Receipt automatic reader

Olga Maslova, Louis Klein, Damien Dabernat, A Benoit, Patrick Lambert

Receipt automatic reader

Olga Maslova
AboutGoods Company
Annecy, France
olga.maslova@aboutgoods.net

Louis Klein
AboutGoods Company
Annecy, France
louis.klein@aboutgoods.net

Damien Dabernat
AboutGoods Company
Annecy, France
damien.dabernat@aboutgoods.net

Alexandre Benoit
Univ. Savoie Mont Blanc
LISTIC
74000 Annecy, France
alexandre.benoit@univ-smb.fr

Patrick Lambert
Univ. Savoie Mont Blanc
LISTIC
74000 Annecy, France
patrick.lambert@univ-smb.fr

***Abstract*—Smartphones bring a new way to scan and digitalize written documents by taking pictures. This enables new document analysis applications to emerge. As a counterpart, unsupervised document capturing brings new challenges mainly related to target document localization and high quality text recognition. In this context, this work addresses automatic sale receipt understanding in an industrial context. It relies on the extraction of accurate and essential consumption data even with low quality receipt captures. We propose a tool chain that combines Deep Neural Networks and traditional image processing to ensure accurate automatic data extraction. The proposed workflow is evaluated globally by the analysis of the quality of the text recognition at the end of the processing.**

***Index Terms*—Sale Receipt Reading, Automatic document undesrtanding, Deep Convolutional Neural Networks, Object Detection**

I. Introduction

In the field of mass distribution, the insights on consumer behavior are key data that many companies are seeking for. Indeed, such information has a high added value as it provides accurate consumption statistics that can help in developing effective sale strategies. Currently, such data is manually obtained by recruiting consumers as panelists who are asked to scan the purchased products and to fill out forms. The obtained statistics are thus costly and cannot be applied to large populations, reducing the value and statistical significance.

Nevertheless, important information is present on sale receipts and this information could be obtained by an automatic reading system in order to address a much larger audience and at a much lower cost. In addition, the automatic reading of sale receipts could also be very useful to manage the validation of thousands of discount coupon awards in advertising operations. Consequently, a system able to automatically read receipts is of great interest. However, automatic information retrieval from such a document is somewhat challenging because receipts are, to name a few, not standardized, often damaged before being captured (crumples, tear, etc.), captured in poor acquisition conditions (no vertical alignment, with perspective effect, poor light, etc.) (see Fig. 3). In such a context, classical Optical Character Recognition software (OCR) cannot be applied directly.

This paper proposes an automatic sales receipt reader able to cope with such issues. It relies on a free mobile solution developed by the AboutGoods Company that enables any consumer to take and send a photo of his receipts. The goal is first to provide users with tools to help them in managing their budget, and second, provided user's consent, it enables to collect anonymous user consumption data that allows a large consumer community to be covered.

The proposed solution is based on a complete sale receipt processing workflow as presented in Fig. 2. The originality of this workflow, contrary to the solutions proposed in the literature, lies in the fact that it also works in the "difficult" cases mentioned above (see Fig. 3). The robustness of our solution is mainly obtained thanks to the combination of Deep Neural Networks with traditional image processing. This workflow is composed of two main steps.First, "Receipt Extraction" relates to the receipt localization within the image followed by its smart crop that provides a flat, vertically aligned sub-image restricted to the receipt area. The second step, "Receipt Reading", consists in the receipt text recognition which requires first the detection of text blocks (one or several lines) and second the reading of these blocks with an OCR. A third following step is a semantic analysis of the extracted text. This last step, which could help in correcting OCR errors and provide higher level analysis, is out of the scope of this paper. The presented workflow is validated through the evaluation of the quality of the recognized text since it serves as an input information for the following analysis. A receipt dataset captured by real users of our application is considered to measure the text reading quality.

The remaining of the paper is organized as follows. In section 2, we present related works. Section 3 provides the description of our processing chain. Results and performances are presented in section 4, and we conclude in section 5.

TABLE I: Comparison of online solutions

Solution	Extracted features				Diff. cases	Proc. Time
	Date	Total	Shop	Prod.		
Expensify	✓	✓	✓		✓	10 min
Wave	✓	✓	✓		✓	5 min
Taggun	✓	✓	✓			5 sec
Tiketi	✓	✓	✓	✓		20 sec
Previous work	✓	✓	✓	✓		30 sec
Our proposal	✓	✓	✓	✓	✓	20 sec

II. Related works

In the literature, few works deal with automatic receipt reading. In [1], authors first use classical image processing, then apply an off-the-shelf OCR and finally detect regular expressions. Szabo et al. [2] assume that the receipt image is clean enough to allow easy and efficient segmentation of text lines and characters. They then focus on character recognition which is performed by an SVM classifier with an RBF kernel. In [3], four main steps are considered: ticket localization, character localization, character recognition using an LSTM network and finally text analysis using regular expressions. In [4], classical image processing tools are used to get character blocks before applying the Tesseract OCR. As a general rule, the provided performance levels of all these works are good but do not allow for comparisons. Furthermore, the receipt images are generally assumed to be of good, even very good quality. Let us add to the above list our previous contribution published in [5]. It relies on Convolutional Neural Networks applied to classification and semantic segmentation in order to detect the position of a receipt. The Google Vision OCR was used to finalize text extraction.

Also, several online and commercial solutions are now available, for instance Taggun[1], Expensify [2], Tiketi [3], or Wave for Business [4]. However, the methods behind remain black boxes and present some limitations. They are typically restricted to general information extraction such as the transaction total amount and the date but do not extract the purchased goods detail. This is indeed a difficult task, especially to ensure that the data extracted contains correct and consistent information. In addition, most of them present long processing delays, sometimes up to minutes.

Some image acquisition constraints are also generally imposed to maximize the receipt reading quality. Vertical alignment, bright surrounding light and dark background are standard requirements that allow the available applications to perform well. But many applications still fail when provided with low quality/not standard images such as crumpled receipts, bad picture lighting, tilted receipts, and so on that are actually common cases in real application use scenarios, as illustrated in Fig. 1.

Table I summarizes the current application capabilities and reports the extracted features, processing time and their support for non-standard images reported as the 'difficult case' column. In addition, we report in this table the behaviors of our first proposition [5] and the solution proposed in this paper.

One can then conclude that the currently available solution on the market are too limited to enable customer loyalty.

We propose in this paper a solution that goes one step further. It differs from our previous work [5] by the use of different neural network structures and processing workflows. The two methods will be compared in the performance analysis section making use of a unified evaluation framework to show the improvements.

[1] https://www.taggun.io/
[2] https://docs.expensify.com/en/articles/4100-mobile-app
[3] https://ununuzi.es/servicios/mobile-ticketing
[4] https://www.waveapps.com/receipts

Fig. 1: On the left, a difficult case. On the right an easy case.

III. Workflow

A. Global description

Our workflow is described in Fig. 2. As mentioned in section 1, this workflow shows two main steps : receipt extraction and receipt reading.

In the receipt extraction step, we first perform the receipt localization and detection. The aim is twofold. First, it ensures that the image we process actually contains a receipt, such that one can move to the next processing step. This phase is necessary to eliminate non-receipt images, as users can make mistakes and upload unrelated images (selfies, etc.). Second, this step also detects the position of the receipt within the image thus enabling the removal of the background that is often a noisy and disturbing information that can make the OCR recognize text not related to the receipt. At the end of this first step, we perform a "smart" crop of the ticket. It relates to geometrical transformations in order to get a readable rectangular and vertically aligned frame limited to the receipt area without geometrical distortions.

In the second step, we begin by the detection of text blocks. Those blocks may consist of a single or several text lines. This step is necessary to obtain the optimal behavior of the OCR. Indeed, we found that the commercial OCR we tested obtain their best performances when they deal with small text regions. The use of OCR on the entire receipt or, conversely, on separate characters, does not lead to good results. Each

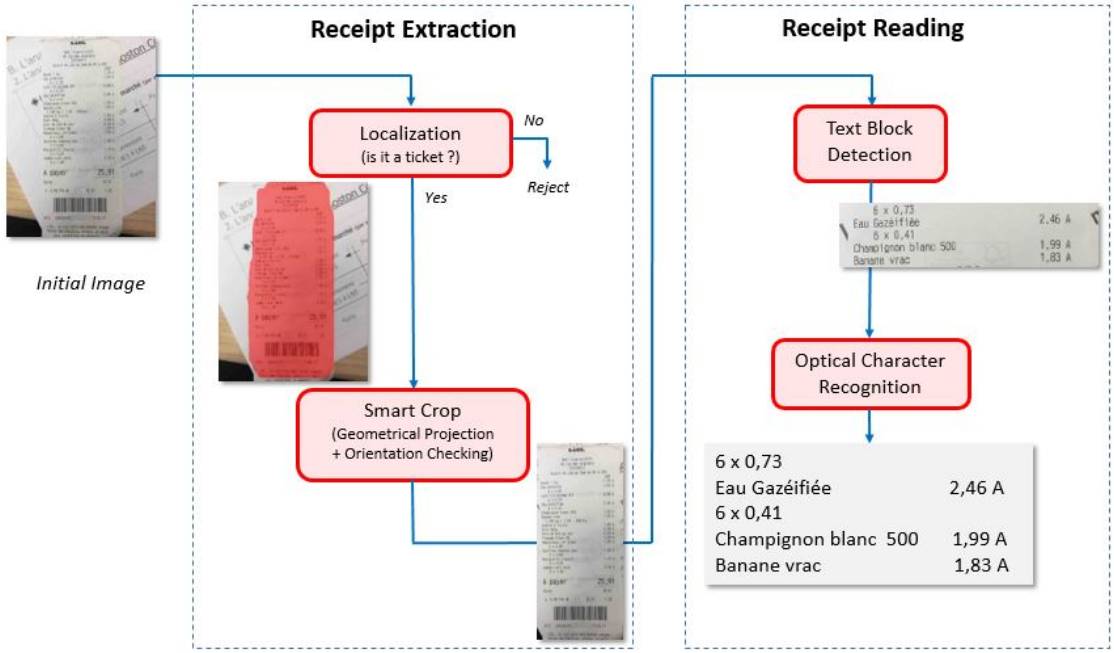

Fig. 2: Complete chain of the automatic reading system of sale receipts

Fig. 3: Bad quality receipts, with various damages and flaws

text block is then given to an OCR to get the text contained in the receipt. As previously mentioned, we used a commercial OCR.

B. Receipt extraction

1) Receipt localization: The first task is to detect and localize the captured receipt in an image taken with a smartphone, and to reject images without receipt. Receipt localization was proposed in our preliminary work [5]. It consisted in a rough receipt area localization followed by a receipt contours detection. However, the considered method leads to too wide bounding boxes that impacted on the text block extraction step coming next. In addition, the method could not manage multiple receipt instances in a single image. Then, in order to improve receipt boundaries detection and support multiple receipt instances, the Mask-RCNN [6] detection and segmentation method is considered. This network is pre-trained on the COCO Image dataset [7], which relies on the ResNet-101 [8] convolutional neural network architecture as a backbone for feature extraction. Only the final layers were adjusted : the Region Proposal Network (RPN) and the segmentation mask heads of the network have been fine-tuned and the bounding box classifier head is modified to comply with our two class problem (receipt/non-receipt). More into the details, in order to refine receipt boundaries that is critical for the following text recognition step, we used a specific Dice loss inspired from [9]. It relies on a specific penalty W applied to the original Dice loss in order to refine the receipt boundaries segmentation. The aim of this penalty is first to improve detection near receipt boundaries. Second, the penalty is enhanced for receipts with a small area w.r.t. the image size. Such a situation is encountered with long receipts captured at a far distance from the camera, which reduces the space (pixel-wise) between the receipt borders and its text areas. We compare this optimization with the classical binary cross entropy and the original Dice loss in the experimental section.

For each receipt instance detected in an image, we can express the new Dice coefficient in the following way : let G be the ground truth Boolean image and P be the Boolean image representing the receipt prediction mask, where True values represent the receipt pixels and False values represent the background. We define A as the result of a sliding average on G with a kernel of size $k \times k$. Experiments showed that $k = 11$ enables for a good precision on the receipt segmentation task for all the considered image sizes. Next we

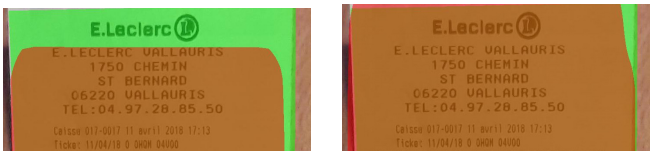

Fig. 4: Comparison of two receipt masks (brown) focused on the top border obtained with two different loss functions: binary cross-entropy (left) and weighted Dice loss (right). Green mask = ground truth.

define the elements $b_i \in B$, a Boolean image representing an enlarged contour of the receipt, as the following:

$$b_i = \begin{cases} 1, & \text{if } a_i > t_l \text{ and} \\ & \quad a_i < t_h \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where i denotes the pixel index (i = 1 to N, N = image size). Depending on the k value, the two above thresholds t_l and t_h are empirically fixed to 0.005 and 0.995. Such thresholding creates a reference receipt boundary mask introducing a tolerance with respect to the true position of the contour that is controlled by k. Penalty W is then defined to force the segmentation network to push the receipt boundary within this B region. The pixel level elements $w_i \in W$ are then defined by:

$$w_i = \frac{2 * b_i + S_g}{S_g} \tag{2}$$

where S_g is the surface area of the receipt, and $b_i \in B$. As a consequence, the smaller the object the stronger the penalty on its mask borders. The weighted Dice coefficient D is finally defined as:

$$D = \frac{\sum_1^N w_i^2 p_i g_i}{\sum_1^N w_i^2 p_i + \sum_1^N w_i^2 g_i} \tag{3}$$

where $g_i \in G$, $b_i \in B$ and $p_i \in P$.

Fig. 4 shows a comparison between the receipt masks obtained from models trained with the classical binary-cross entropy loss function and the proposed weighted Dice loss. We can observe a significant receipt surface gain with limited mask overfilling on the receipt boundaries when weighted Dice loss function is considered. Another advantage provided by the Mask-RCNN model is the multiple receipt instance detection capability. In the training process, in the case of several receipts per image, we compute the final loss function as the mean of per instance loss values. However, in the prediction mode, because of our specific application context that imposes the presence of a single receipt, images where multiple receipt instances are detected are rejected.

2) Receipt smart crop: This step addresses the problem of transforming and cropping a detected mask from an image to obtain a straight rectangular receipt image with the appropriate text orientation. With the use of classical image processing tools available in OpenCV, a multi-step algorithm is applied.

TABLE II: Architecture of CNN for detection of receipt's orientation.

Number	layer	activation
input (512x512 RGB image)		
2	conv3-16	relu
1	maxpool	
1	batch normalization	
2	conv3-32	relu
1	maxpool	
1	batch normalization	
3	conv3-64	relu
1	maxpool	
1	batch normalization	
3	conv3-128	relu
1	maxpool	
1	batch normalization	
3	conv3-128	relu
1	maxpool	
1	batch normalization	
1	global average pooling	
1	FC-layer-4	softmax

First we extract the polygonal representation of the predicted receipt mask boundaries. We then approximate this polygon by a quadrilateral. Finally, a homography is computed to remove the perspective effect. This operation transforms the receipt quadrilateral into the closest straight vertical (or horizontal) and rectangular shaped receipt image and crops it out of the initial image.

The resulting cropped image can be oriented in four directions: 0, 90, 180 or 270 degrees, where 0 degrees corresponds to a vertical receipt with the header on the top of the image. In our case, this orientation is necessary to extract as much information as possible since the text is clearly composed of horizontal lines in the reading direction. Other orientations perform poorly in the OCR step, and finding the correct orientation is thus crucial. For this purpose, a light CNN has been designed and trained from scratch. Its architecture is close to VGG-16 [10], with some tweaks to make it lighter. It uses blocks of Convolutionnal layers with kernels of size 3x3 combined with MaxPooling layers of size 2x2, with Rectified Linear units (ReLU) as activation functions. Batch normalization is used, and the final fully connected layers of the VGG-16 are replaced by a global average pooling layer combined with logistic regression. The input images are resized to the 512x512x3 dimension while the 4 outputs represent the possible receipt directions. The architecture used is described in table II. The Adam gradient descent optimization strategy is considered to train the model by minimizing the categorical cross-entropy loss.

C. Receipt Reading

1) Text Block Extraction: For this step only, considering that receipt generally has the same width, cropped images are resized to 512 pixel width while keeping aspect ratio to ensure that the processed receipts are homogeneous in terms of feature size (character size, space between text blocks). Next, considering that the resized receipt images are oriented

in the reading direction with horizontal text lines, a process is defined to highlight the text-free areas that are likely to isolate distinct text region blocks. Sobel filters are applied to highlight character boundaries. An automatic thresholding that relies on the classical Otsu method is applied next in order to obtain a binary image mask. A morphological image closing is finally applied to merge neighboring text pixels together. More into the details, a 8×3 structuring element is applied to join letters and words together in a single line or several line text blocks for very dense receipts. Those parameters are geometrically coherent with the resized receipt, and work well in our case study. This method allows us to separate the receipt into text blocks as shown on Fig. 2, that can be sent to an OCR API, in our case Google VISION.

2) Optical Character Recognition: Next step is critical: text extraction with an OCR. Open source solutions currently available like Tesseract do not compete with commercial grade OCR systems from Cloud Providers. However, the text recognition quality of all those solutions strongly depends on the precision of the provided text bounding boxes, and the overall quality of the taken pictures. With our industrial constraints, the best compromise between license pricing and additional engineering costs lead to the choice of the Google Vision OCR. However, certain limitations still exist even with commercial solutions that represent so far the main source of confusion in the semantic analysis that follows this step. Typical errors are generated by some text fonts specific to sale receipts and misalignment between the purchased goods and their price. For these reasons an in-house OCR solution can be considered as a part of future improvements.

IV. Performance Analysis

In this section first we focus on the performances of the two neural networks used for receipt detection, localization and orientation detection. The overall performance of the whole workflow is evaluated after the OCR step by the measure of the text recognition accuracy since it serves as an input for our following semantic analysis pipeline (out of the scope of this paper).

A. Receipt Segmentation

The Mask-RCNN network used for receipt detection and segmentation is trained and evaluated on 1200 images captured from the AboutGoods company applications in real use conditions. 1000 images are used for model fine-tuning while the remaining 200 images are used for performance evaluation (validation dataset). Each image was manually annotated with the precise mask of the receipt. Images are carefully sorted in order to represent each capture conditions in both datasets. The dataset is actually small, but enough to represent a variety of image capture conditions and provides good results in our production pipeline. Data augmentation is used to induce more variety into the training dataset and to obtain a network that generalizes better. The most frequent geometric transformations encountered in our case study are applied: horizontal flipping, scaling, small rotations around the vertical alignment of the receipt, shearing and translation. As the images we process come from smartphones, we can assume sufficient quality of resolution and lighting conditions. Table III reports the classical Intersection over Union (IoU) with 99% confidence interval computed on the test images for 3 different fine-tuning experiments, each considering a different training loss function.

TABLE III: Comparison of different loss functions with two metrics

Metrics	**Binary Cross Entropy**	**Dice loss**	**Weighted Dice loss**
IoU	0.920 $\pm$ 0.011	0.915 $\pm$ 0.009	**0.930** $\pm$ 0.007
Border accuracy	-1.50% $\pm$ 0.79%	-2.07% $\pm$ 0.45%	**-1.10%** $\pm$ 0.52%

The receipt detection quality strongly impacts on the following text block detection. The overall Intersection over Union (IoU) measures shows that the proposed Weighted Dice loss provides the highest segmentation quality. The obtained IoU values are very satisfactory considering that this metric is quite strict. Table III also shows the average and standard deviation error on the receipt boundary localization with respect to the size of the receipt. Close to zero negative values show that there is a slight receipt under-filling and once again, model trained with the weighted Dice loss lead to the best compromise with a reduced underestimation of the height and width while providing the smallest standard deviation.

B. Receipt Orientation Detection

Having a receipt oriented in the reading direction is critical to extract meaningful information. We used the CNN described in section III-B2 to classify the receipt orientation. We consider a dataset made of 2500 images of real sales receipts already cropped without any background. Those crops correspond to the output of the Smart crop module. The crops are randomly flipped across the 4 possible orientations in balanced sets. Those 4 orientations correspond to the ones obtained after the Mask-RCNN detection. On a 10-fold cross-validation, the network performed well and provides a 96.6% $\pm$ 1.2 exact match ratio.

Being trained on various images captured by real users, the network shows very good performance. Observed errors correspond to low image quality scenarios.

C. Text Recognition Quality

The final goal of our workflow is to get the best possible results from the OCR we use in this experiment (i.e. Google VISION OCR API). As it is a black box, we need to ensure that all transformations applied in the previous steps of the processing workflow gives the best results. To do this, we use common OCR metrics, the Word Error Rate (WER), and the Character Error Rate (CER) defined below :

$$WER = \frac{n_{substitutions} + n_{insertions} + n_{deletions}}{n_{total}} \quad (4)$$

where substitution is when a word has been replaced by another, insertion is when a word has been added between

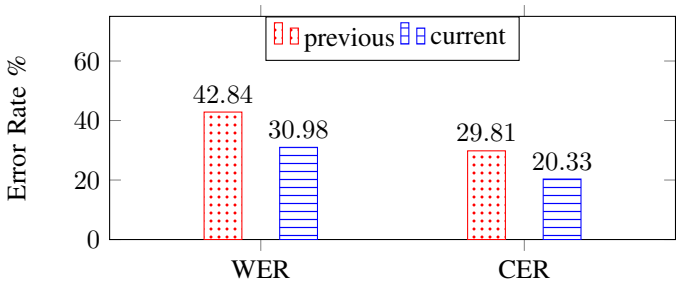

Fig. 5: Comparison between our previous [5] and current workflow

Fig. 6: On the left, the image cropped by our old system. On the right, the image cropped by the approach described in this paper.

two words and deletions is when a word has been removed. The interpretation is the following: the closer the metric from zero, the better the performance. The CER is the same metric, applied at the character level instead of the word level (Levenshtein distance).

Since the method works well in ideal situations with good quality image sensors and careful image capture, we focus on difficult situations that cannot be processed by on the shelf OCR. Those situations are actually frequently observed in the production pipeline. Our dataset is made of a selection of 15 receipt containing "difficult" images: with different exposition, angle and background environment in order to see the improvements brought by the new processing. The mean number of words per receipts in the dataset is 172. The mean number of characters per receipt is 854. This dataset is a good example of real conditions to which our workflow is exposed. However, these receipts are all human readable, and could be processed by a human operator.

To the best of our knowledge, the only available state-of-the-art method to compare with is our preliminary work [5], that is referred as "previous" in the following.

In Fig. 5 we can see that our current workflow described in section III performs significantly better than our previous work. We observe a significant improvement (-12% on the WER metric, and -9% on the CER metric). We should underline that both metric are quite strict that leads to rather high values that we may observe. However, one should note that for our applications some OCR errors (such as missing accents, special characters, dots vs commas confusion) do not penalize the following of our semantic analysis pipeline. As a final note, the proposed measures relate to difficult situations that can be compared to the more favorable situations that show, in average, a WER around 10% and a CER close to 6% that allows us a satisfying textual information extraction.

Since Google VISION is a black box, we can only conjecture how we improved our results. So, it's possible that it comes from the following improvements :

- the receipt boundaries detection is more accurate thanks to the use of state-of-the-art object localization fine tuned to our specific data combined with the weighted dice loss. (see Fig. 6, for a comparison with our old system)
- The receipt smart crop post processing followed by the text block detection provides sufficiently cleaned data to enhance the OCR recognition quality.

V. Conclusion

The proposed method consists in a complete processing tool-chain that detects and reads sale receipts from mobile phones captures. The complete process is evaluated with respect to the quality of the decoded text at the end of the pipeline thus providing a global performance metric. We take advantage of deep learning networks to deal with sale receipt detection and orientation detection. The proposed workflow enables an off-the-shelf OCR system to significantly improve its text recognition quality when applied to unsupervised captures of damaged documents.

Further work will focus on the receipt text semantic analysis and the design of a custom OCR dedicated to sale receipts.

References

[1] Bill Janssen, Eric Saund, Eric A. Bier, Patricia Wall, and Mary Ann Sprague. Receipts2go: the big world of small documents. In ACM Symposium on Document Engineering, 2012.

[2] Roland Szabo. A novel machine learning based approach for retrieving information from receipt images, put online : 15 april 2014. last accessed : 29 april 2019.

[3] Ozhiganov Ivan. Applying ocr technology for receipt recognition, put online : 7 april 2016. last accessed : 29 april 2019.

[4] A. Suponenkovs, A. Sisojevs, G. Mosns, J. Kampars, K. Pinka, J. Grabis, A. Locmelis, and R. Taranovs. Application of image recognition and machine learning technologies for payment data processing. In 5th IEEE Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE), November 2017.

[5] Rizlène Raoui-Outach, Cécile Million-Rousseau, Alexandre Benoit, and Patrick Lambert. Deep Learning for automatic sale receipt understanding. In IPTA conference, Montreal, Canada, November 2017.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. CoRR, abs/1703.06870, 2017.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In European Conference on Computer Vision (ECCV), Zrich, 2014. Oral.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. CoRR, abs/1512.03385, 2015.

[9] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016, pages 565–571, 2016.

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.