ICDSST 2022 PROCEEDINGS – ONLINE VERSION
THE EWG-DSS 2022 INTERNATIONAL CONFERENCE ON DECISION SUPPORT SYSTEM TECHNOLOGY
(editors)
Thessaloniki, UK, 23-25 May 2022

ICDSST 2022 on Decision Support addressing modern Industry, Business and Societal needs

ADVANCE: Automated Document Validation Aid with Nlp and Computer vision for fields Extraction

Martina Roberta Cecchetto, Giulia De Poli, Leone De Marco, Luca Pianta, Claudio Masolo, Matteo Bregonzio
3rdPlace SRL
Foro Buonaparte 71, 20121 Milan, Italy
martinaroberta.cecchetto@3rdplace.com, giulia.depoli@3rdplace.com
leone.demarco@3rdplace.com, luca.pianta@3rdplace.com, claudio.masolo@3rdplace.com,
matteo.bregonzio@3rdplace.com
web-page: www.3rdplace.com

ABSTRACT

Document Intelligence is a complex task which is getting more relevant in the last few years due to the need to efficiently process physical documents. Several businesses from insurance to banking sectors have to spend a large amount of man-hours manually inspecting documents to validate them, extract and transcribe relevant information. In this context, there is growing interest in automatic systems able to automatically process documents to stand in for or to support manual operations. Therefore, in this work we propose a business application able to automatically process tax documents and extract relevant content in a structured manner. The solution consists of an automatic aid for human agents in order to support manual processing. Stacks of documents are automatically classified in their parts and each relevant page is processed to extract relevant information which is then compared to the fields which were manually annotated. This crucial step helps to identify manual errors, resulting in a direct decrease in the time needed for the whole process by reducing the need for human agents to elaborate documents a second time. The system leverages cutting-edge deep learning models for classification and text extraction, applying a mixed approach of visual and text features. Several models have been trained on a multi-language real-world document dataset. The chosen solution shows good performance on both classification and information extraction, as well as the ability to be easily generalisable on future data.

Keywords: Document Intelligence, Deep learning, Document validation, Document OCR, Object detection, Computer Vision.

INTRODUCTION

Nowadays, despite the increase in digitalisation, most businesses are still processing documents manually to classify them, verify compliance or extract relevant content. This is due to the complexity of the task (e.g. content understanding, information retrieval, ..), even if it results in a time consuming and error prone process. In this context, a new trend is growing on building automatic systems able to process, organise documents and extract relevant information in a structured way.

These systems face different challenges as they aim to execute human tasks that some simple hard-coded rules cannot solve. Additionally, with technology available to most people, paper documents can be easily digitised: taking a picture with a smartphone, scanning documents, etc. This phenomenon generates a heterogeneous variety of digital documents with different formats and styles, even for the same original document.

In this paper we propose an innovative system that, leveraging deep learning models, identifies and extracts relevant information from tax documentation while assessing document quality. The three main problems tackled are (1) document quality assessment, (2) document classification from images, (3) extracting relevant key fields such as name, surname, VAT, etc.

The paper is organised as follows: a literature review of document intelligence and interpretation methods is initially presented. Later the proposed document processing system is discussed with a focus on the three main models: quality assessment, classification and field extraction. Finally, the fourth section shows experimental results and the final section summarises achievements and conclusions.

RELATED WORK

Automatic document processing has been widely researched[1] for different businesses in the last few years. GV et al.[2] proposed a framework for insurance documents classification and information extraction, Engin et al. [3] a classification system for banking documents. Along with document processing frameworks, literature shows a broad work on the three specific tasks tackled by our system: image quality assessment, classification and field extraction. In detail, document quality assessment[4] has been treated using different techniques such as defining sharpness images metric[5], applying a deep learning approach[6] or a multimodal quality assessment[7]. Document classification tasks can be based on different techniques, such as image classification or text classification. Recently, image classification for documents focuses especially on deep learning based methods[8]. Different well-known CNN architectures have been proposed: AlexNet[9], VGG-16[10], GoogLeNet[11]. Among them, Afzal et al.[12] suggest VGG-16 as the preferred one for document classification. Ghumade et al.[13] propose a deep learning approach based on textual features, while Bakkali et al.[14] propose a mixed approach using visuals along with textual features for document classification. Finally, different approaches have been proposed for field extraction, from geometric techniques[15] to deep learning methods[16].

AUTOMATED DOCUMENT PROCESSING SYSTEM

The proposed system, Figure 1, is composed of three modules: a) document quality assessment, b) classification and c) field extraction applied to tax documents.

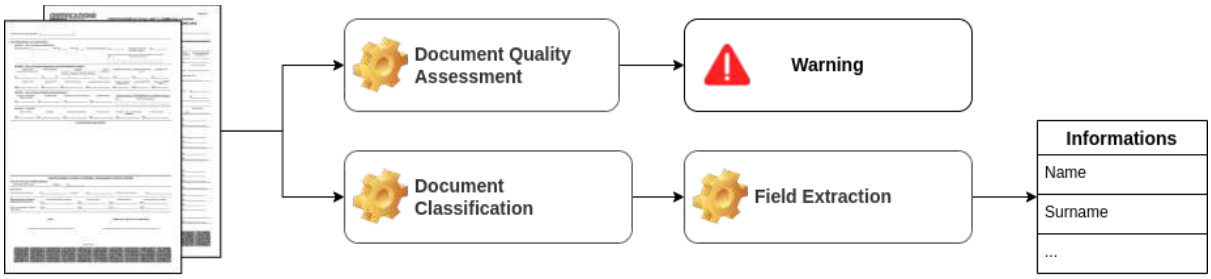

Figure 1: System architecture

Dataset

The dataset contains tax documents in different languages, such as legal documents used to certify employment income, self-employment or other income types; containing both personal data and other information. The document acquisition does not happen in a controlled environment, hence image quality changes based on the acquisition technique employed.

Currently, there are two document layouts that differ by the visual elements they contain. Therefore we divide the dataset into three different groups: layout-1 containing the page of interest for the first layout, layout-2 containing the page of interest for the second layout and other containing any other page as well as pages collected from any other document type. Due to the unbalanced nature of the dataset, data augmentation has been used during training.

Document quality assessment

The quality of the results during field extraction and OCR is strongly dependent on the quality of the images. For this reason, we need to quantify how much the image quality impacts the final result of our system. After manually labelling each image into two categories (high and low quality), we compared different classifiers: SVM, Gaussian Naive Bayes, Random Forest, K-Nearest Neighbors and chose the one with the best results, using a grid search for the tuning of the hyper-parameters of all models.

Given a grey scale image of 224x224 pixels, we use their distribution as input feature for all classifiers. As shown in Figure 2, digitally native document distribution shows a flat distribution with few high peaks compared to scanned document distribution.

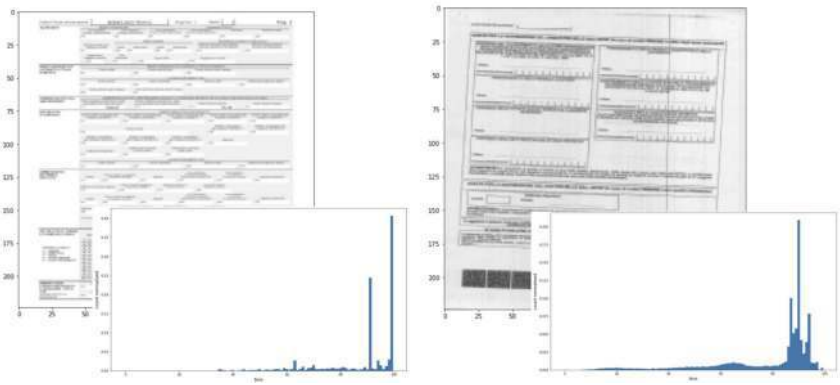

Figure 2: Document quality feature extraction for high and low quality images. Distribution has been computed on 100 bins

Document classification

The second step in our process consists in the identification, starting from a multi-page PDF document, of the exact page in which the relevant information is contained. We decided to frame this task as an image classification problem, in which the interesting pages are labelled as the positive class and all other pages are labelled as the negative class.

Initially, we compared two approaches: (1) SIFT[17], a traditional computer vision algorithm (2) and a more cutting-edge Convolutional Neural Network based method.

The two methods have advantages and disadvantages: traditional methods do not rely on Machine Learning thus are applicable without a true dataset; the drawback is that they are very sensitive to the quality of the images and their complexity. The Scale-invariant feature transform (SIFT) is more a detection algorithm than a classification tool: it extracts key points from a reference image which are then compared to the features of the images for which the detection is desired. It is a very simple algorithm that works well for kinds of images with well defined keypoints. A low quality image, very different from the reference image employed, could make this detection difficult.

On the other hand, Machine Learning methods need a dataset of labelled examples to learn from, and Neural Networks, especially CNNs, are probably the most data-hungry class of ML algorithms. The advantage is that they are very good at generalising, resulting in a more robust classification even for complex and low quality images. For this task, we did not have access to a big enough dataset for training a Neural Net from scratch. For this reason, we chose to exploit an architecture that was pre-trained on millions of images. The more shallow layers of CNNs trained on images act as "detectors" for simple patterns; for this reason, a popular technique is to keep these layers frozen and add a couple of trainable layers at the end of the network specific for the custom application. Specifically, we chose the VGG-16 architecture, trained on the imagenet dataset, which was cut at the fifth block and with three dense layers specific for the classification problem at hand[12]. The model, starting with pre-trained weights, has been further trained with our dataset for 50 epochs using Adam optimizer and categorical cross-entropy as loss function.

For reaching optimal results, we decided to improve the architecture by feeding the model not only images but also information about the text. We used a non optimised model from the tesseract package for a quick OCR of the pages. The output string was summarised by a bag of words model with a reduced vocabulary of 20 words. This vector, combined with the output vector of the second-to-last layer of VGG-16, is directly fed to the last dense output layer of our custom network.

Field extraction

The final step is to extract relevant fields from the identified image. The fields we are interested in are: social security number, name, surname, gender, date of birth, place of birth, province of birth.

Before extracting the relevant information, we need to extract the bounding boxes of each word on the document page. To do so, we pass the image to an OCR engine, in our case Google Vision API, that returns (x,y) coordinate of each point of the bounding box. After the OCR engine returns the bounding box we can finally extract the personal information we need. For this task, we exploit the logical construction of the social security number and search it with a regex. Using its position as anchor, we apply a geometric distance approach to find the other fields of interest.

EXPERIMENT - RESULTS

We tested our models using a total of 509 documents, each composed of different pages reaching a total of 3690 images that are divided as follows. For the quality assessment task 1994 low-quality and 1696 high-quality images, with 80%-20% train-test split. The SVM model shows better performances for the F1 metric and outperforms other classificators especially on precision while detecting high-quality documents and recall for low-quality documents.

Table 1: Comparison of quality classification on the test set in terms of Precision (P), Recall (R) and F1 score (F1)

		Score			
Model	**Class name**	**P**	**R**	**F1**	**Support**
SVM	low-quality	0.89	0.95	0.92	396
	high-quality	0.94	0.87	0.90	342
Gaussian NB	low-quality	0.89	0.58	0.70	396
	high-quality	0.65	0.92	0.76	342
Random Forest	low-quality	0.76	0.89	0.82	396
	high-quality	0.84	0.68	0.75	342
K-Nearest Neighbors	low-quality	0.74	0.88	0.8	396
	high-quality	0.82	0.64	0.72	342

The image classification task dataset is composed of 3 classes: layout-1 375, layout-2 119 and other 3196, with a 84%-8%-8% train-validation-test split. As the results show, the mixed approach outperforms both VGG-16 and SIFT.

Table 2: Comparison of quality classification on the test set in terms of Precision (P), Recall (R) and F1 score (F1)

		Score			
Model	**Class name**	**P**	**R**	**F1**	**Support**
Mixed VGG-16	other	1	1	1	237
	layout-1	1	1	1	31
	layout-2	1	1	1	9
VGG-16	other	0.99	0.97	0.98	237
	layout-1	0.91	0.94	0.92	31
	layout-2	0.69	1	0.82	9
SIFT	other	0.72	0.9	0.8	237
	layout-1	0.64	0.4	0.49	31
	layout-2	0.66	0.14	0.23	9

The field extraction task shows different accuracy metrics depending on the fields, as shown in Table 3. Different accuracies are highly dependent on the document layout since some fields such as surname are positioned in a well-separated area compared to others such as province of birth.

Table 3: Accuracy for field extraction

Fields	Social security number	Name	Surname	Gender	Date of birth	Place of birth	Province of birth
Accuracy	0.98	0.94	0.95	0.91	0.85	0.94	0.85

CONCLUSIONS

This paper proposes an innovative system to automatically process and extract relevant information from documents. The system has been validated on a dataset of Italian tax documents showing promising results, especially in managing different types of document layout and image quality and technique of acquisition.

Future improvements can be achieved by extending the dataset to other types of documents and testing other approaches, such as deep learning for the field extraction task.

REFERENCES

1. Cui, Lei & Xu, Yiheng & Lv, Tengchao & Wei, Furu. (2021). Document AI: Benchmarks, Models and Applications.

2. GV, A. R., You, Q., Dickinson, D., Bunch, E., & Fung, G. (2021, September). Document Classification and Information Extraction framework for Insurance Applications. In 2021 Third International Conference on Transdisciplinary AI (TransAI) (pp. 8-16). IEEE.

3. Engin, D., Emekligil, E., Oral, B., Arslan, S., & Akpınar, M. (2019). Multimodal deep neural networks for banking document classification. In International Conference on Advances in Information Mining and Management (pp. 21-25).

4. Ye, P., & Doermann, D. (2013, August). Document image quality assessment: A brief survey. In 2013 12th International Conference on Document Analysis and Recognition (pp. 723-727). IEEE.

5. Kumar, J., Chen, F., & Doermann, D. (2012, November). Sharpness estimation for document and scene images. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012) (pp. 3292-3295). IEEE.

6. Kang, L., Ye, P., Li, Y., & Doermann, D. (2014, October). A deep learning approach to document image quality assessment. In 2014 IEEE International Conference on Image Processing (ICIP) (pp. 2570-2574). IEEE.

7. Shen, A., Salehi, B., Qi, J., & Baldwin, T. (2020). A general approach to multimodal document quality assessment. Journal of Artificial Intelligence Research, 68, 607-632.

8. A. W. Harley, A. Ufkes, and K. G. Derpanis, “Evaluation of deep convolutional nets for document image classification and retrieval,” in 2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2015, pp. 991–995.

9. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

10. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

11. C. Szegedy et al., “Going deeper with convolutions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

12. Afzal, M. Z., Kölsch, A., Ahmed, S., & Liwicki, M. (2017, November). Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image

13. Ghumade, T. G., & Deshmukh, R. A. (2019). A document classification using NLP and recurrent neural network. Int. J. Eng. Adv. Technol, 8(6), 632-636.

14. Bakkali, Souhail, et al. "Visual and textual deep feature fusion for document image classification." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.

15. Rusinol, M., Benkhelfallah, T., & Poulain d’Andecy, V. (2013, August). Field extraction from administrative documents by incremental structural templates. In 2013 12th International Conference on Document Analysis and Recognition (pp. 1100-1104). IEEE.

16. Yu, W., Lu, N., Qi, X., Gong, P., & Xiao, R. (2021, January). Pick: Processing key information extraction from documents using improved graph learning-convolutional networks. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 4363-4370). IEEE.

17. Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91-110.