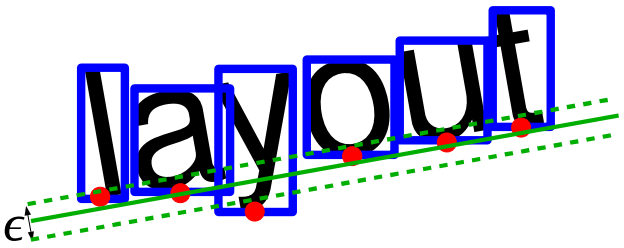


Figure 5: An illustration of the image points used for text line extraction. The effect of ϵ is visualized by the dotted lines: every reference point in between the dotted line contributed to the line.

The left alignment line is defined as a vertical line where left-aligned and justified text-lines have their starting point. The center alignment line is the vertical line that contains the center points of text-lines (centered text-lines as well as justified ones). The right alignment line is analoguesly defined as the vertical line where right-aligned and justified text-lines have their end point.

Finding these three lines is done using RAST line finding [3], similar to the text-line finding explained in Section 3. The key difference is that no descender line has to be modeled and that the parameterization of the line is different. Instead of the (r, θ) parameterization, where r is the y-intercept and θ is the rotation angle with respect to the horizontal, we adopt the polar representation $\vartheta = (|\overrightarrow{n}|, \alpha)$, where $|\overrightarrow{n}|$ is the norm of the vector that is normal to the line and that points to the origin, and α is the rotation angle of $\overrightarrow{n}$.

The advantage of this parameterization is that is well suited for searching for vertical lines, in contrast to the previously presented model used in text-line finding.Consider feature points $\{x_1, x_2, \cdots, x_n\}$: the quality function can then be written as:

$$\hat{\vartheta} := \arg\max_{\vartheta} Q_{x_1^n}(\vartheta) \tag{1}$$

where

$$Q_{x_1^n}(\vartheta) = Q_{x_1^n}(|\overrightarrow{n}|, \alpha) = \sum_{i=1}^{n} (q_{(|\overrightarrow{n}|,\alpha)}(x_i)) \tag{2}$$

and

$$q_{(|\overrightarrow{n}|,\alpha)}(x) = \max\left(0, 1 - \frac{d^2_{(|\overrightarrow{n}|,\alpha)}(x)}{\epsilon^2}\right) \tag{3}$$

and where $d_{(|\overrightarrow{n}|,\alpha)}(x)$ is the Euclidean distance of a point x to the line defined by parameters $(|\overrightarrow{n}|, \alpha)$, and ϵ controls the maximum distance that a component can have from the alignment line in order to be considered as belonging to the line.

The line finding is run three times, each time with different feature points: starting points, center points and end points of text-lines. The left, center and right alignment lines are considered as the respective resulting highest quality lines.

5. DECISION MAKING

Using the alignment lines, the distances of the respective text-line points to the alignment lines can be computed. According to the assumption in Section 2, the distributions of these distances should give a high peak around zero (coming form the aligned lines) and some peaks further away that correspond to indented lines. Also, distances of only a few pixels to the alignment lines should be unlikely. The experimental validation of this hypothesis can be found in Section 5.1.

The distribution of the distances of the forged lines however, cannot be measured due to missing real-world data. However, the following reasonable assumption can be made: the forging person tends to get the forged lines as accurately aligned to the others as possible. Technical variations and also the forging persons inaccuracies will make an exact alignment difficult. Therefore we can model the distribution of the distances of the forged lines as normally distributed with the mean equal to zero and a standard deviation of about 10pix, which corresponds to about approximately half the width of the x character for normal text sizes (around 12pt) and a scan resolution of 300dpi. Higher variations are very easily detectable using *first line inspection* and should thus not be encountered frequently.

Using Bayes theorem, we can formulate the problem as follows:

$$P(f|d) = \frac{p(d|f) * P(f)}{p(d)} \tag{4}$$

where $P(f|d)$ is the probability of having a forged line given a distance d between the line's start, center or end point to the left, center or right alignment line. $p(d|f)$ is the likelihood of observing a distance d given the information that the line is forged. This is considered to be normally distributed with mean μ_f and standard deviation σ_f. Thus:

$$P(f|d) = \frac{\mathcal{N}(d, \mu_f, \sigma_f) * P(f)}{p(d)} \tag{5}$$

The parameter $P(f)$ is the prior for observing a forged line. Unfortunately, this is not known in advance. It can be considered as a sensitivity parameter that is set manually. Similarly, the probability for observing an original line given a distance d can be written as:

$$P(\neg f|d) = \frac{p(d|\neg f) * P(\neg f)}{p(d)} \tag{6}$$

where $P(\neg f) = 1 - P(f)$. The likelihood is computed from the measurements on the training data.

Finally, a line is considered as forged if $P(f|d) > P(\neg f|d)$ and vice-versa. In this setting, $p(d)$, which is actually unknown, can be ignored as it contributes in the same way to both terms.

For each text-line, we obtain three decisions, one for each alignment line and the corresponding text-line point. We consider a line as forged if at least one of the three decisions votes for a forged line. Note that this decision is intended to

mark a document as suspicious so that it can be inspected further by a human operator. Based on the costs involved in making a wrong decision, the decision rule can be made more strict or relax.

5.1 Hypothesis Validation

In order to check the hypothesis, the above mentioned distances have been measured on the synthetic dataset. Figure 6 shows the histograms of the distances between the starting points and the left alignment line, between the center point and the center alignment line and between the end points and the right alignment lines. It can be seen that our assumption is verified quite clearly for the left and center alignment lines. For the right alignment line the distance histogram is also nicely peaked at zero, but it is slightly wider than for the left and center alignment line.

For the histogram of the distances to the left alignment line, a few small peaks with negative values can be observed. These are due to the rare cases where the correct left alignment line is not the same as the computed left alignment line, which is the line where most line starting points lie on. In the left and center histograms, peaks around 80 and 40 respectively can be observed. These come from indented text lines.

6. EVALUATION AND RESULTS

As, to the authors' best knowledge, there is no public dataset with forged documents, a new dataset had to be created: we took publicly available documents, in our case the proposed treaty for establishing a constitution for the European Union [1].

The hypothesis validation in Section 5.1 has been run on the full set of document images (485 document images) that were synthetically generated.

The training dataset which is used to extract the statistics of natural variations of the text-line distances to the alignment lines consists of 20 documents out of the 485 that were printed and scanned again. Binarization and deskewing was run to obtain horizontally aligned text-lines [5, 11, 12].

The test data is split into two parts: part one contains images that have been generated by printing on the same page twice: first the page was printed with its original content. 43 pages containing 938 text-lines[2] were generated using this procedure. In a second pass, additional text is added and printed on the page. The second part contains copies of documents where single lines (one or two per page) have been replaced by pasting a new piece of paper over the original line. This part contains 29 images, where 1030 text-lines were extracted. The ground truth in both cases consists of images containing black rectangles at the positions where forged lines are to be found. No line accurate labeling is done.

[1]http://eur-lex.europa.eu/JOHtml.do?uri=OJ:C:2004:310:SOM:EN:HTML

[2]The text-lines are extracted automatically. Depending on the settings of the text-line extraction, minor variations in the number of extracted lines can be observed.

The evaluation criterion is the number of correct classifications on text-line level: true positive (forged line classified as such), true negative (unforged line recognized as such), false positive (unforged line recognized as forged one) and false negative (forged line not detected as forged line) rates are measured.

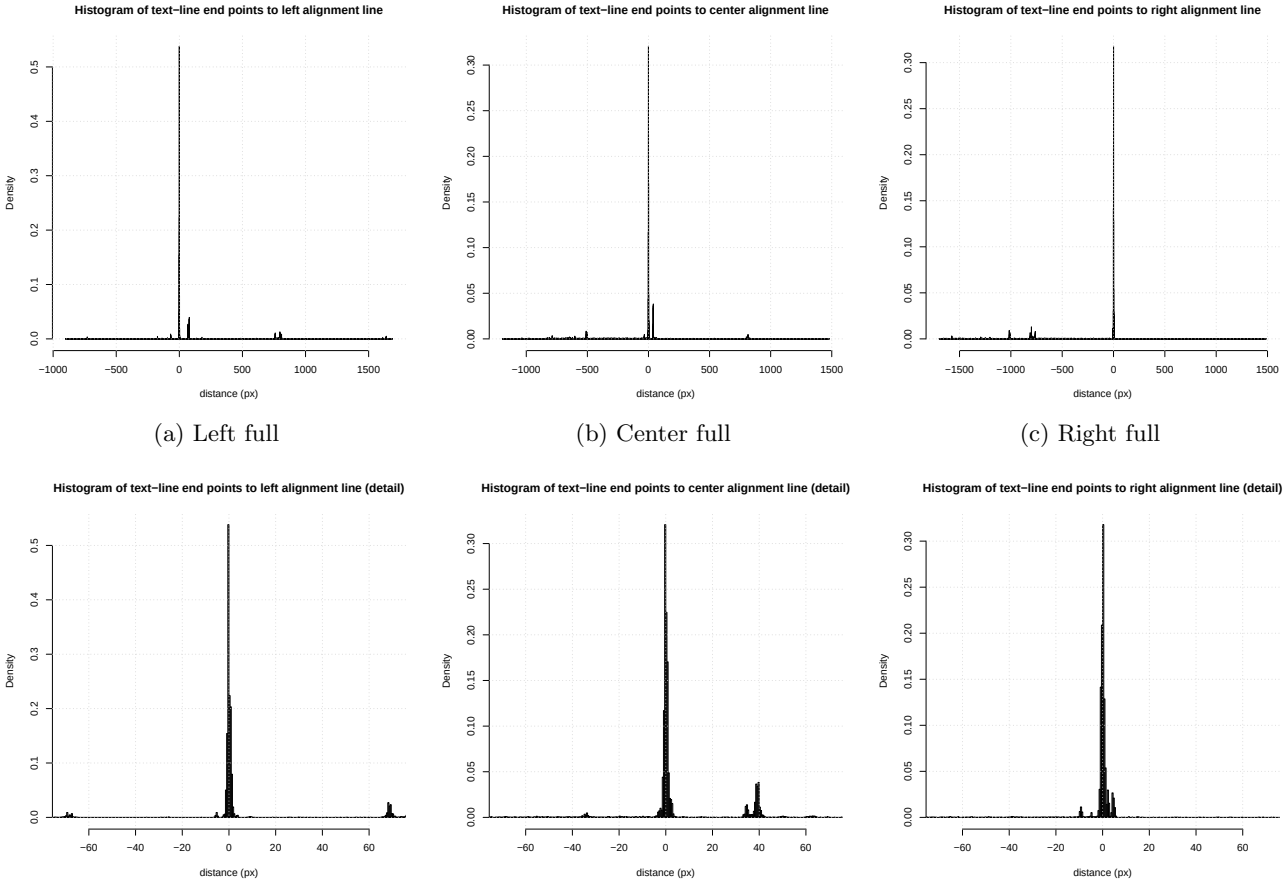
In Figure 10(a) the classification performance of the proposed method on the two pass printing dataset can be found. It can be seen, that for small values of the prior, the best performance can be obtained regarding the number of misclassifications: 90.5% of the lines are correctly classified for $p(f) = 0.001$. For $p(f) = 0.2$, the number of false negatives (forged lines that are not being detected) is low (7%) and does not diminish significantly anymore. It should also be noted that the number of false positives increases dramatically with increasing values for the prior, however, the the number of true negatives does not improve in the same way. An example of a resulting image can be found in Figure 9.

A visual analysis of the errors showed the following reasons for failures:

- "f" at the end of the line: probably due to optical correction reasons, the serif of character "f" extends outside the right margin as can be seen in Figure 7(b). This leads to an increase of false positives.
- near original alignment: in some cases the left, right and center alignment of the forged text-lines is so accurate that a detection using the alignment feature is not feasible. This problem leads to false negative classifications.
- wrong alignment line: when more indented text-lines are present than regular ones or if the number of forged text-lines is higher than the number of original text-lines, the wrong alignment lines may be found leading to misclassifications of unforged lines. This happens however in only a few cases, leading to both false positive and false negative classifications.
- errors in text-line finding: in rare cases text-lines extend to border noise or are split into two parts. Incorrect detection of the descender lines also occurs, but this does not influence the accuracy of the current method.

The problem with the characters extending outside the margin can be dealt by increasing the training set to also cover this problem. To detect forged lines that are accurately aligned to the margins is not possible by the proposed method. But a combination of other features, as e.g. the text-line rotation angles or the line spacing could help to solve these cases. Concerning the problem of detecting the wrong alignment line, the proposed method could be extended to extract more than just one alignment line. However, the distribution of the distances has to be reconsidered if the initial assumptions then still hold.

The results for the test on the second dataset containing copies of manually forged documents are given in Figure 10(b).



(a) Left full (b) Center full (c) Right full

(d) Left detail (e) Center detail (f) Right detail

Figure 6: Histogram of the distances of the line points to the alignment lines.

Compared to the results on the first dataset, is should be noticed, that the overall performance degraded. Also the number of false positives increases significantly even for small increasing in $p(f)$. A visual inspection of the resulting images showed several new problems that were not present in the previous test:

- split lines: the manual pasting of paper over existing lines results in lines with slight curl. The text-line finding splits these lines into two parts where typically one part is not reported as error, as either the start or the endpoint fits to the corresponding alignment line (see Figure 7(a)).
- copy process deformations: it was observed that even unforged text-lines where not exactly parallel. As this could not be observed for the original documents, we suppose, that the scanning process distorts the image in such a way that it makes near pixel accurate measurements impossible. A simple test making a 10th generation copy of a square showed these distortions quite clearly. The upper horizontal line of the 10th generation copy of the square can be seen in Figure 8.

7. CONCLUSION & FUTURE WORK

In this paper we presented a method for automated document security checks using the intrinsic feature of text-line alignment. It could be shown that this feature can be used for document security purposes and that it can be extracted and evaluated by automatic means. Furthermore, an evaluation on a self-generated dataset proved good results.

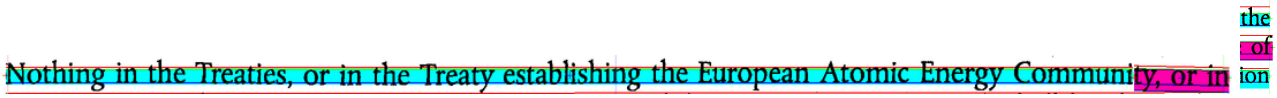
As mentioned before, the alignment feature is useful but it discriminative power is not high enough to detect all forged lines. Therefore we plan to combine this feature with previously published features in the hope to get even better results.

8. ACKNOWLEDGMENTS

This work was partially funded by the BMBF (German Federal Ministry of Education and Research), project PaREn (01 IW 07001).

9. REFERENCES

[1] I. Amidror. A new print-based security strategy for the protection of valuable documents and products using moire intensity profiles. In *Proc. of SPIE Optical Security and Counterfeit Deterrence Techniques IV*, volume 4677, pages 89–100, San Jose, CA, USA, January 2002.



(a) (b)

Figure 7: An example of a misclassified line due a line split that occurred in the text-line finding is shown in Figure 7(a). An example of an error due to an optical correction that leads to characters at the end of the line not being aligned is shown in Figure 7(b).

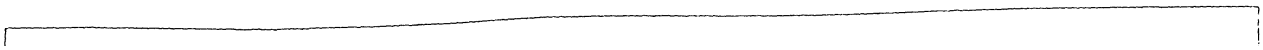
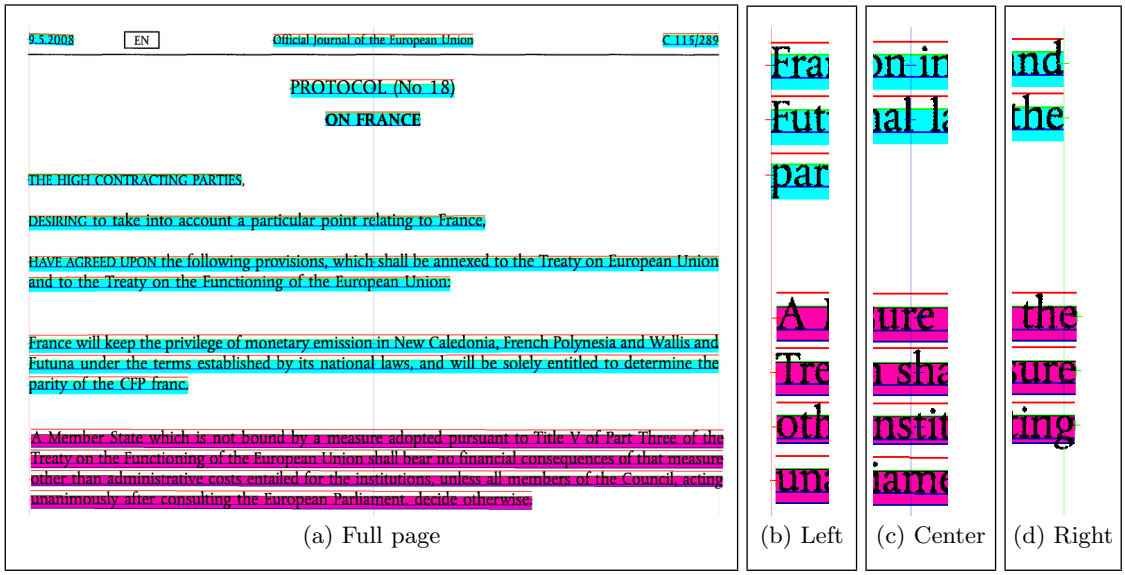
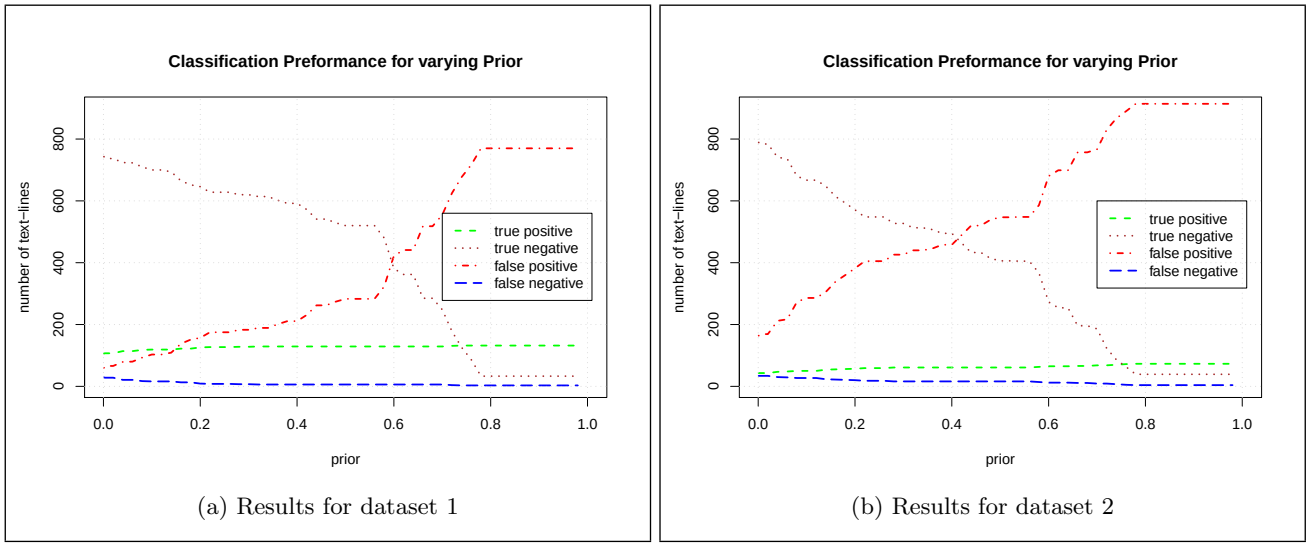


Figure 8: The upper, formerly straight line of a 10th generation copy of a square. It can be noticed, that the distortions are even visible with the bare eye. These distortions are also present in 1st generation copies, but these are normally not visible. However they lead to problems when an accurate measurement is needed.



(a) Full page (b) Left (c) Center (d) Right

Figure 9: Example of an analyzed document. The image on the left shows the full page content (white borders have been cropped). The images to the right show the left, center and right alignment line together with the start, center and end points of the text-lines (crosses). Lines that have been classified as forged are painted in red. Lines classified as original ones are painted in blue.



(a) Results for dataset 1

(b) Results for dataset 2

Figure 10: Plot of the classification performance in relation to the prior $p(f)$. True positive (forged lines recognized as such), true negative (unforged line recognized as such), false positive (unforged line recognized as forged one) and false negative (forged line recognized as original one) are plotted. The left image shows the results for the dataset containing images that passed two print steps. The image on the right shows the results on the dataset containing copies of manually forged documents.

[2] T. M. Breuel. Robust least square baseline finding using a branch and bound algorithm. In *Proc. of SPIE Document Recognition and Retrieval IX*, pages 20–27, San Jose, CA, USA, January 2002.

[3] T. M. Breuel. On the use of interval arithmetic in geometric branch-and-bound algorithms. *Pattern Recognition Letters*, 24(9–10):1375–1384, 2003.

[4] N. A. Hampp, M. Neebe, T. Juchem, M. Wolperdinger, M. Geiger, and A. Schmuck. Multifunctional optical security features based on bacteriorhodopsin. In *Proc. of SPIE Optical Security and Counterfeit Deterrence Techniques V*, volume 5310, pages 117–124, San Jose, CA, USA, January 2004.

[5] J. Sauvola and M. Pietikainen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000.

[6] C. Schulze, M. Schreyer, A. Stahl, and T. M. Breuel. Evaluation of graylevel-features for printing technique classification in high-throughput document management systems. In *Proc. of the 2nd Int. Workshop on Computational Forensics*, volume 5158 of *Lecture Notes in Computer Science*, pages 35–46, Washington, DC, USA, August 2008.

[7] 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*, 30(6):941–954, 2008.

[8] P. J. Smith, P. O'Doherty, C. Luna, and S. McCarthy. Commercial anticounterfeit products using machine vision. In *Proc. of SPIE Optical Security and Counterfeit Deterrence Techniques V*, volume 5310, pages 237–243, San Jose, CA, USA, January 2004.

[9] J. van Beusekom, F. Shafait, and T. M. Breuel. Document signature using intrinsic features for counterfeit detection. In *Proc. of the 2nd Int. Workshop on Computational Forensics*, volume 5158 of *Lecture Notes in Computer Science*, pages 47–57, Washington, DC, USA, August 2008.

[10] J. van Beusekom, F. Shafait, and T. M. Breuel. Automatic line orientation measurement for questioned document examination. In *Proc. of the 3rd Int. Workshop on Computational Forensics*, volume 5718 of *Lecture Notes in Computer Science*, pages 165–173, The Hague, The Netherlands, August 2009.

[11] J. van Beusekom, F. Shafait, and T. M. Breuel. Resolution independent skew and orientation detection for document images. In *Proc. of SPIE Document Recognition and Retrieval XVI*, volume 7247, San Jose, CA, USA, January 2009.

[12] J. van Beusekom, F. Shafait, and T. M. Breuel. Combined orientation and skew detection using geometric text-line modeling. *Int. Jour. on Document Analysis and Recognition*, 2010.

[13] R. van Renesse. Paper based document security-a review. In *European Conf. on Security and Detection*, pages 75–80, London, UK, April 1997.