Automatic Verification of Properly Signed Multi-page Document Images

Marçal Rusiñol, Dimosthenis Karatzas and Josep Lladós

Computer Vision Center, Dept. Ciències de la Computació
Edifici O, Univ. Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain
{marcal,dimos,josep}@cvc.uab.es

Abstract. In this paper we present an industrial application for the automatic screening of incoming multi-page documents in a banking workflow aimed at determining whether these documents are properly signed or not. The proposed method is divided in three main steps. First individual pages are classified in order to identify the pages that should contain a signature. In a second step, we segment within those key pages the location where the signatures should appear. The last step checks whether the signatures are present or not. Our method is tested in a real large-scale environment and we report the results when checking two different types of real multi-page contracts, having in total more than 14,500 pages.

1 Introduction

Nowadays, big bank corporations tend to release their local branches from all the paperwork load by providing efficient ways to digitize and forward the paper documents to their central services for processing. However, bank customers contract daily tens of thousands of financial products such as loans, mortgages, insurances, investments, etc. yielding to huge volumes of document images to be processed. This document processing usually requires manual intervention in tasks such as document classification, information extraction, verification, etc. Document Image Analysis research provides solutions for automating some of these processes with minimal human intervention.

One of the tasks requiring a tedious manual intervention is the verification of properly signed contracts. Before providing a certain service to a customer, the institution has to confirm that the contract has been properly signed. Since the dawn of Document Image Analysis research, many works dealing with signature verification [1–3] have been proposed. However, before verifying that the signature from a given customer is genuine, there should be a step verifying that the document is properly signed, i.e. it contains the correct amount of signatures in the right places.

The method proposed by Zhu et al. in [4, 5], is one of the few contributions in signature localization in document images. The authors propose a detection framework based on analyzing the curvature of contour fragments over multiple

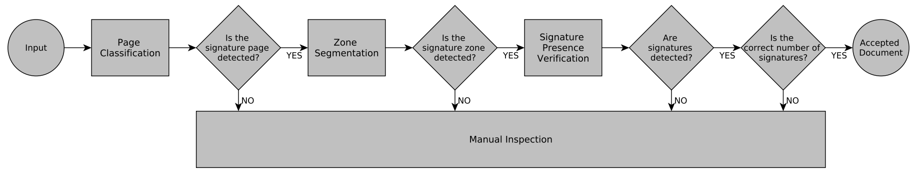


Fig. 1. Overview of the proposed architecture.

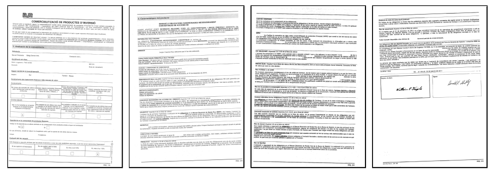
scales. The proposed method performs well at both the localization and matching tasks. Other methods working on signature segmentation are [6, 7]. However, in most scenarios, the incoming documents are usually structured and there is no need to look at signatures over the whole of the document image, instead, focusing at a particular location would be enough, which recasts the problem into a sequential process of segmentation and detection each much simpler than the holistic approach.

In addition, although many works dealing with single-page document representations can be found in the literature (e.g. [8, 9]), works dealing with variable-length multi-page documents, representing a most realistic scenario, are scarce. To our best knowledge, the only works handling a multi-page scenario are devoted to document classification. Examples are the works by Frasconi et al. [10] where a hidden Markov model categorizes documents by looking at sequences of pages. Gordo and Perronnin proposed in [11] a bag-of-pages approach that treats multi-page documents as unordered sets of pages. Finally, Rusiñol et al. proposed in [12, 13] a multi-page document classification system that takes into account both textual and visual cues to categorize incoming documents.

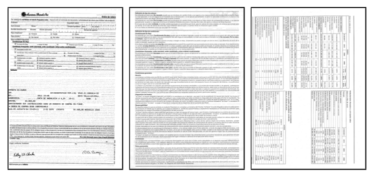
We present in this paper a method that allows the automatic verification of properly signed multi-page structured documents. Our method is tested in a real large-scale environment (more than 14,500 pages have been tested) and we report the results when checking two different types of real multi-page contracts. The remainder of the paper is organized as follows. In Section 2 we overview the proposed method and detail the use cases we focus in. Section 3 gives the details on the page classification step and in Section 4 we present the signature detection strategy. We present the experimental results in Section 5 and we conclude in Section 6.

2 Problem definition and Method Overview

In our particular scenario, the incoming documents to check are multi-page documents. Although we do have knowledge of the document flow structure, i.e. we know when a new document starts and when it ends, the pages can come in a different order and orientation depending on how the operator fed the physical document to the scanner. For a particular document type, we know which pages have to be signed and where on these pages the customer or the bank representative has to sign the document. Given a multi-page document image as input,



Document class A



Document class B

Fig. 2. Example of the multi-page document images considered as use cases.

we want to accept the document if the correct amount of signatures is found in the proper pages at the correct places. Whenever the obtained evidences are too weak, the whole document should be rejected and forwarded to the manual process.

We can see in Fig. 1 an overview of the proposed architecture. We divide our proposed industrial application in three separate stages. First, a page classification step identifies from a page sequence, corresponding to a multi-page document, which are the pages that should be signed. In a second step, we focus on the detected pages and we perform a segmentation step aimed at localizing the particular region on those pages where a signature has to be found. The last step decides whether there is a signature in the document or the signature zone was left blank.

In our scenario it is quite important that the system does not provide false positives. Accepting a non-signed document has a much higher cost than rejecting and sending to manual inspection a signed one. Thus, each of the above steps have associated rejection criteria. In case of doubt, we prefer to forward the document to manual inspection rather than accepting non-properly signed contracts. Hence rejection is configured to work in a conservative fashion.

In this paper we report the obtained results when dealing with two different types of contracts that arrive daily at the central services. We can see an example of those documents in Fig. 2. Since those documents contain private information, we have retouched sensitive parts for presentation here, preserving the overall look and structure of the original documents.

3 Page Classification

Within the document image analysis literature, many descriptors encoding the visual appearance of document images have been proposed. In this work we have used a simple description of document images presented by Héroux et al.

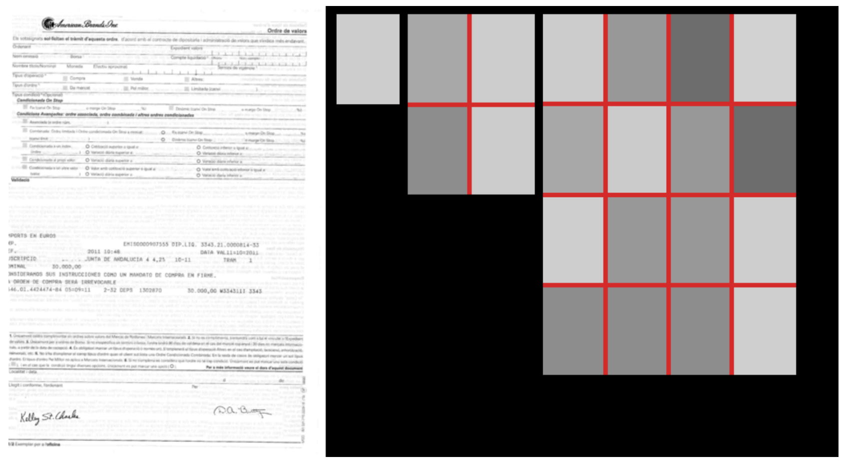


Fig. 3. Example of the multi-scale density descriptor.

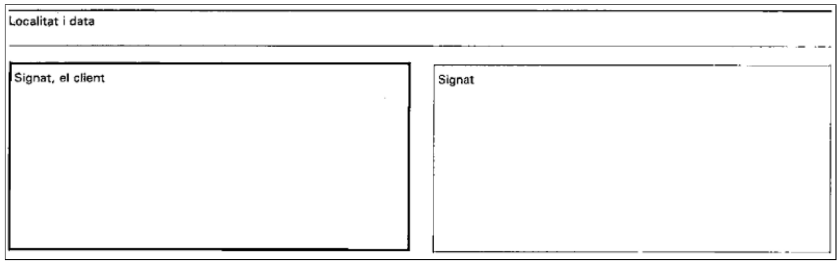
in [14], that encodes pixel densities at different scales. In order to remove small details and noise from the incoming images, a Gaussian smoothing operator is used to blur the images before computing the visual descriptor. Then, the multi-scale descriptor recursively splits each document image into rectangular regions forming a pyramid. In each of the regions the pixel density is computed and stored in the corresponding position of the feature vector. We can see an example of the first levels of the pyramid in Fig. 3. In our experimental setup, we use four scale levels, yielding to an 85-dimensional page descriptor.

Page feature vectors are then $L2$-normalized. The similarity between two pages is assessed by the cosine distance computed using the dot product between both feature vectors. We then use the k-NN classification method over a set of labeled pages in order to decide whether the given page is a page that is expected to bear a signature or not and to which document type the incoming image belongs to. A threshold for the most frequent class in the neighborhood is set as a rejection criteria in order to filter cases where the evidence obtained is weak.

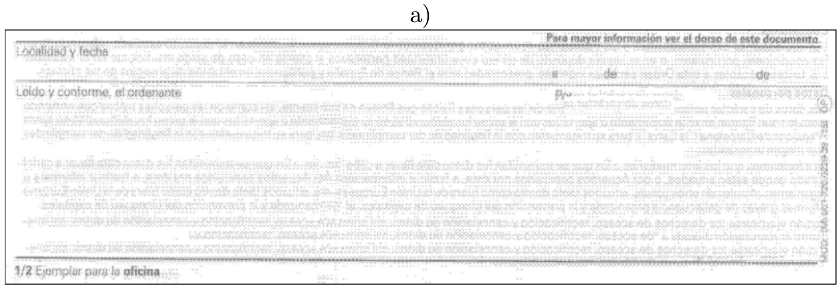
Given a multi-page document, the system either returns a single page where the signature should be found, a negative answer when none of the pages in the document are similar to a signature page or has low confidence on the decision and rejects the document. In the first case, these candidate pages continue the process whereas the rejected or negative documents are forwarded to manual inspection.

4 Signature Detection

Given that the pages that should be signed from the document flow have been identified in the previous step, we aim the following step at detecting the signatures. First we segment the zone of the image in which the signature should



a)



b)

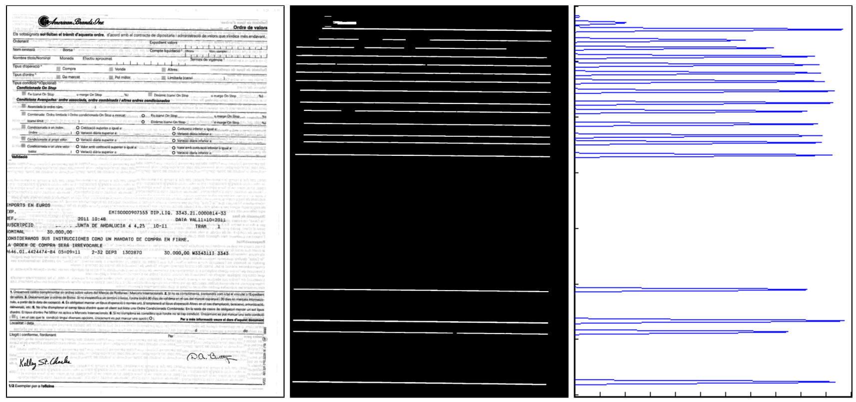
Fig. 4. Example of the zones of interest where the user should sign for the two document types A and B respectively.

be located and then we check whether the zones or interest actually contain a signature or not.

4.1 Zone Segmentation

In all the contracts, the zone within the multi-page document where the customer and the bank clerk have to sign is delimited by some layout structure, e.g. by a box, bold lines, white spaces, etc. In addition, it is usually the case that the zone is also defined by some text indicating where the customers have to sign. Our segmentation framework takes advantage of such graphical (box and line detection) and textual (patterns) layout characteristics, that can be configured for as many different types the system might receive. In the two different types of contracts we deal in our use case we found that these zones of interest are either delimited by a framing box or a bold straight line (Fig. 4).

First, a set of preprocessing steps aim at enhancing the image quality and get rid of the text. A Gaussian smoothing filter is first applied to the document to get rid of punctual noise introduced during the scanning process. The document images are then binarized by applying the Otsu method after a contrast enhancement aimed at stretch the intensity histogram within the 0-255 range. Afterwards, a connected component analysis aims at pruning small objects corresponding to textual characters. A final run-length smearing algorithm is applied to obtain an image where just long straight horizontal and vertical lines are maintained. In the processed image, having only line elements, we per-



a) b) c)

Fig. 5. Original image and graphical part extraction. a) Original image, b) filtered image, c) horizontal projection profiles.

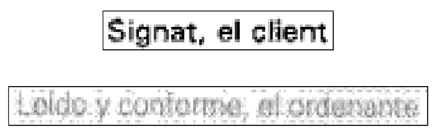


Fig. 6. Example of textual patterns used as rejection criteria for the zone segmentation.

form an horizontal and vertical projection profile analysis [15] in order to locate the zone of the document corresponding to the signature zone layout. We can see an example of the result of such steps in Fig. 5.

Rotated pages are handled as well by looking at all the possible configurations of the sought graphical configurations. Since the localization of the zone of interest is based on projection profiles, the method tends to fail if the pages present severe skew deformations, but tolerates well slightly skewed images.

Finally, in order to ensure that the detected zone is really the location where the signatures should appear, we check whether within the segmented zones we can find certain textual patterns. In all document types, some standard text such as "Customer's signature" or "Read and agree" appears in the zone of interest (see Fig. 6), its existence in the candidate zone can therefore be used as a supporting evidence. A pattern matching implemented through a normalized cross correlation [16] is used as rejection criterion. If in the candidate zone we do not have enough confidence to find those standard texts, we reject the whole image to be processed manually. The locations of those textual patterns are used as well as anchors to refine the zone segmentation.

4.2 Signature Presence Verification

Finally, the last step of the system is to verify whether in the zones of interest there is actually a signature or not. Again some preprocessing steps devoted to reduce the noise are applied to these zones. From the original image we enhance the contrast by stretching the intensity histogram and then threshold the region by applying the Otsu algorithm. Small connected components are pruned in order to get rid of noise provoked by the bleed-through effect. Then, the zones' classification is done by looking at the following features of the remaining connected components.

- **Area**: number of pixels of the connected component.
- **Aspect ratio**: ratio between the height and the width of the connected component.
- **Eccentricity**: ratio of the distance between the foci of the ellipse having the same second-moments as the connected component and its major axis length.
- **Stroke's width**: computed by means of the distance transform.

Here, some experimentally set thresholds established through validation over a training set of 300 pages determine whether a zone contains a signature or not, or rejects the document if we are not confident on the decision.

Given a candidate page, the system either returns a verdict on the signature verification or rejects the page and thus the whole document is forwarded to manual inspection. In our use case, the contracts should contain two signatures each, so the system's answer can be either that both signatures are present, that one of the two is missing or that both are missing.

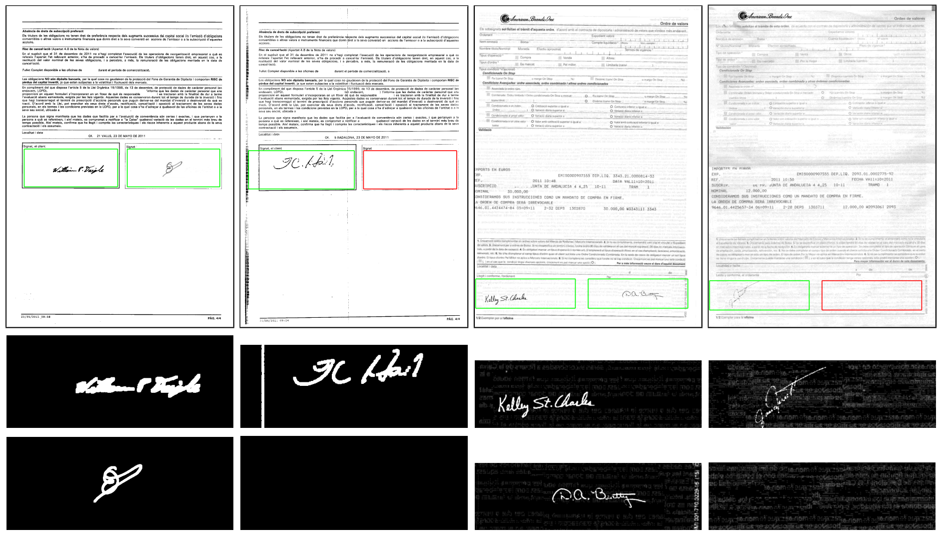
5 Experiments

Table 1. Page classification for documents type A

		System outcome		
		Signature page	Not signature page	Total
True state	Signature page	820	35	855
	Not signature page	16	97	113
	Total	836	132	968

Table 2. Page classification for documents type B

		System outcome		
		Signature page	Not signature page	Total
True state	Signature page	2118	51	2169
	Not signature page	41	130	171
	Total	2159	181	2340



a) b) c) d)

Fig. 7. Qualitative signature verification results. a) and b) properly signed and missing signature documents of type A respectively, c) and d) properly signed and missing signature documents of type B respectively.

Table 3. Signature verification for documents type A

		System outcome				
		Both signatures ok	Only first signature	Only second signature	Neither signature ok	Total
True state	Both signatures ok	698	21	17	2	738
	Only first signature	2	66	0	1	69
	Only second signature	0	0	5	0	5
	Neither signature ok	0	0	0	8	8
	Not signature page	3	9	2	2	16
	Total	703	96	24	13	836

Table 4. Signature verification for documents type B

		System outcome				
		Both signatures ok	Only first signature	Only second signature	Neither signature ok	Total
True state	Both signatures ok	1541	13	22	11	1587
	Only first signature	2	85	0	6	93
	Only second signature	10	1	187	2	200
	Neither signature ok	3	8	3	224	238
	Not signature page	19	6	9	7	41
	Total	1575	113	221	250	2159

Our test dataset consists of 3300 multi-page document images sampled from a real banking workflow consisting of around 14,500 pages. The dataset contains two different types of contracts denoted here as class A and B. The two documents arrive randomly in the document flow. We can see an example of those documents in Fig. 2. We can see in Fig. 7 a qualitative result of the proposed signature verification process. Zones where a signature has been detected are framed in green whereas zones that should contain a signature but it is missing are framed in red.

Concerning the classification step, we can see some results in Tables 1 and 2 for documents of class A and B respectively. The documents labeled as "Not signature page" are forwarded to manual inspection in order to determine whether the document did not contain a signature page or if the system was unable to recognize it. On the other hand, all the document pages labeled as probable signature containers continue the process of zone segmentation and signature presence verification. In that case the proposed system presents an 1.91% and 1.89% of false positives (pages that should not be signed but are labeled as if they should) for class A and B respectively.

After the classification step, the documents with non signature pages have been rejected and only 836 and 2159 pages from classes A and B respectively continue the pipeline. Tables 3 and 4 present the signature verification confusion matrices for documents of class A and B respectively. All the contracts used in the test should contain both the customer and the bank clerk signatures. A document is valid when it contains both signatures and invalid otherwise. In order to reach an accuracy of 92.94% and 94.34% the system rejects 20.26% of documents of class A and 21.37% of documents of class B respectively. However, not all the errors produced by the system have the same effect. Negative answers, that is pages missing one or more signatures, are forwarded to manual inspection, so false negatives (i.e. incorrectly reporting a document as invalid) are not really critical since the human observer will report them as valid ones. Our system delivers a 4.78% and a 2.13% of false negatives per class respectively. Misclassifying an invalid document (e.g. saying that it misses one signature when it misses both) is a recoverable error as well since it is still an invalid image that goes through manual processing. This is the case of a 1.67% and 1.94% of the documents respectively. On the other hand, a false positive (i.e. accepting a document as properly signed when it really lacks one of the signatures) is a critical mistake as it will go through undetected. In that case, our system just yields 0.59% and 1.57% of the images that are incorrectly accepted by the system.

6 Conclusions

In this paper we have presented an industrial application for the automatic screening of incoming multi-page documents in a banking workflow aimed at determining whether the documents were properly signed or not. We have tested our method in a real large-scale environment, handling more than 14,500 pages. The reported results were obtained using two different types of real multi-page contracts. The proposed system is able to automatically verify the signature presence while merely accepting an 1% of critical false positives.

References

1. Plamondon, R., Lorette, G.: Automatic signature verification and writer identification – the state of the art. Pattern Recognition **22** (1989) 107–131
2. Leclerc, F., Plamondon, R.: Automatic signature verification: the state of the art 1989–1993. International Journal of Pattern Recognition and Artificial Intelligence **8** (1994) 643–659
3. Hou, W., Ye, X., Wang, K.: A survey of off-line signature verification. In: Proceedings of the International Conference on Intelligent Mechatronics and Automation. (2004) 536–541
4. Zhu, G., Zheng, Y., Doermann, D.: Signature-based document image retrieval. In: Proceedings of the Tenth European Conference on Computer Vision. (2008) 752–765
5. Zhu, G., Zheng, Y., Doermann, D., Jaeger, S.: Signature detection and matching for document image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **31** (2009) 2015–2031
6. Ahmed, S., Malik, M., Liwicki, M., Dengel, A.: Signature segmentation from document images. In: Proceedings of the. (2012) 425–429
7. Mandal, R., Roy, P., Pal, U.: Signature segmentation from machine printed documents using contextual information. International Journal of Pattern Recognition and Artificial Intelligence **26** (2012)
8. Doermann, D.: The indexing and retrieval of document images: A survey. Computer Vision Image Understanding **70** (1998) 287–298
9. Chen, N., Blostein, D.: A survey of document image classification: problem statement, classifier architecture and performance evaluation. International Journal on Document Analysis and Recognition **10** (2006) 1–16
10. Frasconi, P., Soda, G., Vullo, A.: Hidden Markov models for text categorization in multi-page documents. Journal of Intelligent Information Systems **18** (2002) 195–217
11. Gordo, A., Perronnin, F.: A bag-of-pages approach to unordered multi-page document classification. In: Proceedings of the International Conference on Pattern Recognition. (2010) 1920–1923
12. Rusiñol, M., Karatzas, D., Bagdanov, A., Lladós, J.: Multipage document retrieval by textual and visual representations. In: Proceedings of the International Conference on Pattern Recognition. (2012)
13. Rusiñol, M., Frinken, V., Karatzas, D., Bagdanov, A., Lladós, J.: Multimodal page classification in administrative document image streams. International Journal on Document Analysis and Recognition **17** (2014) 331–341
14. Héroux, P., Diana, S., Ribert, A., Trupin, E.: Classification method study for automatic form class identification. In: Proceedings of the Fourteenth International Conference on Pattern Recognition. (1998) 926–928
15. Likforman-Sulem, L., Zahour, A., Taconet, B.: Text line segmentation of historical documents: A survey. International Journal on Document Analysis and Recognition **9** (2007) 123–138
16. Lewis, J.: Fast normalized cross-correlation. In: Vision Interface. Volume 10. (1995) 120–123