2011 International Conference on Document Analysis and Recognition

Distortion Measurement for Automatic Document Verification

Joost van Beusekom, Faisal Shafait
Multimedia Analysis and Data Mining Group
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
joost.van_beusekom@dfki.de, faisal.shafait@dfki.de

***Abstract*—Document forgery detection is important as techniques to generate forgeries are becoming widely available and easy to use even for untrained persons. In this work, two types of forgeries are considered: forgeries generated by re-engineering a document and forgeries that are generated using scanning and printing a genuine document. An unsupervised approach is presented to automatically detect forged documents of these types by detecting the geometric distortions introduced during the forgery process. Using the matching quality between all pairs of documents, outlier detection is performed on the summed matching quality to identify the tampered document. Quantitative evaluation is done on two public data sets, reporting a true positive rate from to 0.7 to 1.0.**

***Keywords*-document security, forgery detection, scanning distortions**

I. Introduction

Document forgery is a common problem affecting many areas of everyday commercial activities. One prominent example is that of insurance companies: due to advances in document digitization and information extraction, incoming invoices are digitized and information is automatically extracted. The further processing is solely done on the digital data and thus, even plumply tampered documents may lead to a successful defrauding of money. Automatic tools for identifying tampered documents are thus needed.

There are many different ways to generate a forged document. In this paper we focus on two forgery approaches: the approach of re-engineering a genuine document using e.g. word processing software and the set of approaches involving scanning and printing in some stage, as e.g. scanning, digitally editing and printing of a genuine document. These processes may lead to distortions due to imperfection of the re-engineered document or due to the hardware imperfections in the case of scanners. In general, the distortions introduced by the forgery process are not easily visible with the bare eye. Only in direct comparison with a genuine document with identical text parts, these can be detected.

In the afore-mentioned scenario of invoice processing in insurance companies, many invoices of the same source are being processed. In this paper, we use this fact to automatically detect tampered documents by detecting the distortions. Given a set of documents from the same source, without prior information about which document is genuine and which is not, they are all matched against each other. The sum of the matching scores is used as a feature to detect outliers in a cluster of documents from the same source.

The idea behind this approach is the following: invoices often have common text parts that do not vary over different invoices, as e.g. headers and footers. Scanning distortions and imprecise re-engineering will vary slightly the relative position of these text parts. By matching all documents in a cluster against each other, the sum of the matching qualities for each document can be computed. The forged document will have a lower score due to the scanning distortions. As the matching quality depends on the document contents, it will show normal variations even for genuine documents. A flexible approach for detecting abnormal matching sums is taken by using outlier detection methods.

Section II gives an overview of the advances in the domain of automated document authentication and scanning distortions. Section III describes the overall approach. Evaluation and results are presented in Section IV and V. The paper concludes with Section VI.

II. Related Work

Different forgery detection methods have been presented previously. However, most of them use extrinsic security features that are not present in most of the everyday life's documents. A good overview of different kinds of extrinsic security features can be found in [1]. There are also approaches using intrinsic document features, e.g. information about the printer [2], [3] or the printing technique [4].

Scanning distortions have been analyzed in different domains and publications. Seywald [5] analyzed the geometric scanner accuracy of different scanner types. He showed that single line CCD scanners as well as multiple swath line CCD scanners show considerable amounts of geometric deviations. Square array CCD scanners show lower degrees of distortions. These are, however, rarely used for documents.

The distortive effect of the scanning process compared to the digitally rendered document has also been noticed by Kanungo [6] in an attempt to automatically generate ground truth for Optical Character Recognition (OCR).

However, despite the distortions being known for a long time, to the authors' best knowledge, there is no approach using them to detect tampered documents.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.66

In our previous work [7], we presented the first approach to use the distortions to detect forgeries. The approach models the positional variation of connected components in fixed genuine document parts. First, the document images are aligned using only the fixed parts and clustering connected components that share the same position. The variation of this position is modeled. This model is used to verify incoming documents. The main drawbacks of this method are that the fixed areas of a document type have to be selected manually and that, in a supervised manner, genuine documents are needed to learn the model.

In this paper we present an approach to overcome the drawbacks of our previous work by using the matching score in combination with an outlier detection method.

III. Method Description

Despite the promising results in our previous work [7], the drawbacks of needing a manually generated fixed area map and genuine training data makes a fully automatic application impossible. To overcome this problem, a more general approach is presented: given a cluster of invoices claiming to be from the same invoice source, it is assumed that:

- the documents show constant text parts as e.g. headers
- the number of forged documents in the cluster is much lower than the number of genuine documents
- the forged documents have undergone a scan and print process at some point of their generation or have been re-engineered

In this case, due to the distortions introduced by forging to the geometric appearance of the document, when aligning all documents pair-wise, the forged documents will in average fit worse to genuine ones than genuine documents do. This is used to detect outliers that show low qualities of fit.

The first step of the proposed approach (Section III-A) consists of aligning the invoices pair-wisely. Then the summed matching scores are computed (Section III-B). Outlier detection is performed on the summed matching scores (Section III-C) to detect the documents that have suspiciously low matching scores[1].

A. Document Image Matching

To compute how well the contents of two document images can be aligned to each other, the documents have to be matched and a matching score is computed. This matching score can be seen as the number of characters that have the same position in both document images. A previously published method [8] is used for the matching.

The alignment consists of defining the transformation with parameters (t_x, t_y, s, α) that maximizes the matching score (or the matching quality). The initial parameter space S is given by $[t_{x_{min}}, t_{x_{max}}] \times [t_{y_{min}}, t_{y_{max}}] \times [a_{min}, a_{max}] \times [s_{min}, s_{max}]$. To find the optimal parameters set, an optimal branch-and-bound search algorithm is used [9].

For completeness, the main idea of the algorithm is presented here: at start, the algorithm is initialized with the whole parameter space. Next, the parameter space is divided into two subspaces. An upper bound for the matching quality of these subspaces is computed. Let $B = \{b_1, \ldots, b_N\} \in \mathbb{R}^2$ be the set of image points of the first image and $M = \{m_1, m_M\} \in \mathbb{R}^2$ the set of image points in the second image. For each model point m, a bounding rectangle $G_R(m)$, representing the area where m may be transformed to, can be computed using the transformation space to be searched. If the distance $d = min_{g \in G_R(m), b \in B} ||g - b||$ is less than a threshold ϵ, a potential match is found and the upper bound for the quality of the parameter subspace is incremented. As we are interested only in accurate matches, epsilon is set 0.5 to obtain pixel accuracy.

To reduce the search space beforehand as far as possible, initial deskewing of the documents is done [10]. As the documents are all scanned using the same scanner, also the scale parameter can be set to a fixed value of 1.0. Assuming the distortions introduced by the scanner used for verification as being constant, these can be neglected. Thus, only translation in both directions and small amounts of rotation have to be considered, making the matching of two documents efficient. If different scanners are being used for scanning the questioned documents, one model per scanner has to be computed, except for the case that all scanners introduce exactly the same distortions.

Initially, the center points of connected components were used for matching. In controlled scanning environments, these points are stable enough to allow for accurate matching. However, if only these points are used, any content information of the bounding box is lost, leading to many mis-matches (e.g. in tables, different numbers sharing the same position). To avoid this problem, OCR is used to filter the matches in a way that two points may be matched only if the OCR output for both characters is identical. In this case, only connected components representing characters are considered for matching. This approach ignores logos and images, but manipulating textual regions is the far most common scenario in document forgery applications.

B. Summed Matching Score Feature

The matching score resulting from aligning two document images varies depending on the documents' contents. Deciding on a single comparison if one of the documents has been forged is thus a difficult task. Moreover, even if a method detects that both documents differ, it is not possible to decide which of the two is the suspicious one.

The assumption that forged documents are less common than genuine ones is used to solve this ambiguity: in the scenario of a document processing pipeline, the incoming

[1]Distortions will lead to a lower quality of fit, thus only outliers with lower summed matching scores are of interest.

questioned document is matched against a set of previously processed documents. If the incoming document shows distortions, it will fit less well on the other documents and the single matching scores will be in mean lower than the scores of the genuine documents.

Thus, after pair-wise matching of all documents against each other[2], the sum of all the matching scores for each document is obtained. This sum is expected to be lower for the distorted document images than for the genuine ones.

C. Outlier Detection

Depending on the constraints of the problem, several approaches exist for outlier detection [11]. The proposed scenario of the detection system makes it unlikely to have various different outliers in the feature space. Also, due to the nature of the feature, a normal distribution of the summed matching scores can be expected (see Section V for a more detailed analysis of the normality of the data). Therefore, we used Grubbs' method for outlier detection [12]. The one-way test statistic is defined as:

$$T_1 = \frac{\bar{x} - x_1}{s} \tag{1}$$

where $x_1, \ldots, x_n$ are the observed samples, $\bar{x}$ is the mean of all samples and s the standard deviation of the samples. The threshold for rejecting the hypothesis that the lowest value is no outlier is set depending on the sample size and the significance level.

The Grubbs' outlier detection implementation used for the experimental evaluation is provided in the R outliers package. It converts the outcome of the Grubbs' test into a p-value p_G that is then used for deciding whether to reject the null hypothesis or not. If $p_G < p_T$, where p_T is the significance level, the null hypothesis is rejected and the lowest value is considered being an outlier. Grubbs' test can be used to detect multiple outliers by repeating the test and removing the outlier, until no outlier can be found.

IV. Evaluation

The evaluation was done on two different data sets: the *doctor bill* data set consists of 40 dummy doctor bills, 12 forgeries, and 40 first generation copies of genuine bills. The dummy bills were generated by a student using OpenOffice. The 12 forgeries were generated by other students: they were given a genuine bill and their task was to re-engineer the document using a text editor of their choice as accurate as possible. The 40 copies were generated copying five randomly chosen bills on eight different copiers, reaching from simple fax machines to high-end work group copiers.

The second data set[3] also consists of dummy on-line shop invoices created by an automated invoice generating script. It consists of 110 genuine bills from the same invoice source. Ten of these have been used to generate copies on 12 different copiers again reaching from simple fax machines to high-end work group copy machines. In total 120 first generation copies were obtained.

Each data set was scanned using only one scanner. A resolution of 300 dpi was used and images were saved in 8-bit gray-scale format. Binarization was done using Otsu's global thresholding method [13] and for deskewing our previously published method was used [10]. For extracting OCR information, Cuneiform v1.0 was used.

To follow the above mentioned scenario of a stream of documents to be verified and as the data set is limited in size, an n-fold cross validation approach is followed for both data sets: the set of genuine documents is split into n parts. One part is taken as "training set" (the set of documents that the system already processed). The remaining genuine and forged or copied documents are taken as a test set and for each of these documents the following test is run: first, the document is added to the set of currently analyzed documents. Then for each pair of documents their matching quality is computed and summed up. Finally, Grubbs' outlier detection method is run on the obtained summed qualities. The four possible outcomes can be found in Table I; e.g. if the outlier detection was run on a set containing only genuine test images and if an outlier was detected, this is considered as a false positive.

The normality of the feature for genuine documents was verified by QQ-plots and by the Shapiro-Wilks test.

Then, n-fold cross validation has been run with different values of n to see the influence of the size of data set on which the outlier should be detected. As the Grubbs' outlier test is known not to work well on sample sizes of six or smaller [14], n has been set in a way to avoid such small sample sets. For the doctor bill data set cross-validations with $n = \{2, \ldots, 5\}$ have been run, for the second data set, test runs with $n = \{2, \ldots, 15\}$ were performed. If, due to incorrect removal of a non-outlier, the size of the sample set is less or equal than six, no further outliers are being detected and an extra error count is incremented. In this case, the other four metrics stay unchanged. To evaluate the influence of p_T, three different values for the significance level $p_T \in \{0.05, 0.01, 0.005\}$ have been tested.

As evaluation measures, the true positive and the true negative rates are computed as $r_{tp} = \frac{tp}{tp+fn}$ and $r_{tn} = \frac{tn}{tn+fp}$, where tp, tn, fp, fn are the number of true positives, true negatives, false positives and false negatives respectively.

	forged/copied test image	**genuine test image**
outlier detected	true positive	false positive
no outlier detected	false negative	true negative

Table I
Possible Outcomes of a Test-Run.

[2]The complexity of this is obviously quadratic. However, not all available documents are needed to be in the reference set as can be seen in Section V

[3]Data sets can be downloaded from: https://madm.dfki.de/downloads

The ratio of cases where the remaining set for outlier detection was less or equal to six is also given.

V. Results

The results for the normality tests can be seen in Figure 1. The Shapiro-Wilk test returned p-values of $p = 0.3990$ and $p = 0.8264$ and thus in both cases, the null hypothesis that the data is normally distributed is not rejected. From both results it can be concluded that the summed matching score feature is normally distributed for genuine documents.

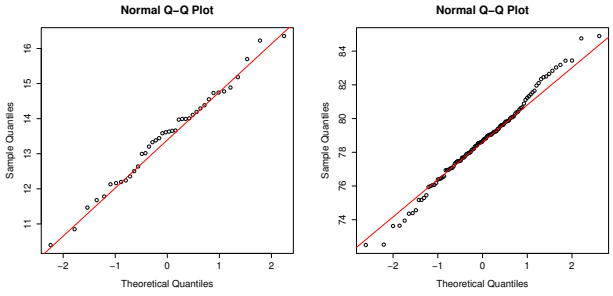
Plots of the true positive and true negative rates for both data sets can be found in Figure 3. It can be seen that for the doctor bills data set the true positive rate varies between 0.7 and 0.2 for the different parameter settings. The true negative rate is nearly zero for values $p_t \leq 0.01$.

For the on-line shop invoices the true positive rate is 1.0 for all tested values of n. Also for this data set it can be observed that with decreasing p_T the true negative rate increases. However, in contrast to the doctor bills data set, this has no influence on the true positive rate.

An analysis of the reason for different performances on both data sets shows in Figure 2 that the feature is more discriminative on the on-line shop invoices data set. The boxplot of the summed matching qualities for the copied documents shows some overlap. For the on-line shop invoices, this overlap cannot be observed. The reason for this is that the position of the footer in the doctor bills data set is relative to the document's content. So only the small header influences the matching quality. An example of a matching result can be found in Figure 4. It can be seen that the copied document matches less well than the genuine one.

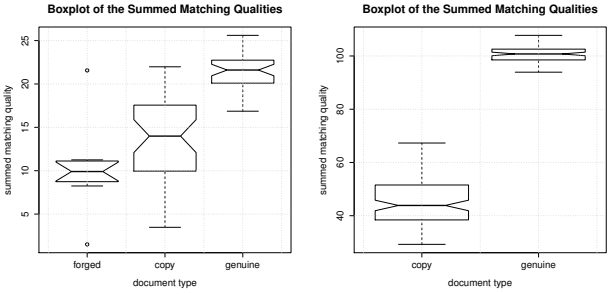
VI. Conclusion

We presented an approach for automatic document forgery detection by detecting distortions that are typical for scanned or re-engineered documents. The detection of the distortions is done on fixed document parts, e.g. headers and footers, that often appear in invoices. By matching the questioned invoices to invoices from the same source, the summed matching quality is computed and used as the feature for subsequent outlier detection. If an outlier is detected, that document is considered as suspicious. Two data sets were generated to evaluate the approach and good results were shown. This permits the usage of the distortions in an unsupervised setup, since no manual intervention is needed.



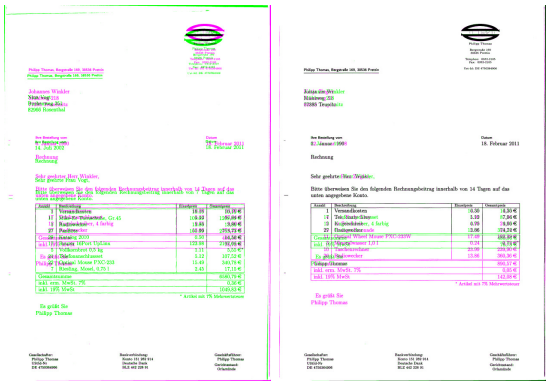
(a) Doctor bills (b) On-line shop invoices

Figure 1. QQ-plots for the summed matching score feature of genuine documents for both data sets.



(a) Doctor bills (b) Online shop invoices

Figure 2. Boxplot of the summed matching qualities for both data sets. For the doctor bills data set the discriminative power of the feature is less compared to copies of the on-line shop invoices.



(a) Copy vs. Genuine (b) Genuine vs. Genuine

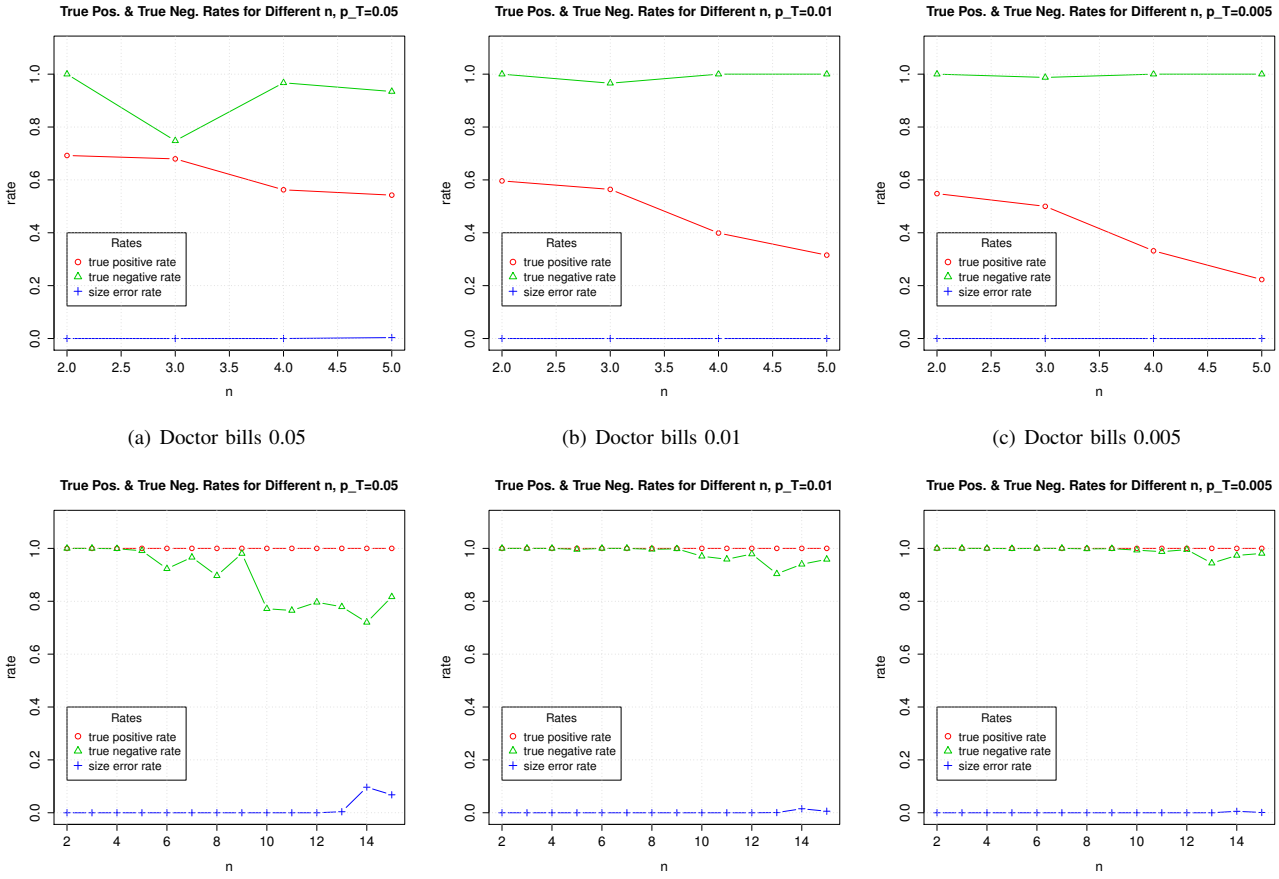
Figure 4. Matching a copied to a genuine document (left) and a genuine to a genuine document (right). One image is drawn in magenta, the other in green, overlapping colored pixels in black. Large parts of the copied document cannot be matched.

References

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

[2] J. van Beusekom, M. Schreyer, and T. M. Breuel, "Automatic counterfeit protection system code classification," in *Proc. of SPIE Media Forensics and Security XII*, San Jose, CA, USA, January 2010.

[3] A. K. Mikkilineni, P.-J. Chiang, G. N. Ali, G. T. C. Chiu, J. P. Allebach, and E. J. Delp, "Printer identification based on graylevel co-occurrence features for security and forensic applications," in *Proc. of SPIE Security, Steganography, and Watermarking of Multimedia Contents VII*, vol. 5681, San Jose, CA, USA, February 2005, pp. 430–440.



(a) Doctor bills 0.05 (b) Doctor bills 0.01 (c) Doctor bills 0.005

(d) Online shop invoices 0.05 (e) Online shop invoices 0.01 (f) Online shop invoices 0.005

Figure 3. True positive and true negative rates for both data sets and different values of p_T. Results on the on-line invoice data set are considerably better than for the doctor bills data set.

[4] C. Schulze, M. Schreyer, A. Stahl, and T. M. Breuel, "Using dct features for printing technique and copy detection," in *Proc. of the 5th Int. Conf. on Digital Forensics*, Orlando, FL, USA, January 2009, pp. 95–106.

[5] R. Seywald, "On the automated assessment of geometric accuracy scanner performance," in *Proc. of the 18th Congress of the Int. Society for Photogrammetry and Remote Sensing 1996*, vol. 31, Vienna, Austria, July 1996, pp. 182–186.

[6] T. Kanungo and R. M. Haralick, "Automatic generation of character groundtruth for scanned documents: A closed-loop approach," in *Proc. of the 13th Int. Conf. on Pattern Recognition*, Vienna, Austria, August 1996, pp. 669–676.

[7] 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*, ser. Lecture Notes in Computer Science, vol. 5158, Washington, DC, USA, August 2008, pp. 47–57.

[8] J. van Beusekom, F. Shafait, and T. M. Breuel, "Automated OCR ground truth generation," in *Proc. of the 8th IAPR Workshop on Document Analysis Systems*, Nara, Japan, September 2008, pp. 111–117.

[9] T. M. Breuel, "A practical, globally optimal algorithm for geometric matching under uncertainty," *Electronic Notes in Theoretical Computer Science*, vol. 46, pp. 1–15, 2001.

[10] 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*, vol. 13, no. 2, pp. 79–92, 2010.

[11] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[12] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, pp. 1–21, 1969.

[13] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[14] V. Chandola, A. Banerjee, and V. Kumar, "Outlier detection: A survey," University of Minnesota, Tech. Rep., 2007.