Lessons Learned from Automatic Forgery Detection in over 100,000 Invoices

Joost van Beusekom, Armin Stahl, and Faisal Shafait(✉)

German Research Center for Artificial Intelligence (DFKI),
Kaiserslautern, Germany
j_v_b@gmx.net, {Armin.Stahl,Faisal.Shafait}@dfki.de

Abstract. Digitization and automatic processing of incoming paper mail is a crucial component of document management systems and is widely adopted in medium and large enterprises. Besides several advantages of this automated processing, it complicates the first line inspection of incoming documents which are often of vital financial or legal relevance. We have developed a number of different techniques to allow automatic detection of forged or manipulated documents over the last years. In this paper, we present an analysis of the application of our methods on a large real-world dataset of invoices to identify the weaknesses of our existing methods and propose some promising directions of future work in the field.

1 Introduction

Even though many traditional paper-based business processes are more and more transferred to pure electronic processing, paper documents still play a vital role in many business scenarios. Since a manipulation or forgery of such documents may lead to significant financial loss or legal problems, an appropriate inspection of incoming documents is often crucial. A typical example is the processing and reimbursement of invoices in insurance companies.

However, the degree of automation in document processing pipelines has dramatically increased in the last decade: data that was previously transferred manually from thousands of documents the day is now being extracted and processed automatically. Due to the automatic processing, observations during manual processing like "that signature looks strange", or "this part was written using a different pen" or even "the layout of this invoice normally looks different" cannot be made anymore. This complicates the identification of questionable documents additionally or even makes it impossible in practice.

To address these issues, some research has been done to pursue the goal of adding automatic authentication methods to replace the formerly manually done first line inspection in document digitization and processing systems. In the context of documents, "authentication" is defined by J. Hails [1] as "*... showing that writing is what it is claimed to be*". In the scenario of incoming invoice processing, this can be further restricted to "*showing that the writing originates*

U. Garain and F. Shafait (Eds.): IWCF 2012 and 2014, LNCS 8915, pp. 130–142, 2015.
DOI: 10.1007/978-3-319-20125-2_12

from where it claims to be from". Thus, if an invoice claims to come from source X, we want to check if this is really the case or not.

Although this will cover already many forgery scenarios, from the definition it becomes clear that not all scenarios are being covered; e.g. if someone uses the hard- and software from source X to generate a forged invoice, or if only small modifications are made directly on the genuine document (like changing a 1 to a 7 to increase the total amount of the invoice).

In this paper, the results and conclusions of an extended test of our previously developed authentication methods on real-world data, and the implication for the ultimate goal of developing an automated system, are presented. This paper does not claim to present new methods, the methods presented here have been published before and are only briefly introduced to better understand the results. The main contribution of this paper is that the authors evaluated authentication techniques on a large real-world data set of invoices. To the author's best knowledge, there is no other work in the domain of document authentication having done an evaluation on such a large and diverse data set. It should be clarified, that these methods are tested independently, no integrated system is yet available, combining the outcomes of the different methods.

The remaining sections of the paper are organized as follows: in Sect. 2 the automatic methods for forgery detection or verification are shortly introduced. In Sect. 3, the evaluation setup is described. The results of the analysis and the conclusions are discussed in Sects. 4 and 5 respectively.

2 Methods

The main idea behind all the presented methods is the following: the incoming documents can be clustered into groups of documents claiming to be from the same source. This can be done, e.g. using optical character recognition (OCR) to read the address of the sender. It is also assumed that many more genuine documents are present in each group than forgeries. The last assumption is that there are some features in a cluster of documents that are similar for genuine documents but that are different for forgeries.

Since the main focus of this paper lies on the results and conclusions of the real-world experiment, the features will be presented here only in a short, high-level manner. Text-line alignment and orientation is presented in Sect. 2.1. Examination of the counterfeit protection system codes from color laser printers and color copiers is presented in Sect. 2.2. Examination of the scanning distortion measurement for detecting possible manipulations is presented in Sect. 2.3. More detailed information about each method can be found in previous publications. References can be found in the concerning sections.

2.1 Text-Line Examination

The idea of this approach is to measure text-line rotation angles and the text-line alignment and to identify text-lines that have an angle that is too far off from

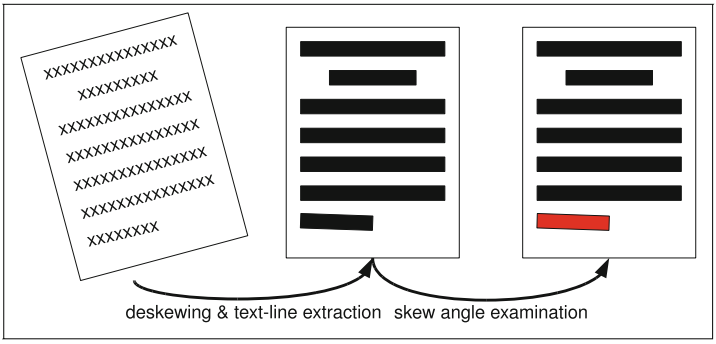

Fig. 1. Visualization of the text-line skew examination: the binarized document is deskewed. The text-lines are examined if their skew angles are abnormally high or not.

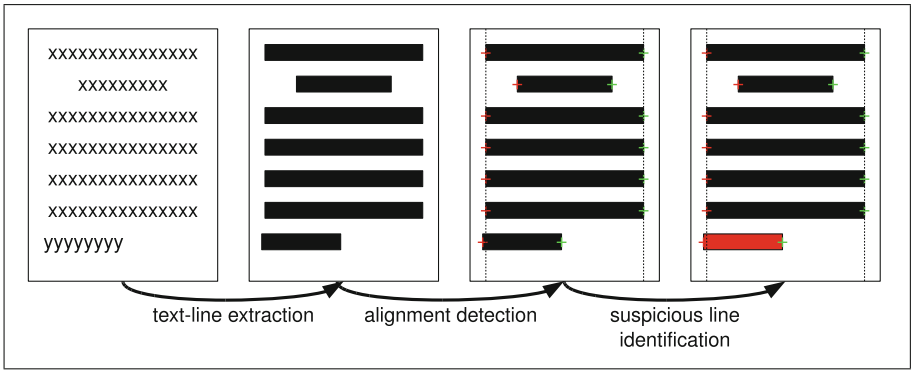

Fig. 2. Visualization of the text-line alignment examination: the text-lines are extracted from the binarized document image. Then the left and right alignment lines are computed. Finally, each text-line is examined whether it shows normal alignment or not.

the normal rotation angle or that show an abnormal alignment. The method is described in detail in [2] and in [3].

An overview of the text-line rotation evaluation method can be found in Fig. 1. First, the document is deskewed [4]. Then, the text-line rotation angles are compared to a previously defined model. If a value differs too much from the model, the corresponding text-line is considered as a suspicious one.

For the alignment of text-lines, a visualization can be found in Fig. 2. After extraction of the text-lines, the left and right vertical alignment lines (margin line) are computed. The distances of the start and end point of the text-lines to the respective alignment lines are used as features to decide if the text-line has a suspicious distance to either of the alignment lines or not.

2.2 Counterfeit Protection System Codes

The image based analysis of counterfeit protection system codes for document authentication has been presented in our previous papers [5] and in [6]. The idea is to use the tiny yellow dots that are generated by many color laser printers to authenticate the document by visually comparing the patterns. This is done by comparing the *prototype patterns* of two documents against each other.

First, the horizontal and vertical size of the prototype pattern are computed by computing the horizontal and vertical pattern separating distance (HPS and VPS distance). Figure 3 gives a rough idea how this step works. Using the so computed width and height, a prototype pattern can be computed from one of the two images whose patterns should be compared. This prototype pattern indicates with what frequency every single dot in the pattern appears in the document. This pattern is then matched to the second image. If a significant difference in frequency of a single dot is detected, the patterns are returned as different. An example of two prototype patterns is given in Fig. 4.

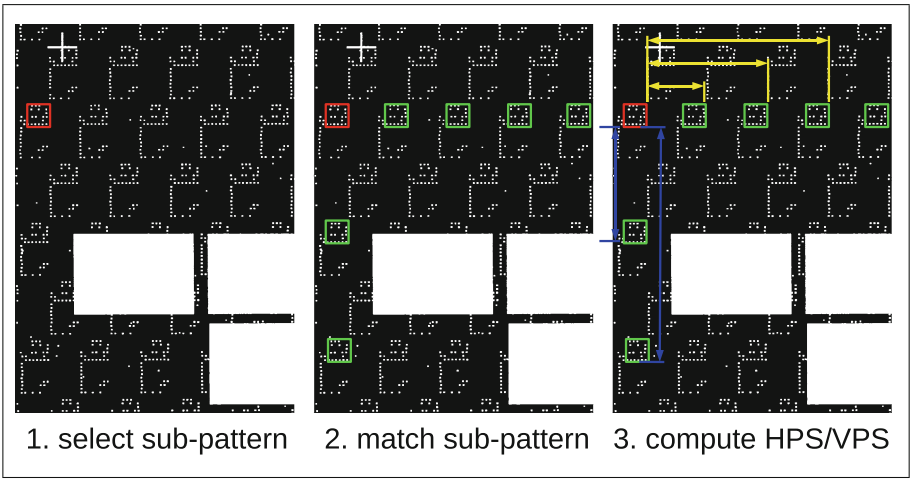

Fig. 3. Computation of HPS and VPS distances: first, a sub pattern is selected. This is matched at different positions in the same column or row respectively. The computed translation parameters in x and y direction are used to extract the HPS and VPS distance of the pattern.

2.3 Distortion Measurement

In this approach, we try to measure the scanning distortions that would be introduced to the forged document through scanning the document (thus, also through the copy process). Details about this method can be found in [7,8].

The main idea of this method is that repeating parts of documents, e.g. headers and footers, should be identical among documents originating from the

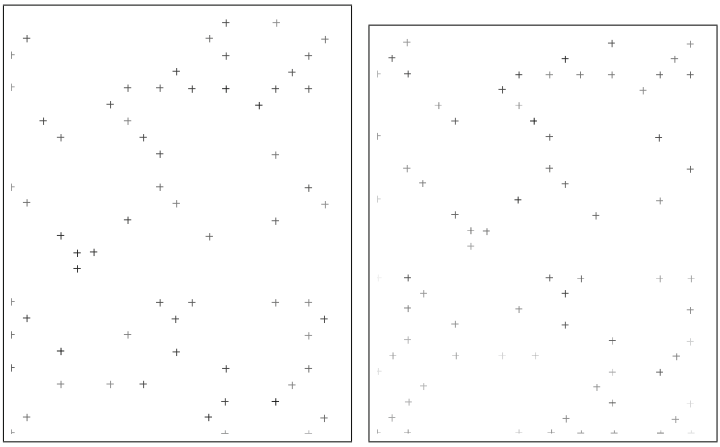

Fig. 4. Examples of extracted prototypes. The darker the cross is the higher its frequency of occurring in a match.

same source and thus, it should be possible to align these pixel-accurately. If this is not the case, the documents are considered as suspicious.

First, all the documents are pair-wisely aligned. From this, a matrix with alignment qualities (also called *matching scores*) is obtained that shows how well each document fits to each other. Distorted documents will fit less well to other documents. This can be detected by computing the sum of the matching scores for each document and running an outlier detection algorithm on the obtained summed scores. Since these values are normally distributed, Grubbs' outlier detection [9] could be used. This also influenced the minimal number of documents that is needed to run the method meaningfully.

3 Evaluation Setup

The data set used for the evaluation contains approx. 143,000 invoice pages. These were scanned using a high-volume automatic document feeding (ADF) scanner using a resolution of 400 dpi and 24 bit color depth. We only considered the first page of every invoice, since in most cases, a forgery attempt will have influence on the first page, even though the main forged part my be found on some other page of the invoice. Unfortunately, the automatic method for detecting the first page did not work at a 100 % accuracy, leading to false alarms by the forgery detection methods. This cases showed up during manual inspection of the automatic results and where recorded as clustering errors.

The data is unlabeled, meaning that we have no information if there are any forgeries or not. A problem with the missing labels is that we are not able to evaluate quantitatively how good or how bad the methods work. Also, no exact information is available about the composition of the data set: from our observations we can conclude the following:

- The vast majority of the documents are machine printed. Only a few consist in majority of handwritten parts. However, signatures are frequent.
- The quality of paper varies greatly: from high-quality watermarked paper to low-quality carbon copies of invoices, nearly every type of paper can be found.
- The quality and type of printing varies considerably: from low-quality ink jet, over middle-quality laser to high quality color laser print-outs. While printing technique itself might be usable as a clue for forgery [10], in some cases multiple printers of different types are used to generate various invoices from one source.

The first pages were clustered according to their source by using OCR and a specialized software. On these clusters, the above methods were executed. Cases where one of the methods threw an alarm were inspected manually. It should be noted here that neither the expert knowledge of a forensic document examiner nor the necessary equipment for doing an in-depth analysis of the documents was available for the manual inspection. However, we know from our customer that forgeries have been detected in the past, mostly by chance, and that these forgeries were of a quite amateurish nature. Also, press releases from other companies having a similar business suggest, that forgeries are to be expected in the invoice dataset that was provided by our customer.

Manual inspection consisted of verification of the cluster's consistency. The digital images and the intermediate debug images of the methods were analyzed. In most cases these could provide an explanation for the detected observation.

In case of doubt, the paper documents were used to verify the results. This was done by one of the authors only, leading to a relative high consistency in the decision making process. Since none of the authors are forensic document experts, we do not claim that from a forensic point of view our analysis meets the requirements of valid forensic analysis. However, since in many cases semantic information was available we were able to reject the hypothesis of forgery in many cases. Also, there were numerous cases where other invoices from the same source, but from different clients were available that showed the same peculiarities. This was also used to reject forgery hypothesis, since we do not assume our customer's clients to cooperate in order to perform fraud.

Only if no reasonable and likely explanation for the observation could be found, the concerned document is labeled as suspicious. This will be finally returned to the client for further investigation on their side.

Due to the missing knowledge about the genuineness or forged nature of the documents, the only outcome of the analysis is either a false positive (an alarm was triggered, although with all available information, a more likely non-fraudulent explanation could be found) or a suspicious document, that might still be a false alarm.

3.1 Test Setup for Text-Line Analysis

For this setup, the text-lines were extracted from the documents and documents that showed large variations of text-line rotation angles compared to a previously learned model were analyzed manually.

3.2 Test Setup for CPS Codes

Two different setups were defined: first, the horizontal and vertical pattern separating distances (HPS and VPS distances) [11] were extracted from each image and a clustering was done based on these distances. Then, the clusters were manually analyzed to see whether the distance clusters actually show CPS codes or if they only represent some noise dots that were mistaken for CPS patterns.

Second, for each document source clusters containing documents with CPS codes, all the CPS codes of the documents were compared pair-wisely against each other. Manual inspection of all the clusters containing differences in CPS codes was done. These differences might be:

- **differing CPS codes:** both documents contain CPS patterns, but they differ in their visual appearance.
- **codes vs. no codes:** one document has CPS codes, the other has not.
- **match:** both documents show the same CPS code.

3.3 Test Setup for the Distortion Measurement

Since Grubb's outlier detection works only reasonably on at least 7 or more values, only the clusters with at least 7 documents were considered in this step. The documents were aligned pair-wisely, the summed matching scores were computed and Grubb's outlier detection was used to detect if there is any outlier or not. If so, the cluster is reported and manual inspection of the results is done. Multiple outliers are detected by removing a detected outlier from the set and rerunning Grubb's method. This process is repeated until either no outlier is detected or the number of remaining samples is equal to 6.

4 Results

Since no ground-truth is available, the results that are presented here base on manual verification and also partially on estimations based on manual inspection.

4.1 Results for the Text-Line Examination

Condensing it to a single sentence: this method in its current form, is not practically useful for forgery detection in real-world invoices. The main reason for this being, that there are too many sources that lead to variations in text-line rotation angle, not to talk of the left and right alignment.

These sources are:

- **pre-printed stationeries** often show variations in the rotation angle of the pre-printed parts in comparison to the actual content part. These are most likely due to the imperfections of the paper paths of the printers that allow the paper to be rotated slightly.

- **unusual layouts and tables** will make the text-line extraction algorithm fail. These unusual layouts show situations that were not expected to be seen in real-world data, as e.g. text-lines with multiple font and font-sizes in one single line. Tables present another problem for the text-line verification approach: the text-line extraction method will find some text-lines where it should not find any and vice-versa.
- **paper cut apart** and pasted together again: while opening the envelopes, some documents are being cut. To allow automatic processing, these are pasted together again by the scanner operator. The rotation angle of the pasted part will diverge from the remaining document part.

On other document types, where long and regular text-lines are more frequent (e.g. contracts and wills), this method might still be useful.

4.2 Results for the CPS Code Verification

The main result of the first evaluation setup was that Tweedy's [11] classification seems still to be up-to-date. No new VPS distances could be found.

The second conclusion is that there were an important number of false detections: CPS points were detected although there were none (e.g. black-and-white print-out) or at least, no regular or repeating pattern could be detected. This problem was solved by adding thresholds on the number of minimally extracted yellow dots: if this number is too small, the extracted dots are considered do be only noise.

The third conclusion is that only an estimated 0.5 % of the invoice documents show CPS dots. This is most likely due to the fact that in this business-related scenario, for financial reasons, most documents are printed in black and white only, thus using either a black and white only printer or using the non-color mode of color laser printers, that in some cases also avoids the appearance of CPS dots on the paper.

The comparison of the CPS codes inside a document cluster lead to some false alarms. In total, 403 clusters containing at least one document with CPS codes were analyzed. These clusters contained a total of 1,181 documents.

In Table 1 the results of the manual verification of the 403 clusters is given.

The meaning of the different result cases are as follows:

- **Correct match** is when all documents in the cluster have the same pattern.
- **Cluster errors** are a frequent source of error for all cluster-based methods, meaning that at least one document in the cluster is from a different source. These are due to the limited resources that were available for tuning the clustering method and manually correcting the results. These errors, however, can be easily removed in a production system, since human operators do correction of the system's clustering results.
- **Method errors** occur due to different problems as e.g. noise, logos or other elements that can eventually lead to yellow dots in the document image. This results in most cases in a imaginary pattern that just consists of noise dots.

Table 1. Results of the verification of the CPS comparison results. In total 403 clusters containing 1,181 files were analyzed

Result	Absolute	Relative [%]
Correct match	156	38.7
Cluster error	127	31.5
Method error	46	11.4
Printer / layout	38	9.4
Suspicious	25	6.2
Copy	6	1.5
Other	5	1.2

In some other cases, a prototype pattern was extracted, but due to noise, it showed differences to the pattern of a clean document image. One reason for these errors are sparse CPS patterns, the reason for these patterns to appear not yet being known.

- **Printer / layout** means that either two or more different printers were consistently used for one document source. In this case, the documents of one cluster could be clustered into different groups with identical patterns. In some cases, the layout of the document was changed by adding some color text or logo, that, most likely, lead to a color print-out instead of a black and white print-out. These clusters thus contained documents both with and without CPS codes.
- **Suspicious** is used when no likely normal explanation could be found on the basis of the data at hand, e.g. when only one single document uses different CPS patterns. These documents are further analyzed by the client. It is reasonable to assume that most of these cases can be reclassified to known cases when background information or other documents are taken into account.
- **Copy** means that one document in the cluster was copied using a color laser copier. Other copy artifacts could be found that vote in favor of this hypothesis, e.g. prints of stamping or staple holes.
- **Other** includes patterns of Xerox printers, where date and time is included into the pattern. This leads to false alarms, since every print-out will show a different time stamp. In some cases the stationeries seem to have been printed using different color laser printers instead of offset print. Also, in a few cases, colored scanner artifacts lead to mis-detection of yellow dots.

4.3 Results for the Distortion Measurement

In total, 2,215 clusters containing 6 or more pages (24,124 pages for all of these clusters) were created. On 88 of these clusters, an outlier was returned by the method. These clusters contained 715 pages. These 88 clusters were verified manually.

An overview over the different, most frequent cases that lead to an outlier being detected can be found in Table 2.

Table 2. Results of the distortion measurement verification. In total 88 clusters containing 715 pages were analyzed

Result	Absolute	Relative [%]
Cluster error	29	32.9
Suspicious	15	17.0
Method error	14	15.9
Layout	8	9.1
Skew	7	7.9
Copy	5	5.7
Document Type	4	4.5
Other	5	1.2

The manual inspection of the results showed, that the main reasons of errors are the following:

- **Cluster errors**, just as for the CPS pattern comparison, are a frequent source of error for all cluster-based methods, meaning that at least one document in the cluster is from a different source.
- **Suspicious** is used when no likely normal explanation could be found on the basis of the data at hand, e.g. if for one document no measurable and visible distortions could be noticed, no sign of a copy could be detected and the document is the only one differing from the source. These documents are further analyzed by the client. It is reasonable to assume that most of these cases can be reclassified to cases when enough background information or further documents are taken into consideration.
- **Method errors** occur due to different problems as e.g. improper matching, noise, improper pre-processing (e.g. problems with binarization) or threshold selection for outlier detection.
- **Layout** stands for varying layouts over time. Document sources tend to change their layout more often than initially expected. This will lead to bad matching, thus to lower matching scores and eventually to an outlier alarm.
- **Skew** was removed before aligning the document images. In some cases the skew of the scanned image was too high. The document could not be unskewed correctly. Thus, the document could not be matched accurately and it would be detected as an outlier.
- **Copy** denoted cases where enough other information could be gathered that increased the likelihood of being a normal copy. Although the data set should not contain copies, the distortion measurement gave some alarms due to copies of most likely genuine documents.

- **Document types** need to be separated for this approach. An invoice should e.g. not be mixed with a formal letter, even if they come from the same source. Although most of the documents were from the same class, some could be found that were of a different class and thus raised a false alarm.
- **Other** includes all other kinds of errors: outliers to the top (e.g. when two exactly identical documents were inside one cluster, e.g. if one invoice was printed twice by the invoice source with exactly the same content); documents that were cut apart and pasted together; stationeries that showed a lot of positional displacement of the main document content and the stationery content and also in rare cases distortions introduced by scanning when the paper was transported in an non-uniform way through the scanner.

5 Conclusions

After processing over 140,000 document pages with the previously mentioned methods and after laborious manual verification of the results, several important conclusions can be drawn concerning the development of automatic document authentication systems.

The main conclusion is that the methods, despite from working reasonably well under laboratory conditions, show weaknesses on real world data. The main problem is that some of the basic assumptions made during development and testing the methods did not hold in their entirety in practice: we assumed that one source uses one printer or type of printer, that layout changes do not occur frequently, scanning distortions are not large enough to cause false alarms when aligning two genuine documents, that paper always remains intact, etc.

The reasons for the assumptions not to hold are not going to be solved in a way to make the authentication methods work error-free. Thus, we think that much more resources should be spent in the logic that comes after the authentication methods: this should bring together all possible information from the analysis methods as well as background information to make a decision whether a document should be finally marked as suspicious or not.

One such an extension would be the **combination of method outcomes**: this would not only be useful in increasing the confidence of forgery detection, but it could also be used in other ways: if e.g. the CPS codes match, other processing steps that could lead to false alarms can be skipped.

Another extension is the allowance of different document **sub-sources** inside one source: the multiple printers and layouts of one company would need to be clustered into different sub-sources. Then the authentication step of an incoming document would be done on the sub-source level.

The most important extension, however, is a **time-relative modeling** of the invoice source: changes in layout, printer hard- and software or stationeries should be visible in most cases if a chronological time-line model of the documents would be available. Then, e.g. a change in layout would not trigger alarm immediately, but only if after a certain amount of time, no layout of the same type is seen by the system. The same procedure could be used for varying printer settings (e.g. black and white print versus color print) and changing printers.

Concerning the research in the area, the following conclusions were drawn: there is missing **understanding of the many factors that influence the document generation**: where do sparse patterns originate from, what influence have software and printer on the visual appearance of the document on the paper, what is the influence of the scanner on the digital image and thus on the results of the image, etc. Some of these questions might have already been discussed in the forensics community, however, they are not explored in the computer science community so far.

Most important for further research is the **availability of real-world data** – genuine as well as forged documents. This would give the possibility to evaluate the methods against real data and not against "what computer scientists think the real data and forgeries could look like". This would also help to get an impression about which forgery methods are being used. However, information security and copyright issues are major obstacles in creating a public data repository for such purpose. As an alternative, researchers should publish their home-brew, synthetic data sets to allow other researchers to compare their methods against and to bring more new ideas into this research area.

Acknowledgment. This project was partially funded by the Rheinland-Palatinate Foundation for Innovation, project AnDruDok (961-38 6261 / 1039).

References

1. Hails, J.: Criminal Evidence. p. 150, Thomson Learning, Boston (2004)
2. van Beusekom, J., Shafait, F., Breuel, T.: Automatic line orientation measurement for questioned document examination. In: Geradts, Z.J.M.H., Franke, K.Y., Veenman, C.J. (eds.) IWCF 2009. LNCS, vol. 5718, pp. 165–173. Springer, Heidelberg (2009)
3. van Beusekom, J., Shafait, F., Breuel, T.M.: Text-line examination for document forgery detection. Int. J. Doc. Anal. Recogn. **16**(2), 189–207 (2013)
4. van Beusekom, J., Shafait, F., Breuel, T.M.: Combined orientation and skew detection using geometric text-line modeling. Int. J. Doc. Anal. Recogn. **13**(2), 79–92 (2010)
5. van Beusekom, J., Schreyer, M., Breuel, T.M.: Automatic counterfeit protection system code classification. In: Proceedings of SPIE Media Forensics and Security XII, San Jose, CA, USA, January 2010
6. van Beusekom, J., Shafait, F., Breuel, T.M.: Automatic authentication of color laser print-outs using machine identification codes. Pattern Anal. Appl. **16**(4), 663–678 (2013)
7. van Beusekom, J., Shafait, F.: Distortion measurement for automatic document verification. In: Proceedings of the 11th International Conference on Document Analysis and Recognition, Beijing, China, September 2011
8. Ahmed, A., Shafait, F.: Forgery detection based on intrinsic document contents. In: Proceedings of the 11th IAPR Workshop on Document Analysis Systems, Tours, France, pp. 252–256. April 2014
9. Grubbs, F.E.: Procedures for detecting outlying observations in samples. Technometrics **11**, 1–21 (1969)

10. Elkasrawi, S., Shafait, F.: Printer identification using supervised learning for document forgery detection. In: Proceedings of the 11th IAPR Workshop on Document Analysis Systems, Tours, France, pp. 146–150. April 2014
11. Tweedy, J.S.: Class characteristics of counterfeit protection system codes of color laser copiers. J. Am. Soc. Questioned Doc. Examiners **4**(2), 53–66 (2001)