2014 11th IAPR International Workshop on Document Analysis Systems

Printer Identification using Supervised Learning for Document Forgery Detection

Sara Elkasrawi
German Research Center
for Artificial Intelligence DFKI GmbH
D-67663 Kaiserslautern, Deutschland
sara.farouk-elkasrawi@dfki.de

Faisal Shafait
School of Computer Science and Software Engineering
The University of Western Australia
6009 Crawley, Perth, Australia
faisal.shafait@uwa.edu.au

***Abstract*—Identifying the source printer of a document is important in forgery detection. The larger the number of documents to be investigated for forgery, the less time-efficient manual examination becomes. Assuming the document in question was scanned, the accuracy of automatic forgery detection depends on the scanning resolution. Low (100-200 dpi) and common (300-400 dpi) resolution scans have less distinctive features than high-end scanner resolution, whereas the former is more widespread in offices. In this paper, we propose a method to automatically identify source printers using common-resolution scans (400 dpi). Our method depends on distinctive noise produced by printers. Independent of the document content or size, each printer produces noise depending on its printing technique, brand and slight differences due to manufacturing imperfections. Experiments were carried out on a set of 400 documents of similar structure printed using 20 different printers. The documents were scanned at 400 dpi using the same scanner. Assuming constant settings of the printer, the overall accuracy of the classification was 76.75%.**

I. Introduction

Despite the integration of computer systems in most offices nowadays, paper-based documents still play an important role in everyday life ranging from direct-value paper (money) transactions to governmental paperwork, trade deals, insurance papers or different reports and receipts. A lot of these documents do not contain secure marks which allows their forgery using low-cost commercial devices like scanners and printers. Companies and agencies where large amounts of paper-based documents (such as receipts and bills) are processed automatically without proper verification face similar problems. Consequently, forged documents in such cases are handled by the system without investigation which results in possible financial losses. Due to financial and practical constraints, embedding security features in low-security documents is rarely adopted. Examination through sophisticated tools or with utilization of high quality scanners is also impractical for companies processing large amounts of paper.

One of the scenarios to identify forged documents is by identifying the source printer. According to Ali et al. [1] and Gebhardt et al. [2] printers produce noise depending on their printing techniques and brands. Aging mechanical parts also affect the quality of printed documents. Filtering noise in the printed documents and finding relevant features that can characterize this noise using low-resolution scans can be used as a key feature for detecting forged documents.

Most of the document forgery detection approaches can be categorized into two main groups. One tackles printer identification in order to verify if the document in question has been printed by the original printer (see [1], [2], [3], [4], [5] and [6]). The other examines the document for irregularities that might have occurred during modification or fraud ([7], [8] and [9]).

The approach by van Beusekom et al. [3] detects embedded yellow point patterns in a document that are manufacturer-specific characteristics of the printer. These patterns however appear only on colored prints. Another method (see [10]) makes use of quasiperiodic banding effects on the printed paper to identify the printer type. This approach is not applicable for text-only documents due to the lack of a wide range of grey-levels. In [1], the author extended this approach by projecting signals from each letter and applying a classifier to classify the printer. The tests however were applied to a limited number of printers (six known printers and one unknown) and the documents contained from 40-100 letters which is approximately ten lines depending on the font. Our approach in comparison to this one is independent of the number of text-lines present in the document.

Further approaches for printer recognition have been presented by Mikkilineni et al. [4] where a graylevel co-occurence matrix is used to obtain texture features. Those are used to classify documents from different printers. This approach depends on scanning at a high resolution of 2400 dpi requiring high quality scanners. In our approach we target low-resolution scans as high quality scanners are less common.

Several methods for printing technique recognition have been implemented by Schulze et al. [5] that examine the quality of the printed characters based on the assumption that different printing methods produce different effects on the printed material. Another method implemented by Schreyer et al. [6] uses discrete cosine transform (DCT) features to recognize whether documents have been printed using an inkjet/laserjet printer or photocopied. Tests were performed using only one source document, printed from each of the examined printers. A similar approach [2] examines the character edges for high variance in the gray-level and classify the documents into either laserjet or inkjet using unsupervised approaches. Their work is based on the assumption that edge roughness is a characteristic of the printer, which is captured by the variance of the grey-levels at the edges of characters.

978-1-4799-3243-6/14 $31.00 © 2014 IEEE
DOI 10.1109/DAS.2014.48

Detecting document irregularities has been tackled by [7] where document text lines are examined for misalignment to detect a fraudulent modification. Similarly font differences and over-similarity of characters within a document have been examined in [9] to detect forgery.

A system for detecting forgery in camera images based on scanner noise analysis has been implemented by Khanna et al. [11]. The system assumes a unique noise pattern for each scanner brand and selects statistical features of imaging sensor pattern noise. The method shows promising results for detecting forgeries made through a combination of different images. These features have been used in our approach for printer identification.

The rest of this paper is organized as follows. In Section II, we present features extracted from printed areas, followed by the Section III where details of the experiment are explained. Then, we present the experiment results in Section IV and finally conclude the evaluation of the experiment in Section V.

II. Printer Specific Feature Extraction

To determine features characteristics of each printer, we first separate the printed area from the non-printed area for a closer examination of the printer-generated noise. Our assumption is that the printers would not leave any ink marks on the blank areas of the document, hence these areas are not relevant for extracting printer specific features. This assumption might not hold for defected printers, but most of the functional printers satisfy this criteria. Furthermore, we assume that all pages have the correct skew and orientation [12] and non-text elements have been removed from the documents using text/image segmentation [13]. The main reason for considering text-only documents is to focus on examining printed areas originating from vector graphics. Printed half-tone regions have additional factors that influence their printed quality, hence we chose to ignore such regions in this work. Text lines in the printed document are thresholded into background and foreground pixels. The original image is then subtracted from the thresholded image. The difference image represents the noise and was used for extracting features to train a Support Vector Machine. In the following subsections each step is explained.

A. Text Line Extraction

To examine printed areas in textual documents, text lines were segmented with the help of Tesseract [14]. A sample of the text line boundaries is presented in Figure 1.

B. Image Filter and Noise Extraction

Image filtering was performed to obtain a clean image from which the original image is subtracted afterwords to obtain noise patterns. Filtering the printed area is done by first calculating the Otsu's threshold and getting the median gray-level for the foreground as well as the median gray-level for the background pixels from the original image, using the Otsu binarized image as a mask. Hence, a clean bi-level image is generated that has the gray-level values of all foreground / background pixels set to the median foreground / background values thus calculated. A sample of a clean bi-level image and its original image is presented in Figure 2a and 2b.

To obtain an image representation of the noise, from which features are extracted, the original image is subtracted from the clean image. A sample of a noise image is presented in Figure 2c.

$$I_{\text{noise}} = \begin{cases} I_{\text{clean}} - I_{\text{original}} & \text{if } I_{\text{clean}} \geq I_{\text{original}}, \\ 255 + I_{\text{clean}} - I_{\text{original}} & \text{otherwise} \end{cases} \quad (1)$$

C. Feature Extraction

Feature selection from the noise image is based on the work of Khanna et al. [11], where the author uses statistical features to describe pattern noise produced by flatbed scanners. The reason for using statistical features is to be independent of image content or size. Pattern noise introduced by scanners is mainly caused by imperfections during the manufacturing process which affect scanner sensors. In flatbed scanners, an image is translated by a linear sensor array along the length of the scanner, resulting in each row of the scanned image being generated by the same sensor. Thus, the average of all rows gives an estimate for the fixed "row-patterns" [15].

To draw similarities between the scanning process and the two considered printing processes (i.e. inkjet and laserjet), an understanding of the printing process is necessary.
Inkjet Printers place very small drops of ink on the paper, whose diameter ranges from 50 to 60 microns. They have a print-head that moves back-and-forth horizontally while dissipating ink onto paper as it moves through the printer.
A *laserjet printer* has a drum that is initially positively charged. When a document is to be printed, a laser beam discharges certain areas on the drum, that correspond to the content being printed (e.g. letters). Afterwards, positively charged ink is placed on the drum, and is drawn only by the discharged areas. The printing paper is charged negatively to attract ink from the drum and then discharged for it not to cling to the drum. Finally the paper is heated up to melt the ink on the paper [16].

Laserjet and inkjet printing processes are similar in that they proceed horizontally. Analogously, scanning also proceeds horizontally. Building on the analogy, features extraction is done as explained below.

Let I_{noise} denote an $M \times N$ (M rows and N columns) noise image. Averages of the image columns and rows, denoted by I^r_{noise} and I^c_{noise} respectively, are calculated as follows:

$$I^r_{\text{noise}}(j) = \frac{1}{M} \sum_{i=1}^{M} I_{\text{noise}}(i,j); 1 <= j <= N \quad (2)$$

$$I^c_{\text{noise}}(i) = \frac{1}{N} \sum_{j=1}^{N} I_{\text{noise}}(i,j); 1 <= i <= M. \quad (3)$$

Note that $I^r_{\text{noise}}(j)$ is N-dimensional and $I^c_{\text{noise}}(i)$ is M-dimensional. The correlation between each row of the noise image and the average of all columns, denoted by $p_{\text{row}}(i) = C(I^r_{\text{noise}}, I_{\text{noise}}(i,.))$, as well as that of each column and average of all rows, denoted by $p_{\text{col}}(j) = C(I^c_{\text{noise}}, I_{\text{noise}}(.,j))$ are calculated.

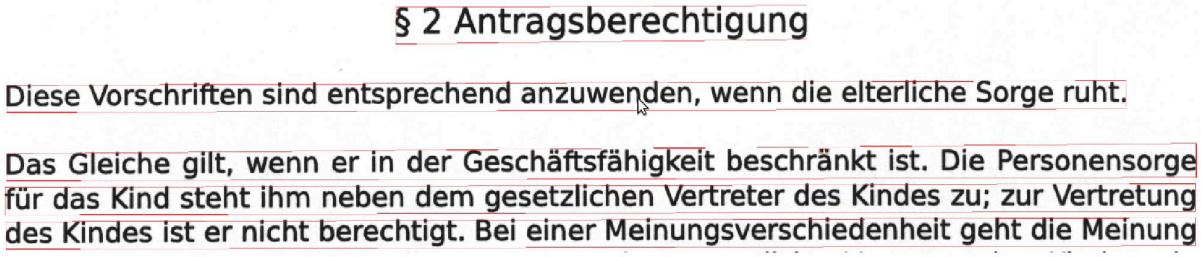

Fig. 1: A sample result of line boundaries

(a) Scanned original document image

(b) Cleaned bi-level document image after replacing all foreground / background pixels in the original image with their median values.

(c) Noise image obtained after subtracting the original image from the cleaned bi-level image.

Fig. 2: Samples from original, filtered (clean) image and noise image

Selected are 15 features from a grayscale image representing printing noise; The mean, standard deviation, skewness and kurtosis of p_{row} and p_{col} are the first 8 features. The standard deviation, skewness and kurtosis of I^r_{noise} and I^c_{noise} represent features 9 to 14. Feature number 15 is a measure of relative periodicity between noise in columns and rows:

$$f_{15} = (1 - \frac{\frac{1}{N}\sum_{j=1}^{N} p_{\text{col}}(j)}{\frac{1}{M}\sum_{i=1}^{M} p_{\text{row}}(i)}) * 100. \quad (4)$$

After extracting those features, an SVM is trained with different parameters to achieve the best possible classification.

D. Example on Feature Extraction

Consider the given excerpt of a sample noise image:

94	222	252	0	2	218	139	22
16	96	152	156	164	100	40	0
12	80	100	106	106	108	96	46
44	172	234	240	244	242	234	164
44	182	252	2	4	6	6	248
44	176	250	2	2	6	8	6
108	228	2	4	4	6	8	8
18	180	254	2	4	4	6	8

From equations 2 and 3, $I^r_{\text{noise}} = [47 \quad 167 \quad 187 \quad 64 \quad 66 \quad 86 \quad 67 \quad 62]$ and $I^c_{\text{noise}} = [118 \quad 90 \quad 81 \quad 196 \quad 93 \quad 61 \quad 46 \quad 59]$. Thus the correlation vector of each row and the mean column noise is $p_{\text{row}} = [0.76541908 \quad 0.40335458 \quad 0.30764977 \quad 0.22812077 \quad 0.63556492 \quad 0.9453724 \quad 0.37110938 \quad 0.96699924]$ and the correlation vector of each column and the mean row noise is $p_{\text{col}} = [-0.03410681 \quad -0.02542542 \quad 0.38378662 \quad 0.73388494 \quad 0.73277127 \quad 0.85778067 \quad 0.91560419 \quad 0.52314518]$.

Calculating the 15 feature for this sample image: $[0.57794877 \quad 0.27238116 \quad 0.40142544 \quad -1.44735627 \quad 0.51093008 \quad 0.35072175 \quad -0.91527649 \quad -0.75345361 \quad 49.6027973 \quad 1.75227986 \quad 0.20677763 \quad 44.40720662 \quad 2.18270397 \quad 1.83316294 \quad 11.59595677]$

III. Experimental Setup

In this work, we used a dataset consisting of 400 document images, involving 20 different printers. Each printer was used to print 20 different pages. Out of 20 different printers used, 13 are laserjet printers[1] and seven are inkjet printers[2].

This dataset is a subset of the dataset developed by Gebhardt et al. [2], containing only the contract document category. A contract consists of horizontal text lines only and

[1](1) laserjet1= 'Samsung CLP 500', (2) laserjet2= 'Ricoh Aficio MPC2550', (3) laserjet3= 'HP LaserJet 4050', (4) laserjet4= 'OKI C5600', (5) laserjet5= 'HP LaserJet 2200dtn', (6) laserjet6= 'Ricoh Afico Mp6001', (7) laserjet7= 'HP Color LaserJet 4650dn', (8) laserjet8= 'Nashuatec DSC 38 Aficio', (9) laserjet9= 'Canon LBP7750 cdb', (10) laserjet10= 'Canon iR C2620', (11) laserjet11= 'Hp Laserjet 4350 o.4250', (12) laserjet12= 'Hp Laserjet 5', (13) laserjet13= 'Epson Aculaser C1100'

[2](1) inkjet1= 'Officejet 5610', (2)inkjet2= 'Epson Stylus Dx 7400', (3) inkjet3= unknown (4) inkjet4= 'Canon MX850', (5)inkjet5= 'Canon MP630', (6)inkjet6= 'Canon MP64D', (7) inkjet7= unknown

of different page length, however most documents contain more than 20 text-lines. The other two document categories in [2] are invoices and scientific papers. As mentioned earlier, we focused in this work on printer noise present in text lines only. Therefore, we ignored the other two categories due to the presence of logos, half-tones, tables, and graphics.

For the first experiment, contracts from six printers were manually selected from the dataset mentioned above (120 documents). The six printers were chosen of different brands to minimize the probability of them having similar mechanical parts. Three of those printers, namely Samsung CLP 500, Ricoh Aficio MPC2550 and HP LaserJet 4050, were laserjet printers whereas Officejet 5610, Epson Stylus Dx 7400 and Canon MX850 were inkjet printers.

The second experiment was performed in three different settings. The first one included documents printed by the laserjet printers only (260 documents), the second one included documents printed by the inkjet printers only (140 documents), and the third setting included the whole dataset (400 documents).

For the evaluation of classification results, the accuracy was chosen as a metric. Accuracy is defined as:

$$\text{accuracy} = \frac{tp + tn}{tp + fp + tn + fn} \tag{5}$$

where true positives (tp) stands for the number of correctly classified samples, false positives (fp) for the number of wrongly classified samples, true negatives (tn) for the number of correctly rejected samples and false negatives (fn) for the number wrongely rejected samples.

For evaluation of the classification per class, precision and recall measures were used:

$$\text{precision} = \frac{tp}{tp + fp} \tag{6}$$

and

$$\text{recall} = \frac{tp}{tp + fn}. \tag{7}$$

The SVM training and testing were done using the RapidMiner[3] SVM package. Features used for the SVM were first normalized prior to training to achieve higher classification accuracy. One SVM was trained for each Experiment set. The kernel function we chose was a polynomial kernel. To select the best parameters for the kernel a grid search was performed.

IV. Results

In the first experiment, the overall accuracy was 90%. Values of precision and recall for the different printers are shown in Figure 3. The figures show no significant distinction between inkjet printers classification and laserjet printer classification. Missclassification occurred between printers of similar technology. For example, most missclassified documents of the Samsung laserjet printer were classified as Ricoh samples, which is a laserjet printer as well.

The second experiment on the whole dataset, resulted in an accuracy of 76.75%. Values of precision and recall for each printer are show in Figure 4. Naturally the accuracy degrades as documents from more printers are included in the classification. The lowest values for precision and recall, namely those of laserjet3 and laserjet10, suggest that laserjet printers might be harder to identify than inkjet printers.

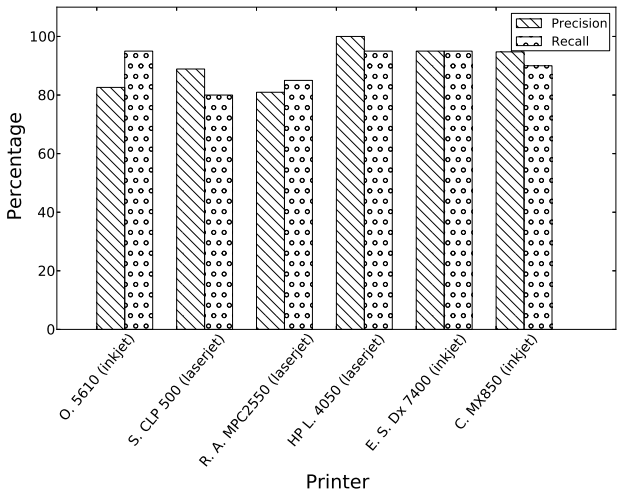

Fig. 3: Precision and Recall for the 6-printers dataset

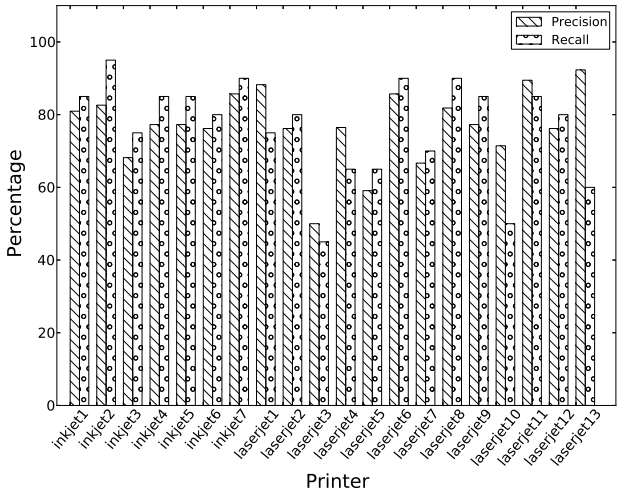

Fig. 4: Precision and Recall for the 21-printers dataset

Comparing the results of classifying the printer only among the laserjet or inkjet printers, we see that inkjet printers are better identified than laserjet with an accuracy of 93.57% (Figure 5). Laserjet-printed documents classification resulted in an accuracy of 78.46% (Figure 6).

Inkjet printers are better distinguishable than laserjet ones due to the fact that inkjet printers, as opposed to laserjet printers, produce more unsharp edges [2]. This allows for more noise to be printed which represents the distinguishing features of the source printer.

[3]http://www.rapidminer.com/

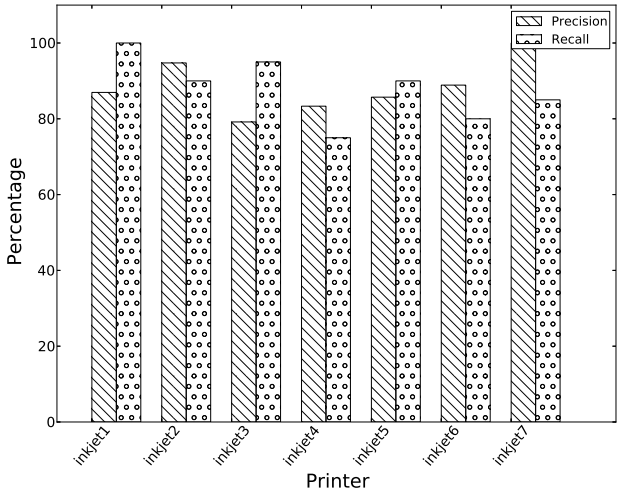

Fig. 5: Precision and Recall for the inkjet printers dataset

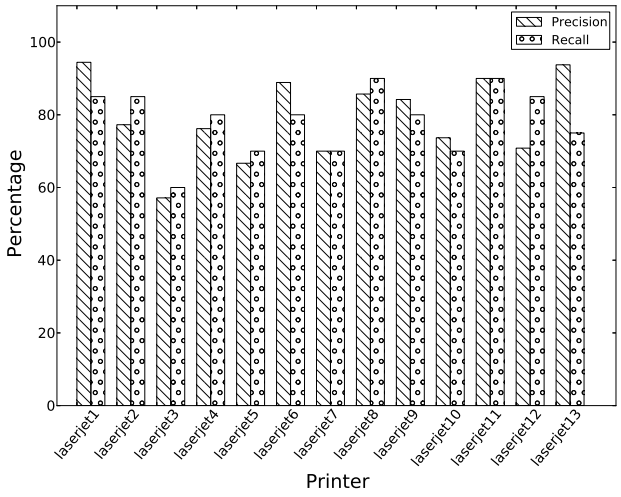

Fig. 6: Precision and Recall for the laserjet printers dataset

V. Conclusion

Using a dataset of documents from printers of different types and brands, our approach has shown promising results in identifying their source printers. The overall classification accuracy for the whole dataset was 76.75%. For inkjet-printed documents, a classification accuracy of 93.57% was achieved. Our approach performs better in identifying source inkjet printers as opposed to laserjet ones, as documents printed by the former type contain more characteristic noise than those of the latter.

Further enhancement could be achieved by finer segmentation of the printed area aiming to improve noise analysis. Experimenting on documents of different formats (e.g. tabular, graphic) would also be useful for testing our approach.

Acknowledgments

This research work was partially funded by The University of Western Australia's FECM research grant.

References

[1] G. N. Ali, A. K. Mikkilineni, P.-J. Chiang, J. P. Allebach, G. T. Chiu, and E. J. Delp, "Application of principal components analysis and gaussian mixture models to printer identification," in *International Conference on Digital Printing Technologies*, vol. 20, 2004, pp. 301–305.

[2] J. Gebhardt, M. Goldstein, F. Shafait, and A. Dengel, "Document authentication using printing technique features and unsupervised anomaly detection," in *Proceedings of the 12th International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 8/2013, pp. 479–483.

[3] J. van Beusekom, F. Shafait, and T. M. Breuel, "Automatic authentication of color laser print-outs using machine identification codes," *Pattern Analysis and Applications*, vol. 16, no. 4, pp. 663–678, 2013.

[4] 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 *Security, Steganography, and Watermarking of Multimedia Contents*, 2005, pp. 430–440.

[5] 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 Int Workshop on Computational Forensics*. Springer, 2008, pp. 35–46.

[6] M. Schreyer, C. Schulze, A. Stahl, and W. Effelsberg, "Intelligent printing technique recognition and photocopy detection for forensic document examination." in *Informatiktage*, 2009, pp. 39–42.

[7] J. van Beusekom, F. Shafait, and T. M. Breuel, "Text-line examination for document forgery detection," *Int Jour on Document Analysis and Recognition*, vol. 16, no. 2, pp. 189–207, 2013.

[8] J. van Beusekom, F. Shafait, and T. Breuel, "Automatic line orientation measurement for questioned document examination," in *Proc Int Workshop on Computational Forensics*. Springer, 2009, pp. 165–173.

[9] R. Bertrand, P. Gomez-Kromer, O. Ramos Terrades, P. Franco, and J. Ogier, "A system based on intrinsic features for fraudulent document detection," in *Proceedings of the 12th International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 8/2013.

[10] G. N. Ali, P. Chiang, A. Mikkilineni, J. P. Allebach, G. T.-C. Chiu, and E. J. Delp, "Intrinsic and extrinsic signatures for information hiding and secure printing with electrophotographic devices," in *Proceedings of the IS&Ts NIP19: International Conference on Digital Printing Technologies*, vol. 19, 2003, pp. 511–515.

[11] N. Khanna, G. T. Chiu, J. P. Allebach, and E. J. Delp, "Scanner identification with extension to forgery detection," in *Electronic Imaging 2008*. International Society for Optics and Photonics, 2008, pp. 68 190G–68 190G.

[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*, vol. 13, no. 2, pp. 79–92, 2010.

[13] S. S. Bukhari, M. I. A. Al-Azawi, F. Shafait, and T. M. Breuel, "Document image segmentation using discriminative learning over connected components," in *Int Workshop on Document Analysis Systems*, 2010, pp. 183–190.

[14] R. Smith, "An overview of the Tesseract OCR engine," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2. IEEE, 2007, pp. 629–633.

[15] N. Khanna, A. K. Mikkilineni, and E. J. Delp, "Scanner identification using feature-based processing and analysis," *Information Forensics and Security, IEEE Transactions on*, vol. 4, no. 1, pp. 123–139, 2009.

[16] M. Schreyer, "Intelligent printing technique recognition and photocopy detection using digital image analysis for forensic document examination," Master's thesis, University of Mannheim, 2008.