Receipt Dataset for Fraud Detection

Chloé Artaud, Antoine Doucet, Jean-Marc Ogier
L3i, University of La Rochelle
Avenue Michel Crépeau, 17042 La Rochelle, France
Email: {Chloe.Artaud, Antoine.Doucet, Jean-Marc.Ogier}@univ-lr.fr

Vincent Poulain d'Andecy
Yooz
Parc d'Andron - Le Sequoia, 30470 Aimargues, France
Email: Vincent.PoulaindAndecy@yooz.fr

***Abstract*—The aim of this paper is to introduce a new dataset initially created to work on fraud detection in documents. This dataset is composed of 1969 images of receipts and the associated OCR result for each. The article details the dataset and its interest for the document analysis community. We indeed share this dataset with the community as a benchmark for the evaluation of fraud detection approaches.**

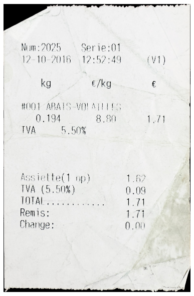
I. Introduction

Recent research in document forensics are mostly focused on the analysis of images of documents. One of the frequent tasks consists in retracing the course of document images [1], identifying the use of different printers or scanners for a single document to detect inconsistencies [2]. Some other tasks concern the analysis of the contents of images. Printed text is analyzed to find all abnormalities: characters with identical shapes [3] or irregular fonts [4], or lines that are skewed, misaligned, bigger or smaller than others [5]. Graphical elements of documents are another subject of research, such as signatures, logos or stamps [6], for instance.

While image analysis is the main field of document forensics, we believe that Natural Language Processing (NLP) and Knowledge Engineering (KE) could be used to improve the performance of fraudulent document detection. Document is not only an image: it contains textual information that can be processed, analyzed and verified. The aim of our work is to provide an Image-Text parallel corpus with a view to create a toolbox that could be combined with previously presented tasks.

In order to build a representative and sound framework for this kind of research, a dataset of genuine documents is needed. In the case of synthetically-created document datasets, the documents fields are generally filled with random information and are thus unrealistic. Such datasets of document images with randomly generated forgeries are good to apply image-based methods, but do not satisfy the needs of information-based methods since information is fake and inconsistent.

Therefore we decided to create a real-life dataset of documents with both images and texts and share it with the community[1]. Section 2 presents the collection of document images while Section 3 describes the OCRed text corpus. Finally, we discuss the use of such a corpus for the document forensics community (Section 4).



```
Num:2025     Serie:01
12-10-2016   12:52:49        (V1)

    kg          €/kg          €

#001 ABATS-VOLAILLES
   0.194         8.80        1.71
TVA      5.50%


Assiette(1 op)          1.62
TVA (5.50%)             0.09
TOTAL...........        1.71
Remis:                  1.71
Change:                 0.00
```

Fig. 1. Example of receipt

II. Document image dataset

Creating a public administrative document dataset is a difficult task: most of documents are personal or sensitive, they contain private information about individuals, administrations or companies and people want to keep originals [3]. One type of document that people easily accept to share is receipts, since they preserve anonymity and for most people and in most cases, saving them is essentially useless.

From December 2016 to June 2017 we collected around 2,500 documents by asking members of the L3i laboratory, families and friends. After removing receipts which are not French, not anonymous, not readable, scribbled or too long, we captured 1969 images of receipts. To have the best workable images, we captured receipts with a fixed camera in a black room with floodlight. Receipts were placed under a glass plate to be flattened. Each photography contained several receipts, so we extracted and straightened each one of them to obtain one receipt per document without much borders. The resolution of these images is 300 dpi.

These pictures are split into 3 categories corresponding to their origin:

- 454 receipts from one same franchise with homogeneous size (23% of total corpus),
- 148 receipts from other shops of the same retail company (8%),
- 1367 receipts from other stores (69%)

Images have different sizes because of the nature of receipts: it depends both on the number of purchases and on the store that provides the receipt. The first subset is homogeneous from

[1] The dataset is available online: receipts.univ-lr.fr

TABLE I
AVERAGE OF CHARACTERS AND CORRECTIONS PER RECEIPTS TYPE

	Characters/Receipts	Corrections/Receipts
Same-shops-type Receipts	413.7	6.7
Same-brand Receipts	811.0	13.3
Others Receipts	719.0	6.6

the layout and size point of view, while the second one is homogeneous at the layout level. The third one contains very different receipts, from different stores or restaurants, with different fonts, sizes, pictures, barcodes, QR-codes, tables, etc. There is a lot of noise due to paper type, the print process and the state of the receipt, as they are often crumpled in pockets or wallets and generally handled with little care. Noises can be folds, dirts, rips. Ink is sometimes erased, or badly printed. This is a very challenging dataset for document image analysis.

III. TEXT DATASET

To extract text from images, we applied Abbyy Finereader 11's Optical Character Recognition engine. Since image quality is not perfect, so are the OCR results. The dataset we propose is actually the result of an automatically corrected corpus, in which we corrected the most frequent errors, such as "€" symbols at the end of lines or "G" characters (for "grammes") after sequences of 2 or 3 digits. The Table I shows the average of characters and automated corrections per documents for each subset. Of course, after these corrections there are still OCR errors, and it would be interesting for the community to improve quality of low resolution documents OCR. A participative platform will be implemented to correct texts and get a sound ground truth.

On the first subset, receipts are generally very small and contain 17 lines on average. On these 17 lines, we can always find:

- 4 or 5 lines with the name of shop and its contact information
- 1 line for the table header
- 1 line about the total amount
- 2 lines about payment information
- 1 line about receipt details and time and date information
- 1 or 2 lines of thanks

To these lines are sometimes added a few lines about fidelity card or to encourage clients to sign up for it. Finally, the other lines detail the price and title of the purchased products. This subset stems from very little shops and most of the purchases consist in meals and other staples.

As to the other two subsets, the length and content of receipts is more variable: the average number of lines is 30. In France, sellers have no obligation to provide receipts to private persons for the sale of products. Consequently, standards are not strictly followed regarding the content of receipts, which is thus rather irregular, sometimes not even containing the shop's name or that of products as shown in Figure 1.

The text of the receipts is very challenging to extract, analyze, model and verify because of the way the information is expressed in them. In general, approaches to address document understanding are based on NLP techniques using lexicon and syntactic rules, which generally requires well-formed sentences consisting of dictionary words, and following grammars.

Working with the texts found in receipts does not allow the use of such approaches. We have to work with the partial structure of these "semi-structured" documents to create a context and to be able to tag each part of document. We can not apply Part of Speech tagging but we can for example detect recurrence and patterns to tag and extract the price or the name of a product.

Another difficulty of this type of documents is the need to be concise to put all information on a small paper. Indeed, receipts contain numerous acronyms and abbreviations, in addition to named entities, which provides a very hard challenge for information processing. For example, we can find expressions such as "BRK SQR 1L PJ POMM 1.54€" for "Brick square 1 litre Pur Jus de Pomme 1.54" (i.e. "one liter brick of pure apple juice").

IV. A DATASET TO BE FALSIFIED

This parallel dataset of images and texts is intended to undergo realistic forgeries. By "realistic forgeries", we mean modifications that could happen in real life, as in the case of insurance fraud when fraudsters declare a more expensive price than true for objects that were damaged or stolen. Victims of fires or theft have to provide evidence of purchases, namely receipts or invoices, to prove their existence. Falsifying a receipt to earn more money from insurance is very tempting and quite simple.

Each part of receipt information is verifiable with external knowledge, and we think that suspicious details can be detected by semantic comparison. To prove this assertion, we need fake information, therefore fake - but real-life-based - texts. A fraud campaign with non-specialists of fraud will be organized to get realistic falsifications (price raises, changes of product titles, hotel address changes, etc.).

Having real and original documents lets us know the truth before falsification. This is very useful to analyze the falsification process and compare versions. So far, image-based and text-based methods have already proven useful to detect distinct cases of fraud [7]. Image-based methods can detect poor falsification, but not the perfect imitations; text-based methods can detect unlikely information, but not the reasonable frauds. The combination of the two types of approaches should be able to allow further improvement to fraudulent document detection. The receipts dataset will provide a unique benchmark to test and evaluate such approaches.

V. CONCLUSION

This new dataset, composed of 1969 images of receipts and their associated OCR results is a great opportunity for the computational document forensics community to evaluate and combine image-based and text-based methods for the detection of fake, forged and falsified documents.

References

[1] S. Shang, N. Memon, and X. Kong, “Detecting documents forged by printing and copying,” *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, p. 140, 2014.

[2] S. Elkasrawi and F. Shafait, “Printer identification using supervised learning for document forgery detection,” in *11th IAPR International Workshop on Document Analysis Systems*. IEEE, 2014, pp. 146–150.

[3] R. Bertrand, P. Gomez-Kramer, O. R. Terrades, P. Franco, and J. M. Ogier, “A system based on intrinsic features for fraudulent document detection,” in *12th International Conference on Document Analysis and Recognition, (ICDAR)*, 2013, pp. 106–110.

[4] R. Bertrand, O. R. Terrades, P. Gomez-Kramer, P. Franco, and J.-M. Ogier, “A conditional random field model for font forgery detection,” in *13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 576–580.

[5] J. van Beusekom, F. Shafait, and T. M. Breuel, “Document inspection using text-line alignment,” in *9th IAPR International Workshop on Document Analysis Systems*, ser. DAS ’10. ACM, 2010, pp. 263–270.

[6] B. Micenková, J. van Beusekom, and F. Shafait, “Stamp verification for automated document authentication,” *Computational Forensics*, vol. 8915, pp. 117–129, 2015.

[7] F. Goasdoué, K. Karanasos, Y. Katsis, J. Leblay, I. Manolescu, and S. Zampetakis, “Fact checking and analyzing the web,” in *2013 international conference on Management of data (SIGMOD)*, 2013, p. 997.