

Find-it! Fraud Detection Contest

C. Artaud, N. Sidère, A. Doucet, J.-M. Ogier, V. Poulain d'Andecy

ICPR2018 Contest Session, Beijing

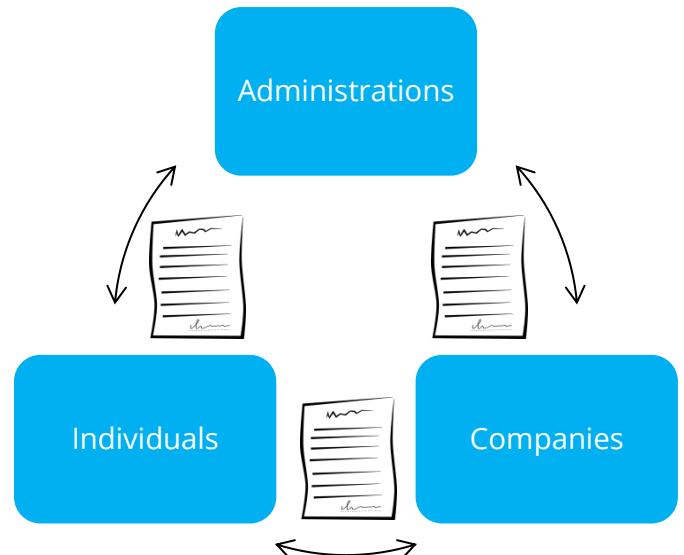


Find-it! Fraud Detection Contest

Introduction - Context

↗ Context

- > Huge loss of money due to frauds
- > Fraudulent documents are the support of numerous frauds
- > Documents have role of proof or attestation
- > Fraudulent documents can be :
 - Counterfeited: imitation of true documents with fake information
 - modified genuine documents: alteration of image



↗ Website : <http://findit.univ-lr.fr/>

Find-it! Fraud Detection Contest

The contest

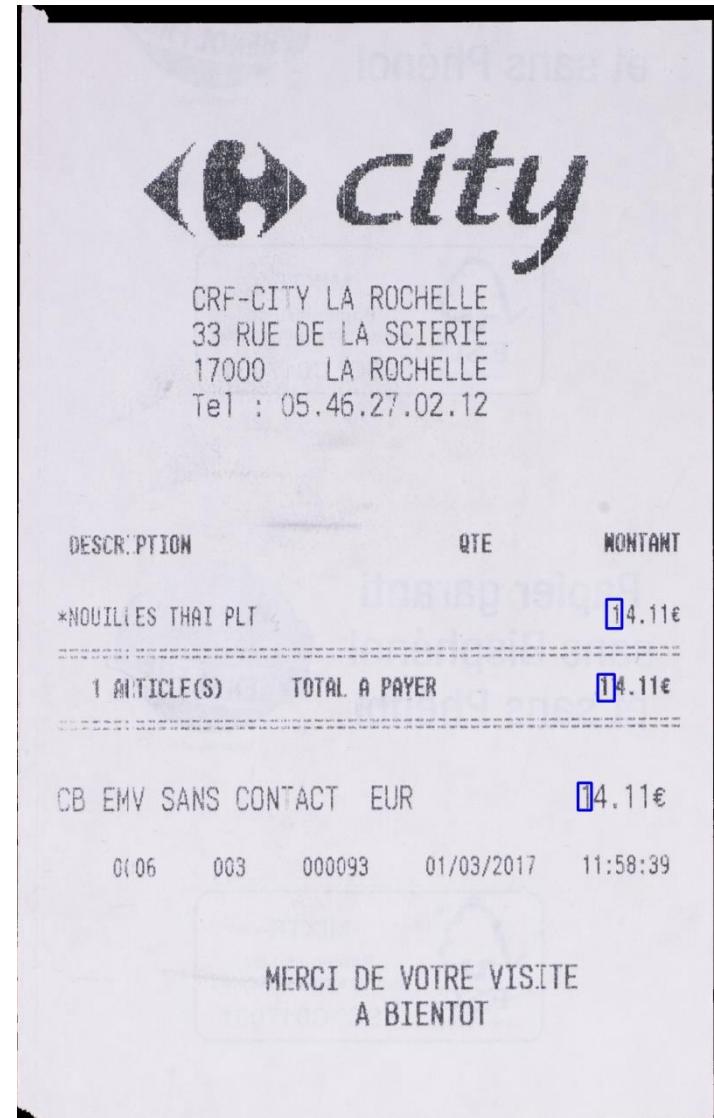
Goal of the contest

- > To have a common benchmark for document forensics community
- > For both images and texts checking approaches

2 tasks

- > Classification of documents:
 - Genuine documents
 - Forged documents
- > Localization of the fake information/alteration

A new dataset of forged and genuine documents



Find-it! Fraud Detection Contest

Need in document forensics

- Some works about alteration detection in images of documents
 - > Synthetically-created document images datasets used
 - > Random information
 - > Random frauds (often)
 - Good for image-based approaches, but not for semantic/text-based methods
- Real-life documents dataset : cash receipts
 - > Anonymous documents (not as salary slip)
 - > Prevalent documents (not as tax notice)
 - > Without sensitive data (not as medical certificates)
 - > With realistic fraud cases:
 - Insurance fraud
 - Mission expenses report

Find-it! Fraud Detection Contest

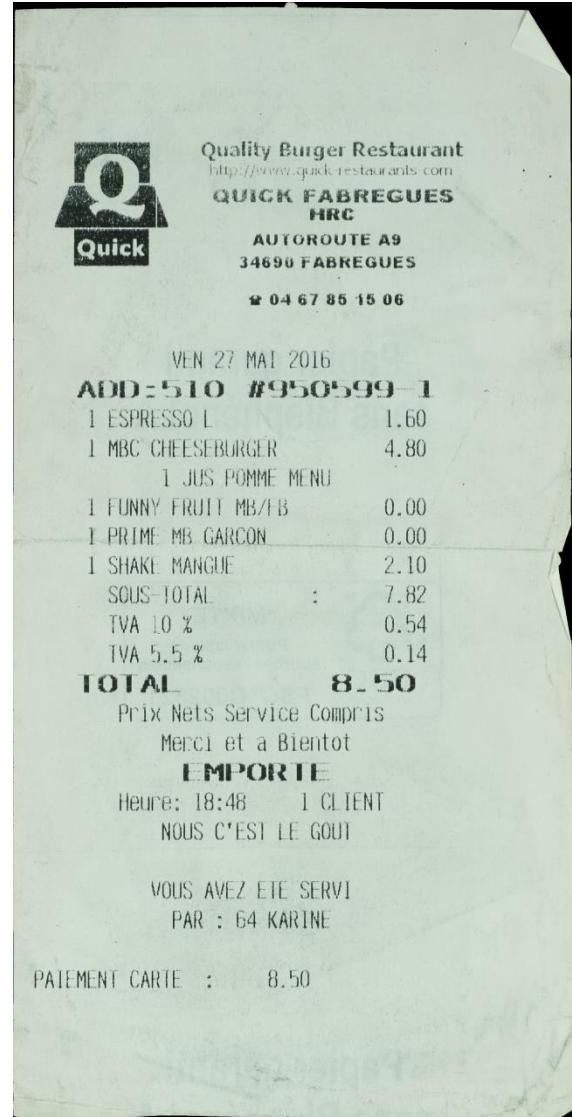
Dataset characteristics

➤ Receipts

- > From different stores, restaurants...
- > Different sizes
- > Variable quality
- > Only French language

➤ Images

- > Flattened receipts captured by one camera
- > Resolution: 300 dpi
- > Cropped and straightened



Find-it! Fraud Detection Contest

Dataset characteristics

→ Texts

- > Optical Character Recognition (OCR)
 - Abbyy Finereader 11
- > Automatic corrections of frequent errors
- > Crowd-sourced manual corrections :
<http://receipts.univ-lr.fr>
- > Still some errors



Chloé

Quality Burger Restaurant
hUp://■•/mv.guk.k iestduianls.com
gUKK FABREGUES
HHC
Qwick AU(OROUTE Ad
346ÜÜ FABREGUES
* 04 07 05 15 06
VEN 27 MAI 2016
Ann = * x i o fMsosgy i
1 ESPRESSO E 1.60
1 MBC CHEESEBURGER 4.80
1 JUS POMME MENU
1 EUNNY ERUII MM B 0.00
1 PRIME MB GARÇON 0.00
1 SHAKE MANGUE 2.10
SOUS 101AE : 7.82
TVA 10 % 0.54
TVA 5.5 % 0.14
101 AI 8-50
Prix Nets Service Compris
Merci et à Bientôt
I MI 'OKTE
Heure: 18:48 1 CLIENT
NOUS C'EST EE GOUT
VOUS AVEZ ETE SERVI

Submit

UNIVERSITE

La Rochelle

Find-it! Fraud Detection Contest

Dataset characteristics

↗ Fraud session

- > 25 one-day fraudsters
- > Paint, GIMP, Photoshop...
- > Images and texts modified

↗ Many types of fraud

- > In the choice of image manipulation
 - Copy-Move Forgeries (inside the document),
 - Splicing (a part from another document),
 - Imitation of font with a textbox tool
 - Deletion of a part...
- > In the choice of false information
 - Name of shop, address, phone number, prices, total amount, total paid, dates and hours...



Find-it! Fraud Detection Contest

Training phase - material

➤ Training datasets

- > **Task 1:** 500 documents including 30 forged
- > **Task 2:** 100 forged documents

➤ Ground Truth

- > **Task 1:** one XML file with the name and the forged value (1=forged, 0=genuine) of all documents
- > **Task 2:**
 - **Images:** one XML per document with the coordinates of the bounding-box of each alterations
 - **Texts:** one XML per document with the location (line and column) of modified words

```
<?xml version='1.0' encoding='utf-8'?>
<document>
  <fraud height="74" width="35" x="1420" y="2006" />
  <fraud height="62" width="25" x="1517" y="1540" />
</document>
```

```
<?xml version='1.0' encoding='utf-8'?>
<document>
  <fraud line="7" col="19" forged_value="18.14€"/>
  <fraud line="9" col="0" forged_value="" />
  <fraud line="9" col="0" forged_value="SALADE VERTE 50,64€"/>
</document>
```

Find-it! Fraud Detection Contest

Evaluation

↗ Evaluation metrics

- > **Task 1:** Precision/Recall/F-Measure
- > **Task 2 (image):** average of Jaccard Index computed for each document :

$$\frac{|A \cap B|}{|A \cup B|} \text{ where } A = \textit{set of the predicted forged pixels}$$

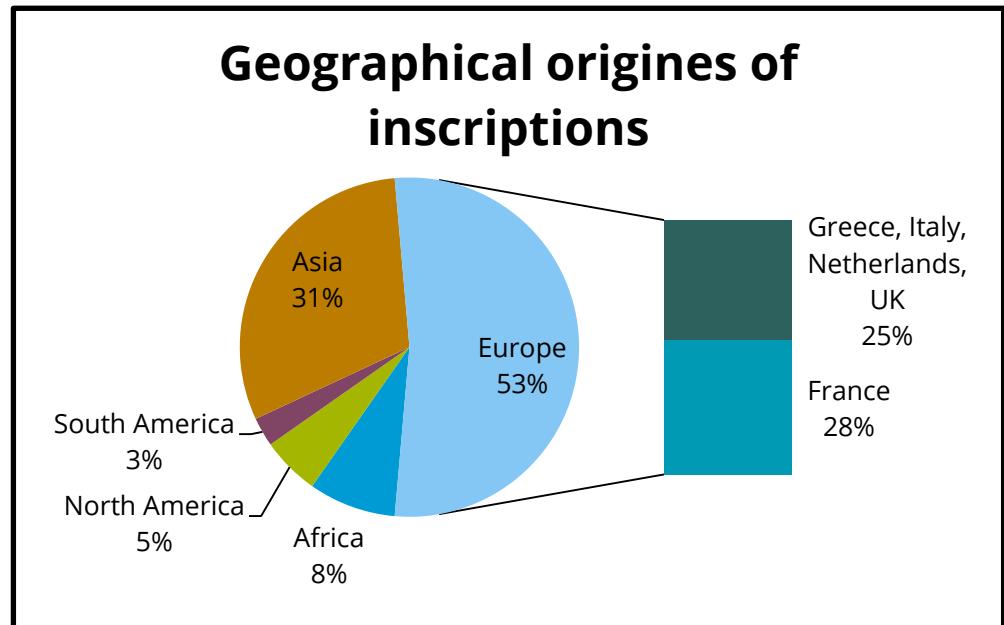
and $B = \textit{set of the ground truth forged pixels}$

↗ Evaluation tools and Ground Truth for training dataset was available for the participants

Find-it! Fraud Detection Contest

Interest in the contest

- 36 inscriptions
 - > 11 countries
 - > 5 continents
- Different affiliations
 - > 66% Academic
 - > 19% Industry
 - > 14% Other : Police...



- Interest for the 2 tasks and the 2 formats

| | Task 1 | Tasks 1+2 | Total |
|-------------------|----------|-----------|-----------|
| Only image | 5 (14%) | 4 (11%) | 9 |
| Only text | | 3 (8 %) | 3 |
| Both text + image | 2 (6 %) | 22 (61 %) | 24 |
| Total | 7 | 29 | 36 |

- But only 5 participants for the Task1 and 2 for the Task2

Find-it! Fraud Detection Contest

5th: Fabre's team – Coffreo, France

↗ Method

- > Combination of deep-learning with fraud detections techniques on images
 - "Error Level Analysis",
 - "Discret Wavelet Transform",
 - "Grayscale" images.
- > Deep neural network: Resnet152

↗ Results

| | Precision | Recall | F-Measure |
|-------|-----------|--------|-----------|
| Fabre | 0.364 | 0.933 | 0.523 |

- > Mostly forged documents are detected (good recall), but a lot of genuine ones too (bad precision)

Find-it! Fraud Detection Contest

4th: Cruz's team – L3i, France

➤ Method

- > Copy-Move Forgeries Detection on images
 - Discrete Cosine Transform

➤ Results

| | Precision | Recall | F-Measure |
|-------|-----------|--------|-----------|
| Fabre | 0.364 | 0.933 | 0.523 |
| Cruz | 0.857 | 0.4 | 0.545 |

- > All documents don't contain CMF (bad recall), but the ones who have it are mostly detected

Find-it! Fraud Detection Contest

3rd: Clausner's team – University of Salford, UK

↗ Method

- > Text-based modules:
 - Price outlier
 - Inconsistencies in article prices and the total to pay
 - Keywords which imply a specific piece of information, but that information is missing
 - Inconsistencies in discounts
 - Inconsistencies in Quantity*Articles = Sum
 - Invalid dates
- > Image-based modules:
 - 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
- > Threshold on the sum of the 9 scores to determine whether the document is fake or not

↗ Results

| | Precision | Recall | F-Measure |
|----------|-----------|--------|-----------|
| Cruz | 0.857 | 0.4 | 0.545 |
| Clausner | 0.882 | 0.5 | 0.638 |

Find-it! Fraud Detection Contest

2nd: Verdoliva's team - University Federico II of Naples, Italy

↗ Method

- > 3 image-based modules:
 - CMFD and inpainted manipulations : dense-field techniques
 - Noiseprint: extraction of camera signature through a deep net
 - StegoFeatures: local image features from steganalysis and linear SVM classification.
- > An image is declared forged if at least one method detects a manipulation

↗ Results with only T1 dataset and with T1+T2 datasets

| | Precision | Recall | F-Measure |
|-----------------|-----------|--------|-----------|
| Verdoliva T1 | 0.906 | 0.967 | 0.935 |
| Verdoliva T1+T2 | 0.935 | 0.967 | 0.951 |

Find-it! Fraud Detection Contest

1st: Zampoglou's team – CERTH-ITI, Greece

➤ Method

- > Steganographic features extracted, SVM classification
- > Features trained separately with bagging
- > Majority voting

➤ Results

| | Precision | Recall | F-Measure |
|-----------------------|-----------|--------|-----------|
| 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 |

- > T1 uses same settings than T1+T2 for training
- > T1-balanced uses a more representative class balancing during training

Find-it! Fraud Detection Contest

Overview of results

↗ Human Baseline

- > 5 humans (non experts)
- > For each image of the test dataset: say true or false

| | 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 |
| Zampoglou T1 | 0.964 | 0.9 | 0.931 |
| Verdoliva T1 | 0.906 | 0.967 | 0.935 |
| Zampoglou T1 balanced | 1.0 | 0.9 | 0.947 |
| Verdoliva T1+T2 | 0.935 | 0.967 | 0.951 |
| Zampoglou T1+T2 | 1.0 | 1.0 | 1.0 |
| Human5 | 0.45 | 0.33 | 0.38 |
| Human4 | 0.55 | 0.37 | 0.44 |
| Human3 | 0.69 | 0.37 | 0.48 |
| Human2 | 0.64 | 0.47 | 0.54 |
| Human1 | 0.75 | 0.5 | 0.6 |

Find-it! Fraud Detection Contest

Task 2

➤ Localization is a very difficult task

- > Only 2 participants send their results
- > We computed:
 - the average of the Jaccard Indexes of all documents
 - The average of the non-zero Jaccard Indexes (where one fraud was at least a little well located)

| | Average | Standard deviation |
|---------------------|--------------|--------------------|
| Clausner | 0,091 | 0,222 |
| Verdoliva | 0,426 | 0,261 |
| Clausner without 0 | 0,287 | 0,315 |
| Verdoliva without 0 | 0,461 | 0,24 |

- > Difficulty for CMF in choosing which is genuine and which is fake
- > Difficulty with combined method to find the localization on image of the found forged text

Find-it! Fraud Detection Contest

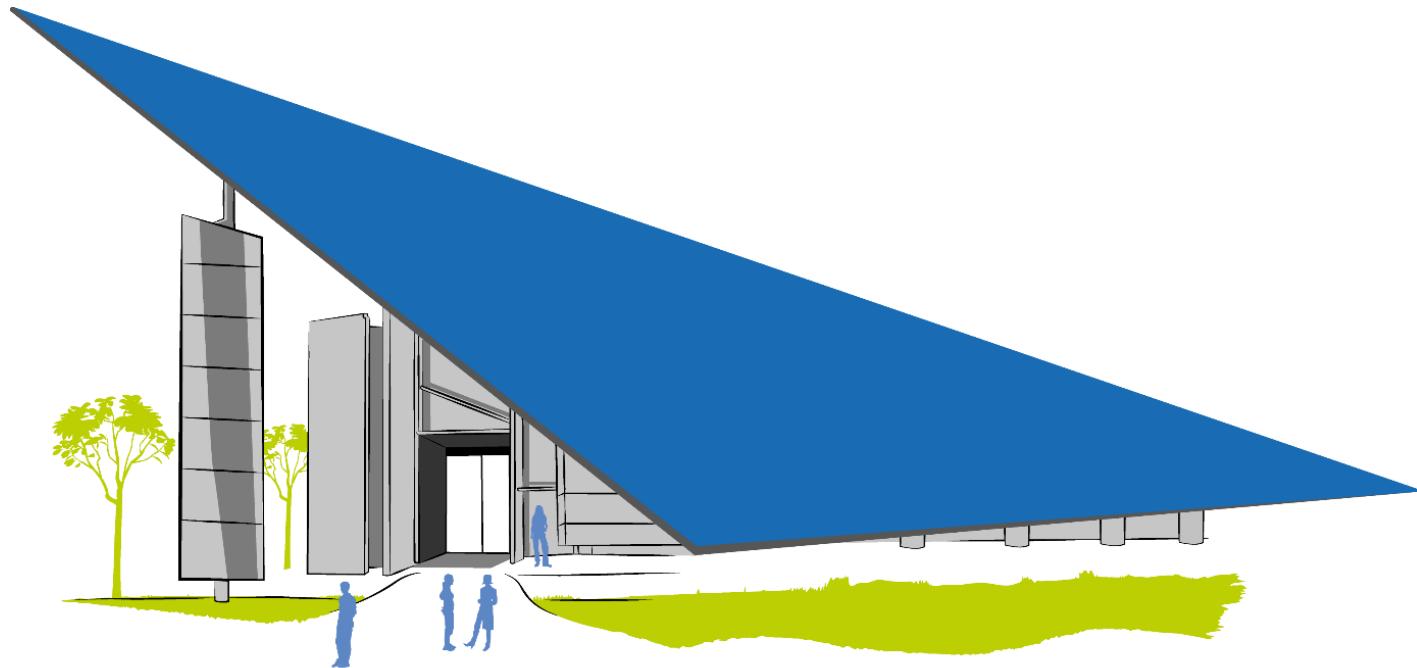
Conclusion & Perspectives

- A new dataset for both image and text approaches
 - > Available on demand on <http://findit.univ-lr.fr/download>
 - > With Images, Texts, Ground Truth and Evaluation tools
- Surprising results
 - > Image-based method for classification is already outperformed
 - > But localization of forgeries remains a non-resolved problem
- Diversity of the proposed methods
 - > But overall image-based methods
- Perspectives
 - > Adding noises to images (print&scan, filters) should decrease the results of classification task



**Thank you for your attention!
Any questions?**

Université de La Rochelle



L'Université nouvelle génération

23 avenue Albert Einstein - BP 33060 - 17031 La Rochelle
www.univ-larochelle.fr