Find it! Fraud Detection Contest Report

Chloé Artaud, Nicolas Sidère, Antoine Doucet, Jean-Marc Ogier, Vincent Poulain d'Andecy

Find it! Fraud Detection Contest Report

Chloé Artaud, Nicolas Sidère, Antoine Doucet and Jean-Marc Ogier
L3i
University of La Rochelle
La Rochelle, France
Email: firstname.name@univ-lr.fr

Vincent Poulain d'Andecy
Yooz
Aimargues, France
Email: Vincent.PoulaindAndecy@yooz.fr

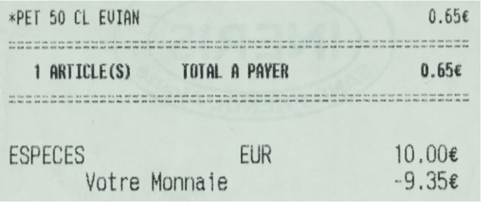
***Abstract*—This paper describes the ICPR2018 fraud detection contest, its data set, evaluation methodology, as well as the different methods submitted by the participants to tackle the predefined tasks. Forensics research is quite a sensitive topic. Data are either private or unlabeled and most of related works are evaluated on private datasets with a restricted access. This restriction has two major consequences: results cannot be reproduced and no benchmarking can be done between every approach. This contest was conceived in order to address these drawbacks. Two tasks were proposed: detecting documents containing at least one forgery in a flow of documents and spotting and localizing these forgeries within documents. An original dataset composed of images and texts of French receipts was provided to participants. The results they obtained are presented and discussed.**

I. Introduction

In the last decades, the explosion of the volume of digital document images and the development of consumer tools to forge these images has led to a huge increase in the number of corrupted documents. The development of many tools and methods to detect modifications has also increased but benchmarking remains a challenge.

With all the benefits it comes with, development of technologies has also an important side effect. For instance, number of amateur fraudsters has increased. Actually, it is quite accessible for many people to scan any document (payslip, bill, etc.), modify one or several critical field (name, date, amount of money, etc.) and print it with common devices and basic softwares (MS Paint, Gimp, etc.)

For institutions and companies, fraud can be considered as a real plague as shown by the many studies that were undertaken during previous years. A first one conducted in 2016 by Price Waterhouse Coopers reveals that 36% of the 5125 companies are victims of frauds. The estimated economic cost exceeds 50k$ for 53% of the surveyed companies. Beyond this financial cost, the reduction of the employees' motivation and, the cost on the company's relations with the public and its partners are major collateral damage directly impacting the company's trust capital. ICAR had a study conducted on banking fraud in Spain stating that almost half of Spanish users have experienced attempts to defraud. Out of the victims of successful fraud, 33% of them did not recover the money they lost and frauds cost on average 218 for each of them. In one hand, all these studies show the necessity to act in fraud prevention and detection. But on the other hand, as far as we know, no contest is organized on fraudulent documents and not so many studies are leaded on this topic. We hope that this contest will help to make the focus on this domain.



```
*PET 50 CL EVIAN                              0.65€
===================================================
   1 ARTICLE(S)      TOTAL A PAYER            0.65€
===================================================
ESPECES                  EUR                 10.00€
        Votre Monnaie                        -9.35€
```

Fig. 1. Genuine receipt

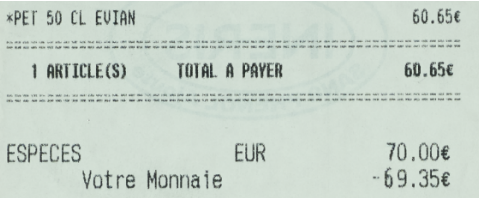
If these figures present fraud in a global way, one cant ignore frauds on documents. Because they can be easily modified with common and usual tools, frauds on documents have seriously increased. For instance, a dishonest person can be easily attempted to modify amount of purchases on types of document admitted as evidence, such as invoices or receipts, in order to earn more money from insurance in case of theft or fire. Receipts can also be provided as expense report by employees. We can imagine there are cases of falsifications of name of purchased products, to respect constraints of reimbursement, or of the address of restaurant to prove the presence at the good place. For all these reasons, we choose for this contest to focus on this type of documents.

Recent research in document forensics are mostly focused on the analysis of images of documents. However, 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 this contest proposal is to provide an Image-Text parallel corpus and an unique benchmark to test and evaluate image-based and text-based methods.

II. Dataset

A. Corpus Collection

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



*PET 50 CL EVIAN 60.65€

1 ARTICLE(S) TOTAL A PAYER 60.65€

ESPECES EUR 70.00€

Votre Monnaie -69.35€

Fig. 2. Fake receipt

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, and we extracted and straightened each one of them to obtain one receipt per document. The resolution of these images is 300 dpi.

The size of images differs because of the nature of receipts: it depends both on the number of purchases and on the store that provides the receipt. The dataset 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.

To extract text from images, we applied Abbyy Finereader 11s Optical Character Recognition engine. Since image quality is not perfect, so are the OCR results. We automatically 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. Then, we proposed an online participative platform to correct OCR results and get a sound ground truth. The crowd-sourcing of the human correction is still in progress.

B. Alteration

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. We need to get realistic falsifications (price raises, changes of product titles, hotel address changes, etc.).

Synthesizes this step, by an automatic algorithm that randomly changes some characters for instance, would mechanically induce a bias that we absolutely wanted to avoid. From this statement, the only way to meet the previous criteria is to organize workshops with volunteers to become one-day fraudsters. To increase the diversity of quality and precision of forgeries, workshops are open to PhD students and post-docs from various labs with various skills with image tools. To have a representative sample of real-like fraudsters, it is quite important not to restraint this job only to members of our computer science lab but to enlarge the scope of our project to a non-expert public, at least people who are not used to work with digital documents or image processing tools.

As we said, the aim of these workshops is to try to reproduce real forgeries in real conditions of fraudsters. All fraudulent acts are made using common and widespread material or tools, ie. a standard computer equipped with Windows 10 and several image editing softwares, at the users choice. For each receipts, image and text was modified simultaneously. We obtained 250 altered receipts, containing several types of modification, on all receipt information.

C. Corpus of the training phase

Data to process was organized as a set of couples image and text files :

- An image file formatted in png, representing one receipt that can contain one or several forgeries
- A text file containing a textual transcription of the content of the receipt

Participants can use only images, only texts, or both images and texts.

For the first task, we provided a set of 500 documents, containing 6% of altered documents. An XML file of ground-truth shows the name of the documents and whether they are genuine or fraudulent.

For the second task, we provided 100 altered document (images and texts), with 2 XML files defining ground truth:

- For images, the XML file contains the coordinates of each modification, as follows: x and y are horizontal and vertical coordinates of the top-left point of the rectangle bounding box that have height and width. All measurements are in pixels.
- For texts, XML file contains the tokens, delimited by spaces in the text file, that are forged and the line and the column where they are located. If the whole line is modified (or append), the forged value contains the complete line. If the forged value is empty, it means that information has been deleted.

There was no overlap in the fraudulent documents between the corpus of Task 1 and the corpus of Task 2, so participants could add the 100 documents from Task 2 to improve learning in Task 1.

D. Evaluation tools

We provided to participants three tools to help them to train their algorithms:

- EvalT1.py evaluates the detection of modified documents among others in a set of documents containing both genuine and modified documents. The evaluation script produces a CSV file with the Precision, Recall and FMeasure results, and, for their information, the ID of each receipt and its status (True Negative, False Negative, True Positive, False Positive).
- EvalT2-img.py evaluates the spotting of one or several modifications in a set of document images. The evaluation script produces a CSV file with, for each receipt, its name and the Jaccard index between the set of pixels covered by their localization results and those of the Groundtruth.

- EvalT2-img.py evaluates the spotting of one or several modifications in a document OCR output (text file). The evaluation script produces a CSV file with, for each receipt, its name, and 3 measures of Jaccard index corresponding to 3 different sets:
  - set of the lines covered by their localization results,
  - set of the lines and column covered by their localization results,
  - set of the lines, column and length of token covered by their localization results.

This choice is due to the complexity to localize a fake information in a line, and leave the possibility to have a less severe metric.

E. Corpus of the test phase

The corpus provided for the test phase of task 1 respected the same proportions as that of the training phase, i.e. 30 false documents out of 500 (6%).For the corpus of Task 2, we provided only 80 false documents.

Participants had to send us an XML file with their results and we calculate their scores.

III. Submitted Methods

In total, 36 teams registered to the competition. 5 of them submitted results to Task 1, while only 2 did so for Task 2. The following subsections provide the descriptions written by participants to describe the approaches they experimented to tackle the first task.

A. Fabre

We combined deep-learning with fraud detections techniques to achieve more than 85% good guesses. We did not however directly fed our network with the genuine images. We pre-processed them before with Tampered Image Detection methods. We tried a few, but the best result were done by combining: ”Error Level Analysis”, ”Discret Wavelet Transform” and ”Grayscale” images. We then fed the three dimensions matrices, of the three methods, to a well known deep neural network, Resnet152.

B. Clausner

The CFraudChecker consists of nine check modules, each looking at a specific type of fraud. Each module returns a fraud likelihood value between 0 and 1. The method reports a detection if the sum of all values is greater or equal to 1. Therefore, fraud is detected if either one module is very confident or multiple modules return a small value. Text-based modules:

- Price variation check: Looks for price outliers
- Total to pay check: Looks at inconsistencies in article prices and the total to pay
- Missing text check: Looks for keywords which imply a specific piece of information, but that information is missing
- Discounts check: Looks for inconsistencies in discounts
- Quantities check: Looks for inconsistencies in Quantity * Article = Sum
- Date check: Looks for invalid dates

Image-based modules (using OpenCV):

- Colour check: Looks for unnatural saturation, blackness, or pepper noise
- Erased parts check: Looks for unnatural white areas or large homogeneous areas (which have no noise)
- Copy + paste check: Looks for identical copies of connected components in binarised image

A challenge was the noisy textual input data (partially corrected OCR output). A basic text normalisation is performed (remove spaces in prices, fix decimal points etc.) but this can be extended to cover more inconsistencies. The free parameters were tuned manually but this can be automated in future.

C. Verdoliva

We have implemented fusion strategies for both the detection and localization tasks.

We use three methods for detection task:

- *CMFD.* The first method has been recently proposed by our team [1] for copy-move forgery detection but can be also applied to detect inpainting-based manipulations.
- *Noiseprint.* The second method extracts a camera signature, called Noiseprint [2], [3], through a deep net which removes the high-level image content. If the image has been tampered with, an anomaly arises which can be discovered by comparing the image Noiseprint with a reference Noiseprint extracted from a set of pristine images.
- *StegoFeatures.* The third method, proposed in [4], performs forgery detection based on local image features and linear SVM classification. The local features, originally proposed in steganalysis [5], capture expressive micro-patterns in the high-pass filtered image.

Fusion. An image is declared forged if at least one method detects a manipulation.

D. Zampoglou

Our submission was based on steganographic features extracted from the entire image and used to train an SVM classifier ensemble to discriminate between tampered and untampered images. Steganographic features have demonstrated strong performance in similar challenges in the past [4]. In [4], a set of steganographic filters are passed over the image, and a co-occurence matrix descriptor is formed for the entire filtered image. [5] presented a set of 39 filters. In the approach of Cozzolino et al, each filter is evaluated via cross-validation, and the features produced by the most successful filters are concatenated into a final classifier.

We followed a similar approach, where we used cross-validation on the training set to find the most successful filters from [5] for the dataset, but instead of concatenating we trained individual classifiers for each filter feature, and the final result is produced by majority voting over all individual classifier outputs. The steganographic filters from Fridrich et al are that demonstrated the best performance in cross-validation are:

- s5x5_spam14hv_q1
- s5x5_minmax22v_q1
- s3x3_minmax22v_q1
- s3x3_minmax24_q1
- s3_spam14hv_q1
- s3_minmax34v_q1
- s3_minmax22v_q1
- s2_spam12hv_q1
- s1_spam14hv_q1

Each model is trained using bagging. The output for each model is calculated by averaging all bag outputs, and the final result is drawn using majority voting over all the models.

E. Cruz

This method aims at detecting parts of document image that are duplicated, for instance in case of modification of a string by copy-pasting some characters. Based on some previous works applied on natural scene images ([**?**]), developped method is based on the discrete cosine transform (or DCT). DCT is often used in image compression algorithm because of its abality of projecting an image (or part of image) with excellent properties of grouping energy level, and consequently allows to held major informations on only few coefficients. In our algorithm, we use this property to detect and identify areas of image with similar coefficients meaning a identical information.

IV. Results and Discussion

A. Detection Task

To calculate the candidates scores for the first task, we used the usual metrics for classification: precision, recall and f-measure.

Some of our candidates used Machine Learning algorithms. The first results they submitted were based from algorithms that had been trained on the learning corpus of Task 1 to which they had added the fraudulent documents from Task 2. This corresponds to 130 false documents for 470 true documents, that is about 22%, instead of 30 false documents out of a total of 500 (6%). This was not prohibited, so we asked participants to re-train their model only on the learning corpus of Task 1.

Table I presents the results obtained by the different participants, with and without the documents of Task 2 included in the learning process. Zampoglou send us two results : the first uses identical settings to his original full run, while the second uses a more representative class balancing during training to account for the reduction of tampered training samples.

The last result in Table I shows a perfect detection score: the method used finds the 30 fraudulent documents perfectly. This surprising result is certainly due to the fact that the corpus is very specialized. Indeed, the documents were all scanned by the same camera, with almost identical parameters. It would therefore be interesting to see whether this method obtains equivalent scores on a corpus made up of images from different cameras under different lighting and inclination conditions. We can cite the use of smartphones, for example, to scan cash register receipts, as part of applications for reimbursing mission expenses or fidelity accounts.

TABLE I
Results of Task 1

Candidate	Precision	Recall	F-Measure
Fabre	0.364	0.933	0.523
Cruz	0.857	0.4	0.545
Clausner	0.882	0.5	0.638
Verdoliva T1	0.906	0.967	0.935
Verdoliva T1+T2	0.935	0.967	0.951
Zampoglou T1	0.964	0.9	0.931
Zampoglou T1 balanced	1.0	0.9	0.947
Zampoglou T1+T2	1.0	1.0	1.0

B. Human Baseline

In order to compare these results to a human baseline, we asked 5 people to detect false documents on the test corpus of Task 1. To do this, an interface provided them with a receipt image of the test corpus and they had to click on the ”true” or ”false” button, which allowed them to display the following image. Each annotator processed the 500 images in the corpus, proposed in a random order, knowing the rate of fraudulent documents in this corpus. The annotators had several days to process the entire corpus, and had feedback on their results and those of the other annotators at mid-term. Table II shows their Precision, Recall and F-measure scores.

These scores show that it is difficult for a non-specialist human to detect a false document (many false negatives, hence a low recall), and that many documents appear suspicious even though they are authentic (many false positives, hence low precision).

We observe that the average processing time of a sales receipt under these conditions is 20 seconds per ticket, being concentrated on this single task. Indeed, beyond the quick inspection to detect visible anomalies (characters of an abnormal color or font, strange streaks...), the annotators checked if the information was coherent between them (sums of the prices corresponding to the total and the payment, good number of displayed articles, etc.) and if there was no aberrant information.

Of the 500 documents in the corpus, the annotators do not agree on the authenticity of 49 of them. In addition to this, they are all wrong about the classification of 9 other of these 500 receipts (all false negatives, i.e. undetected forged documents). We also calculated the Fleiss Kappa between our annotators, which is 0.4375. This measure is used to calculate the inter annotator agreement and is in an interval of 0 to 1. The Kappa in this case shows that the four annotators only moderately agree on the frauds they detect. In other words, some see fraud where others do not.

TABLE II
HUMAN BASELINE

Candidate	Precision	Recall	F-Measure
Human 1	0.75	0.5	0.6
Human 2	0.64	0.47	0.54
Human 3	0.69	0.37	0.48
Human 4	0.55	0.37	0.44
Human 5	0.45	0.33	0.38

TABLE III
TYPES OF FORGERIES IN TEST CORPUS

Type of forgery	Number of documents concerned
CPI	13
CPO	3
IMI	3
CUT	2
CPI + CUT	6
CPI + IMI	1
CPO + CUT	1
IMI + CUT	1

C. Forgeries Types

The test corpus of Task 1 contained on average 3.7 alterations on the image, which corresponds to 3.5 frauds on the text transcriptions. These modifications on the images were made using 4 procedures :

- CPI (copy and paste inside the document)
- CPO (copy and paste from an other document)
- IMI (creation of a text box imitating the font)
- CUT (deletion of one or more characters/words)

Table III shows the types of changes made to the image in the false documents.

A detailed comparison of our candidates' results shows that the errors of classification do not concern the same documents. Indeed, all false documents were detected by at least two methods and raised false positives were by only one method. Thus, 70 documents are misclassified by only one candidate, 8 by 2 candidates and 3 by 3 candidates. The latter 3 were also misclassified by 5, 4 and 2 humans respectively. They contain CPI+CUT fraud types for two of them and CPI+IMI for the last one. We can also observe that 5 false documents are perfectly detected as false by the 5 candidates methods, when one of them isn't by four of the five humans.

V. SECOND TASK

The second task of our contest brought together only two participants: Clausner and Verdoliva. If this does not really make it possible to compare the results, it seemed important to us to reward the effort of these two participants by presenting them all the same. The second task was to locate forgeries in documents. The methods used are as follows:

a) Clausner: Each check module [presented in III.B.] adds fraud areas to a shared binary image. Then, the bounding boxes of connected components in that image are reported as fraud. The text-based modules make use of a Tesseract OCR result to match the given input text with OCR output (which has location data).

TABLE IV
RESULTS OF TASK 2

Candidate	Mean	Standard Deviation
Clausner	0,091	0,222
Verdoliva	0,426	0,261
Clausner without 0	0,287	0,315
Verdoliva without 0	0,461	0,240

b) Verdoliva: Besides CMFD and Noiseprint [presented in III.C.], we use a further method, proposed in [6], which exploits double JPEG compression artifacts. *Fusion.* If double compression is detected, the related localization map is used without further information, as it is very reliable. Else, if only CMFD or only Noiseprint detect a forgery, the corresponding map is used. Else, if both CMFD and Noiseprint detect a forgery, we keep the connected components of the CMFD map which overlap the Noiseprint map.

We evaluate these results by a Jaccard index between the set of pixels found and the set of pixels of our field truth for each document. Table 4 therefore presents the mean and standard deviation of the coefficients taking into account all the documents, as well as the mean and standard deviation of the indexes which are not equal to 0, i.e. documents where fraud has been at least a little located.

VI. CONCLUSION

One of the main lessons of this contest is the low number of participants compared to the number of registered and interested persons. We explain this by the low state of the art scores of most known methods of Forgeries Detection compared to the results presented in the first Image Forensics Challenge organized in 2013 by the IEEE Signal Processing Society and presented in WIFS 2014. Indeed, we believe that some registered people preferred not to submit their results.

Another lesson concerns the preponderance of image-based methods over text-based methods. Indeed, as the receipts are not in natural language and have various structures, it is complicated to extract features for the Machine Learning methods. Moreover, the corpus is in French, which does not help international linguists extract information to process it effectively.

In conclusion, we can say that the received results show that the task of detecting false documents on this corpus is a solved problem thanks to computer vision features. Nevertheless, the fine detection of frauds, and their location in documents, is not so obvious, as shown by the few results received for the second task and the partial location of frauds.

The excellent results of two participants for the first task lead us to think that it would be interesting to propose a more complex corpus to treat, which would perhaps reduce the possible biases of our corpus. Indeed, perfectionist fraudsters could try to print the image and re-scan it, which would probably change the results of the proposed approaches. We are also thinking of extending the corpus with images taken by other cameras than the one used for the captures and other shooting parameters, which would perhaps complicate the detection of anomalies in the corpus.

Acknowledgment

We would like to thank the participants of this competition for their motivation and cooperation. We would also like to thank all those who participated in the creation of this corpus, whether for the collection of receipts, for the correction of OCR outputs, for the falsification of documents, or for the manual detection of false documents in the test corpus. Without them, neither the corpus nor the competition could have existed.

References

[1] D. Cozzolino, G. Poggi, and L. Verdoliva, "Efficient dense-field copy–move forgery detection," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 11, pp. 2284–2297, 2015.

[2] D. Cozzolino and L. Verdoliva, "Noiseprint: a cnn-based camera model fingerprint," *submitted*, 2018.

[3] D. Cozzolino and L.Verdoliva, "Camera-based image forgery localization using convolutional neural networks," in *submitted*, 2018.

[4] D. Cozzolino, D. Gragnaniello, and L. Verdoliva, "Image forgery detection through residual-based local descriptors and block-matching," in *IEEE International Conference on Image Processing*, 2014, pp. 5297–5301.

[5] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[6] T. Bianchi, A. D. Rosa, and A. Piva, "Improved dct coefficient analysis for forgery localization in jpeg images," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 2444–2447.