Document Signature Using Intrinsic Features for Counterfeit Detection

Joost van Beusekom[1], Faisal Shafait[2], and Thomas Breuel[1,2]

[1] Technical University of Kaiserslautern, Kaiserslautern, Germany,
[2] German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany,
{joost.van-beusekom,faisal.shafait, tmb}@dfki.uni-kl.de
http://www.iupr.org

Abstract. Document security does not only play an important role in specific domains e.g. passports, checks and degrees but also in every day documents e.g. bills and vouchers. Using special high-security features for this class of documents is not feasible due to the cost and the complexity of these methods. We present an approach for detecting falsified documents using a document signature obtained from its intrinsic features: bounding boxes of connected components are used as a signature. Using the model signature learned from a set of original bills, our approach can identify documents whose signature significantly differs from the model signature. Our approach uses globally optimal document alignment to build a model signature that can be used to compute the probability of a new document being an original one. Preliminary evaluation shows that the method is able to reliably detect faked documents.

1 Introduction

Document signatures for paper documents, features on a document to prove its originality, have always been a critical issue, even in ancient times, where the number of paper documents was rather limited compared to the number of documents used today. In these days, where modern technologies enable a broad mass of people to easily counterfeit documents and bills, it becomes more and more important to assure that the document comes from the expected source and that it has not been faked or altered. The signet rings from the monarchs that were used to sign the documents in ancient times have nowadays been replaced with all kind of watermarks: specialized paper, holographic images [1], specialized printing techniques [2] and other physical and chemical signatures [3].

All these methods are in one way or another enhancements of the information medium (often this medium is paper) or the printing process, which results in increased costs compared to the use of standard paper and printing devices. Furthermore, in applications involving a high number of different sources of bills or vouchers (many different invoicing parties), these techniques may be impractical due to the high number of invoicing processes that would need to be adapted.

A typical example of such a use-case involving many different invoicing parties can be found in the tax office: the annual tax declarations are often joined by vouchers from many different invoicing parties. Checking the originality for each voucher is not feasible for tax inspectors. Still it would be important to check for faked bills, as these may be used by tax dodgers to pay less money to the tax office.

In this paper we present a method that allows identifying faked vouchers and bills using a signature obtained from intrinsic features of the documents, namely bounding boxes of connected components. Our method has the advantage, that no extra security features have to be added either to the paper or to the printing process. The main class of forgeries that can be detected by this approach is the imitation of existing bills, which can easily be done by persons capable of using text processing software.

The approach works as follows: observing a number of original bills from one invoice party allows to build a model signature of the non-variable part of a bill. A new bill is then checked against this model signature and if it is significantly different, it is considered as a potentially faked bill. The doubtful bill could then be given to a human operator for further inspection.

The rest of this paper is organized as follows: Section 2 presents our approach in detail and provides an overview of the intrinsic features used in this work. Evaluation and results are shown in Section 3. Section 4 concludes the paper with a short summary and outlook.

2 Description of the approach

As mentioned in the introduction, we focus on identifying fakes of every day documents, e.g. vouchers and bills. The class of falsification methods we are aiming at is the case of home-made pseudo-copies of the bills, which are created by trying to remake the document using a text processing software. Although the faked documents are often at a first glance quite similar, it is very difficult to obtain exactly the same layout conserving the same spacings. Therefore, we want to identify differences in positions of characters in the static part of the bills. The static part of bills and vouchers are the regions of the page that contain always the same information for one invoicing party, e.g. headers, bank account information, and the contact information of the invoicing party.

The layout of a document can be viewed at different abstraction levels, starting from pixels, over to connected components, and finally lines and paragraphs. We choose connected components, more precisely the bounding boxes of connected components as a suitable description of the document images: connected components are well defined, easy to compute and quite stable, as the scanning process can be influenced to deliver reasonable quality for binarized images of the documents.

After a preprocessing step performing binarization [4] and skew correction [5], first we need to construct document signature. This is done in two steps: first, images of all original bills from an invoicing party are aligned with pixel-accuracy

to one reference bill from that party (Section 2.1); second, a signature for this invoicing party is built based on an analysis of variations in positions and sizes of connected components among the aligned bills (Section 2.2). Once a signature is constructed for an invoicing party, the originality of the new bills from that party can be verified by comparing it to the signature (Section 2.3).

We currently focus on the non-variable parts of the documents (e.g. headers, footers, source address and phone number). The distinction between variable and non-variable parts is currently done manually using a bit mask defining which image regions belong to the non-variable part and which do not. The mask can easily be created from one reference document. In future we plan to extract the mask in an automated way by analyzing the bills from one invoicing party using layout analysis methods [6].

2.1 Alignment

The first step in our method is to accurately and robustly align the images. The alignment of two document images aims at identifying the transformation parameters that allow to overlay both images.

Different techniques have been proposed in literature for image registration and alignment. The approaches for general image registration [7] are not well suited for binary document image registration because binary documents lack the color and texture features that are typically used in image registration. Nakai et al. [8] and Liang et al. [9] have proposed image registration techniques for document images, but they handle alignment/registration of the same document under different kinds of distortions. Another way could be to use page frame detection [10] first and then align the page frame of two documents. However, page frame detection is error prone and small errors in the detected page frame will lead to large alignment errors.

In this work, we use the image-matching technique described in [11] for aligning two images from the same invoicing party. This technique is tolerant to changes in the two documents to be matched and hence is a good candidate for use in this scenario. It uses an optimal branch-and-bound search algorithm, called RAST [12] (Recognition by Adaptive Subdivision of Transformation Space). This method allows robust and accurate finding of the globally optimal parameters describing the transformation needed to align both images. Since the RAST algorithm finds the globally optimal alignment of the two images, it is expected that it will align the two images based on their static part.

The quality function used in this case is defined as the number of model points matching an image point under the error bound ϵ.

The RAST algorithm uses a branch-and-bound search for quickly finding a global optimum, which s in our case a maximum for the quality function. The method uses a priority queue containing parameter subspaces in order of their upper bound quality. The highest upper bound quality subspace is divided into two new subspaces, by splitting it into two parts of equal size. For each part, the new upper bound quality is determined and both subspaces are added into the priority queue. These steps are repeated until a stopping criterion is met. In

our case the method stops when the size of the remaining parameter sub space is smaller than a given threshold.

For applying RAST, first an initial parameter space (also called transformation space) has to be defined. Let $[tx_{min}, tx_{max}] \times [ty_{min}, ty_{max}] \times [a_{min}, a_{max}] \times [s_{min}, s_{max}]$ be the initial search space, where tx stand for translation in x direction, ty translation in y direction, a for the rotation angle and s for the scale.

Next, computation of the upper bound quality has to be done. Let $B = \{b_1, \ldots, b_N\} \in R^2$ be the set of image points of the scanned image and $M = \{m_1, m_M\} \in R^2$ the set of image points of the synthetic image, also called "model points" (in order to stick to the original notation of the RAST algorithm). For each model point m, a bounding rectangle $G_R(m)$ can be computed using the transformation space to be searched. This rectangle represents the possible positions where a model point m may be transformed to, using all possible transformations from the current transformation subspace. If the distance d, defined as $d = min_{g \in G_R(m), b \in B} ||g - b||$ is less than a threshold ϵ, the quality of the parameter subspace is incremented. A more detailed description of RAST can be found in [12, 13].

As image points we choose the centers of connected components, as they are relatively stable and easy to compute. In order to speed up the computation of the upper bound for the quality, a filtering step is added before the branch-and-bound search: to avoid comparing bounding boxes that are not similar at all, Fourier descriptors for the contour of the connected components have been extracted [14], describing the shape of the connected component. In order to be invariant to scale and rotation, the images of the connected components are downscaled to a fixed size and the phase is discarded to obtain rotation invariance for the Fourier Descriptors. For each model point (connected component) only the 50 most similar image points are considered for the quality estimation. The value of 50 was chosen manually and showed to work quite well for standard documents. A more detailed description of the filtering step can be found in our previous work [11].

2.2 Building the Model Signature

We follow a probabilistic approach to compute the probability of a document being original. Let ω_o and ω_f denote the two classes of "original" and "faked" documents respectively. Let X be the observed document image consisting of bounding boxes of connected components $x_1, \ldots, x_n$. The posterior for the observed image to be an original one can then be written as:

$$\begin{aligned} p(\omega_o|X) &= \frac{p(X|\omega_o)p(\omega_o)}{p(X)} \\ &= \frac{\Pi_{i=1}^{n} p(x_i|\omega_o)p(\omega_f)}{p(X)} \\ &= \frac{\Pi_{i=1}^{n} p(x_i|\omega_o)p(\omega_f)}{p(X|\omega_o)p(\omega_o) + p(X|\omega_f)p(\omega_f)} \end{aligned} \tag{1}$$

For Equation 1 we assume independence of the observed connected components, which is not always true, but as documents, bills and vouchers have quite diverse types of fonts, font sizes and layouts, the assumption is reasonable.

Problems appear for the other parameters: the prior for having a fake has to be estimated from the dataset. If this not possible, it could be set by the operator to tune the sensitivity of the method. Another problem is the estimation of the $p(X|\omega_f)$. Finding a set of faked documents to train on is quite cumbersome and not feasible in practice.

Therefore, instead of using the Bayesian view, we follow the classical frequentist view of probabilities. We choose to use only $p(X|\omega_o)$. This can be used to determine the probability of a document being original. The value for a document differing to much from the ones from the training set is a strong hint that it may be faked.

The next step is to model $p(x_i|\omega_o)$, the probability of observing a given connected component given the fact that the document is original. A connected component is defined by four parameters x_l, y_l, x_h, y_h defining the lower left and upper right corners of the bounding box of the connected component.

To avoid modelling the probability in the four dimensional space, we do an implicit clustering step allowing to reduce the dimensionality of the resulting histogram: two components are considered being the same, when their normalized overlap is greater than a given threshold T. The normalized overlap is defined by:

$$D_{\text{ov}}(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{2 \times \text{Ov}(\mathbf{x}_i, \mathbf{x}_j)}{\text{area}(\mathbf{x}_i) + \text{area}(\mathbf{x}_j)}$$

where $\text{Ov}(\mathbf{x}_i, \mathbf{x}_j)$ is the number of overlapping pixels of both connected components and $area(c)$ is the number of pixels of connected component x.

We obtain a 2D histogram where the bins represent the positions (the sizes of connected components are included implicitly) and the height the number of connected components of similar size at similar positions.

The procedure to extract this histogram from the training images is as follows: first, all images are aligned using RAST, so that the coordinate systems of all document images have same origin and unit vectors. This is needed to allow comparison between the positions of connected components of the different documents. The set of scanned original documents from the same source is denoted as $O = \{B_1, \dots B_n\}$. One document is taken as reference document, e.g. let B_1 be the reference document. For this document the mask defining what regions should be considered as fix is manually created.

For each document, the set of connected components is computed. Let us denote the set of all connected components of all documents as $X = \{x_1, \dots, x_m\}$. Let M be the bins of the sparse 2D histogram. These bins will be represented by connected components together with number of samples in the bin. We start with an empty histogram. Now for each connected component in X, it is checked if there is a connected component in the model for which the normalized overlapping area is greater than a certain threshold T. Without much parameter tuning $T = 0.8$ showed to work fine. If this is the case, the counter for the number of

samples in the bin is increased. If no such component is found, the given component is added to the histogram as a new bin. The bin sizes are then normalized by the number of all components.

The resulting 2D histogram defines the probability of a connected component of a given size being at a certain position. A simplified visualization of the model can be found in Figure 1.

As the connected components depend on print an scan quality, the question of robustness against merging and breaking connected components arises. As the scanning process can be optimized by the operator of the system, the remaining source of merged and broken components is the invoice generation process. It may happen that the printer of the person creating the bill is low on ink or the paper was changed which could lead to more ink smearing. These problems will result in merged and broken connected components and thus the risk of a false positive will increase. But as these cases should be rare, the cost of sending these to an operator will be reasonably low.

2.3 Checking a New Document

To check if a new document is likely to be an original one or not, the scanned version of the document is aligned to the reference document of the model set B_1. Then, the connected components are extracted. The probability of a connected component to be part of the model is computed using the histogram obtained from the model.

For badly faked bills it may happen, that the alignment will fail. This is no problem as the obtained probability will then be even lower and the faked bill will be reported as falsification.

The probability of a document being original is obtained by:

$$p(X|\omega_o) = \Pi_{i=1}^{n} p(x_i|\omega_o) \tag{2}$$

To decide if this value is likely to be an original document, a threshold value for the probability is defined. This can be set by a human operator. If an automatic setting is needed and if no faked documents are available, the 99% confidence interval of the training set values could be used as a decision rule Under the assumption that the obtained probabilities for the training set are distributed normally, the mean and the variance of the probabilities of the training set can be computed. If the probability of a new bill is less than the mean minus three times the variance, it is classified as a fake.

3 Evaluation and Results

In order to test the performance of our approach, a dataset was needed. As to our best knowledge no public dataset is available containing original and faked documents from one and the same invoice party, we created our own dataset of medical doctor bills as a use case. These sample documents were created by a

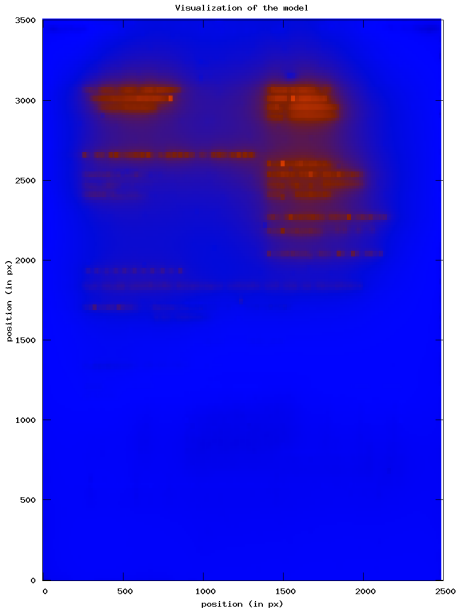


Fig. 1. Visualization of the model. The red regions show regions with stable connected components. The more bluish regions are regions where positions of the connected components are more likely to vary.

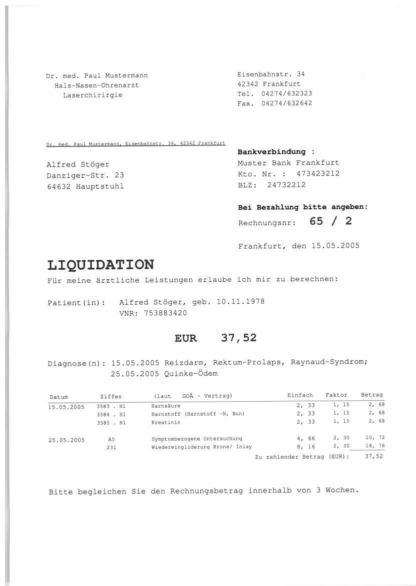
student using Open Office. Next, we picked randomly one document and gave it to other students. Their task was to copy the document as accurate as possible using the text editor of their choice. The number of original documents is 40, the number of faked documents is 12. An example of an original and a faked document can be found in Figure 2.

The set of original bills was split into a test set and a training set of 20 bills each. The model was trained on 20 original documents. The first document was chosen as reference document, where all the other documents were aligned to. Then the model signature was build. For defining the threshold, the 99% confidence interval has been used, computed on the training set.

Then using the model, the probability for the faked bill of being an original for was computed.

The results of the test are shown in Figure 3. It shows the histogram of the probabilities of all the documents, training set and faked test set, for being original. The peak around the right comes from the original documents in the test set. Using the above mentioned threshold, all the faked documents (12 in total) were correctly classified as fakes. 5 out of the 20 original in the test set were wrongly classified as fakes.

A second test has been done in order to measure the performance of the method on a second falsification scenario: instead of remaking the whole docu-



Dr. med. Paul Mustermann
Hals-Nasen-Ohrenarzt
Laserchirirgie

Eisenbahnstr. 34
42342 Frankfurt
Tel. 04274/632323
Fax. 04274/632642

Dr. med. Paul Mustermann, Eisenbahnstr. 34, 42342 Frankfurt

Alfred Stöger
Danziger-Str. 23
64632 Hauptstuhl

Bankverbindung :
Muster Bank Frankfurt
Kto. Nr. : 473423212
BLZ: 24732212

Bei Bezahlung bitte angeben:
Rechnungsnr: **65 / 2**

Frankfurt, den 15.05.2005

LIQUIDATION

Für meine ärztliche Leistungen erlaube ich mir zu berechnen:

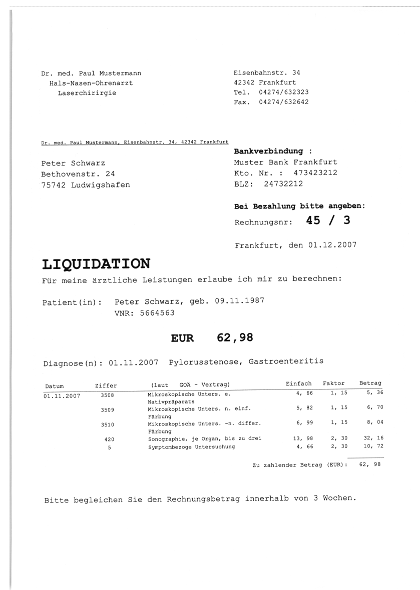
Patient(in): Alfred Stöger, geb. 10.11.1978
VNR: 753883420

EUR 37,52

Diagnose(n): 15.05.2005 Reizdarm, Rektum-Prolaps, Raynaud-Syndrom;
25.05.2005 Quinke-Ödem

Datum	Ziffer	(laut GOÄ - Vertrag)	Einfach	Faktor	Betrag
15.05.2005	3583 . H1	Harnsäure	2, 33	1, 15	2, 68
	3584 . H1	Harnstoff (Harnstoff -N, Bun)	2, 33	1, 15	2, 68
	3585 . H1	Kreatinin	2, 33	1, 15	2, 68
25.05.2005	A5	Symptombezogene Untersuchung	4, 66	2, 30	10, 72
	231	Wiedereingliderung Krone/ Inlay	8, 16	2, 30	18, 76
			Zu zahlender Betrag (EUR):		37,52

Bitte begleichen Sie den Rechnungsbetrag innerhalb von 3 Wochen.



Dr. med. Paul Mustermann
Hals-Nasen-Ohrenarzt
Laserchirirgie

Eisenbahnstr. 34
42342 Frankfurt
Tel. 04274/632323
Fax. 04274/632642

Dr. med. Paul Mustermann, Eisenbahnstr. 34, 42342 Frankfurt

Peter Schwarz
Bethovenstr. 24
75742 Ludwigshafen

Bankverbindung :
Muster Bank Frankfurt
Kto. Nr. : 473423212
BLZ: 24732212

Bei Bezahlung bitte angeben:
Rechnungsnr: **45 / 3**

Frankfurt, den 01.12.2007

LIQUIDATION

Für meine ärztliche Leistungen erlaube ich mir zu berechnen:

Patient(in): Peter Schwarz, geb. 09.11.1987
VNR: 5664563

EUR 62,98

Diagnose(n): 01.11.2007 Pylorusstenose, Gastroenteritis

Datum	Ziffer	(laut GOÄ - Vertrag)	Einfach	Faktor	Betrag
01.11.2007	3508	Mikroskopische Unters. e. Nativpräparats	4, 66	1, 15	5, 36
	3509	Mikroskopische Unters. n. einf. Färbung	5, 82	1, 15	6, 70
	3510	Mikroskopische Unters. -n. differ. Färbung	6, 99	1, 15	8, 04
	420	Sonographie, je Organ, bis zu drei	13, 98	2, 30	32, 16
	5	Symptombezoge Untersuchung	4, 66	2, 30	10, 72
			Zu zahlender Betrag (EUR):		62, 98

Bitte begleichen Sie den Rechnungsbetrag innerhalb von 3 Wochen.

(a) (b)

Fig. 2. The left image is an "original" document from the dataset. The image to the right represents a sample of the faked documents.

ment using a word processor, the forger could just scan an original bill, make some changes using an image editor and print the changed bill. As scanning and printing an original bill distorts the bill slightly, our method should be able to detect these cases. An example of such a distortion can be found in Figure 4.

To simulate this scenario we chose 35 bills to train the model. The remaining 5 original bills were copied on different multi function printers (MFP) to simulate effects of scanning and printing the original bill. In total 8 different MFPs were used to obtain 40 copied bills. The copies were then scanned using the same scanner as for the original bills. The threshold is computed in the same manner as done in the test before. Using our approach, all copies were correctly recognized as fakes. From the 5 original bills only one was wrongly classified as faked. The other 4 were correctly classified as original bills, which is equal to an error rate of 2.2%. On the training set, only 2 false positives were registered.

A plot of the sorted log-likelihoods can be found in Figure 5. It can clearly be seen that most originals have high probability of being original, whereas copied bills lie in between the faked ones and the original ones. The two outliers to the left are due to two copies where the toner was nearly empty.

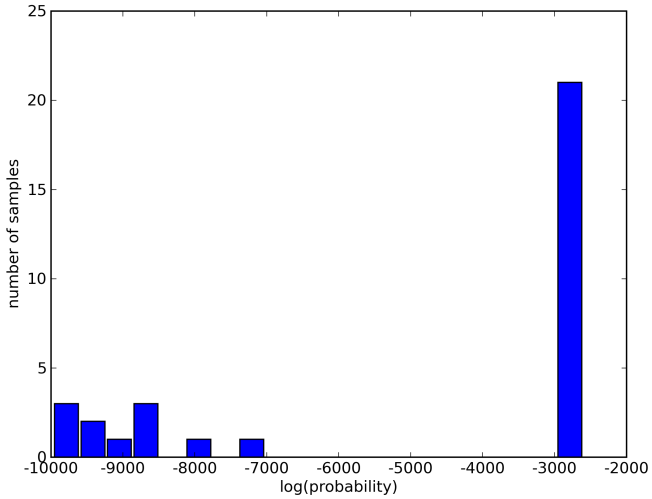


Fig. 3. Histogram of the log-likelihoods. The peak at the right results from the training images. The probabilities of most of the documents are widely different from the ones of the original documents. Only on document had a probability close to the original ones, but still less than all the originals on (around -3000).

4 Conclusion and Future Work

In this paper we presented a novel approach of document falsification detection using intrinsic document features. Using document alignment a connected component based signature could be computed allowing to estimate the probability of a document to be original.

Our approach was tested on a manually created dataset of doctor bills containing a small number of faked documents. The preliminary results proved that the method works reasonably well. A second test on copies of original doctor bills showed that the approach is even able to detect copies with reasonably high accuracy.

One main conclusion of this test is that it is not easily feasible to exactly counterfeit a bill. Although at a first glance copies and fakes look very similar, more detailed analysis shows, that the small variances are unavoidable due to imperfection of the hardware (MFP in our case). The hypothesis, that would need to be investigated in much more detail is that for bills generated using PCs, exactly faking a bill is not feasible unless the same operating system, the same word processing software and the same printer is used.

This approach could be combined with other intrinsic features, e.g. the features used for printing technique classification [15]. This could be incorporated into the model and allow a more accurate modeling of the invoice party, reducing the risk of missing counterfeits.

One important part of future work is to test the approach more thoroughly. If no data sets with appropriate data can be found, these will have to be generated

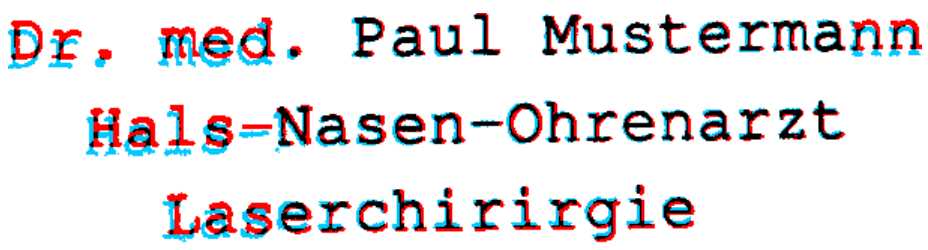


Fig. 4. Example of a distortion induced by copying an original bill. The originally black pixels from the copies bill are painted in blue, the black pixels from the original bill are painted in red. If blue and red pixels overlap, these are painted black. It can be seen that the copying process seems to move blocks up and down: left part the copy is to far down, the middle part fits quite well and the right part is again to far down.

manually. One weakness currently is the missing modelling of the connected component distribution for faked documents. Furthermore, the method needs to be adapted to work on the whole document and not only on the invariant parts. One example could be the detection of incorrect line spacing in the body part of the document, which could result from belated adding of lines.

References

1. Smith, P.J., O'Doherty, P., Luna, C., McCarthy, S.: Commercial anticounterfeit products using machine vision. In van Renesse, R.L., ed.: Optical Security and Counterfeit Deterrence Techniques V. Edited by van Renesse, Rudolf L. Proceedings of the SPIE (2004). Volume 5310 of Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference. (jun 2004) 237–243
2. Amidror, I.: A new print-based security strategy for the protection of valuable documents and products using moire intensity profiles. In: Optical Security and Counterfeit Deterrence Techniques V. Edited by van Renesse, Rudolf L. Proceedings of the SPIE. Volume 4677. (jun 2002) 89–100
3. Hampp, N.A., Neebe, M., Juchem, T., Wolperdinger, M., Geiger, M., Schmuck, A.: Multifunctional optical security features based on bacteriorhodopsin. In: Optical Security and Counterfeit Deterrence Techniques V. Edited by van Renesse, Rudolf L. Proceedings of the SPIE. Volume 5310. (jun 2004) 117–124
4. Bernsen, J.: Dynamic thresholding of gray level images. In: Proc. Int. Conf. on Pattern Recognition, Paris, France (1986) 1251–1255
5. Breuel, T.M.: The OCRopus open source OCR system. In: SPIE Document Recognition and Retrieval XV, San Jose, USA (Jan. 2008) 0F1–0F15
6. Shafait, F., Keysers, D., Breuel, T.M.: Performance evaluation and benchmarking of six page segmentation algorithms. IEEE Trans. on Pattern Analysis and Machine Intelligence **30**(6) (2008) 941–954
7. Zitova, B., Flusser, J.: Image registration methods: a survey. Image and Vision Computing **21**(11) (October 2003) 977–1000
8. Nakai, T., Kise, K., Iwamura, M.: A method of annotation extraction from paper documents using alignment based on local arrangements of feature points. In: Int. Conf. on Document Analysis and Recognition, Curitiba, Brazil (Sep. 2007) 23–27

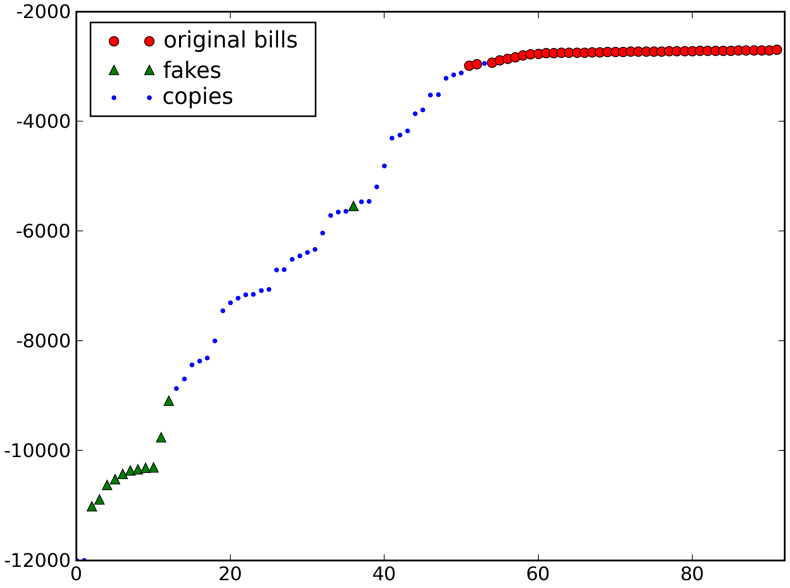


Fig. 5. Plot of the sorted log probabilities (y axis) together with the type of the bill. Circles represent the original bills, triangles the faked bills and dots the copied bills. The x axis represents the plotted bills (92 samples in total)

9. Liang, J., DeMenthon, D., Doermann, D.: Camera-based document image mosaicing. In: Int. Conf. on Patt. Recog., Hong Kong, China (Aug. 2006) 476–479
10. Shafait, F., van Beusekom, J., Keysers, D., Breuel, T.M.: Page frame detection for marginal noise removal from scanned documents. In: SCIA 2007, Image Analysis, Proceedings. Volume 4522 of Lecture Notes in Computer Science., Aalborg, Denmark (June 2007) 651–660
11. van Beusekom, J., Shafait, F., Breuel, T.M.: Image-matching for revision detection in printed historical documents. In: DAGM 2007, Pattern Recognition, 29th DAGM Symposium. Volume 4713 of Lecture Notes in Computer Science. (2007) 507–516
12. Breuel, T.M.: A practical, globally optimal algorithm for geometric matching under uncertainty. Electronic Notes in Theoretical Computer Science **46** (2001) 1–15
13. Breuel, T.M.: Implementation techniques for geometric branch-and-bound matching methods. Computer Vision and Image Understanding **90**(3) (jun 2003) 258–294
14. Granlund, G.H.: Fourier Preprocessing for Hand Print Character Recognition. IEEE Trans. on Computers **C–21**(2) (1972) 195–201
15. Lampert, C.H., Mei, L., Breuel, T.M.: Printing technique classification for document counterfeit detection. In: Computational Intelligence and Security (CIS) 2006, Ghuangzhou, China. Volume 1. (nov 2006) 639–644