Verification Tasks through Deep Learning in a Semantic Information Extraction System

Angel Luis Garrido
University of Zaragoza
Universitat Politecnica de Valencia
Zaragoza, Spain
Email: garrido@unizar.es

Norman U. Bellorín
InSynergy Consulting S.A.
Madrid, Spain
Email: nbellorin@isyc.com

Alvaro Peiró
InSynergy Consulting S.A.
Madrid, Spain
Email: apeiro@isyc.com

Eduardo Mena
I3A, University of Zaragoza
Zaragoza, Spain
Email: emena@unizar.es

***Abstract*—Nowadays, the use of applications that validate identity documents is already widespread. The problem is when, for multiple reasons, the image sent to these systems is impossible to process, either due to lack of definition, because the image is incomplete, or simply because it is not the correct document, among other reasons. To bypass this type of errors, a prior image verification process is proposed, which avoids the superfluous use of resources for an unattainable validation. For this, we have designed a working methodology that combines several Artificial Intelligence techniques: computer vision, deep learning, and semantics tools. The proposal has been implemented and evaluated in a real environment with promising results.**

Keywords-Computer vision; Knowledge-Based Systems; Deep Learning.

I. Introduction

In any company, document management is a necessary task for developing business functions, but it requires much time and dedication. The reason is not because of its difficulty, but because of the need to be methodical and exhaustive, which may become impractical if working with many documents. To solve this situation, companies use Document Management Systems (DMS) to store, share, track, and manage files or documents. Among other functions, the DMS software can usually handle document digitization processes for verification tasks. For instance, users attach the images of identity cards captured by a camera, and their validation is typically done using external commercial tools. After the analysis, the validation software returns the information embedded in the image. The identity document may include the identifier, name, surname, and validity date. All this data allows identity verification by comparing it with the user's data stored in the system. An analysis carried out during six months with 359,885 identity documents for validating resulted in 24% of invalid documents that could not be processed by a validation software. The study was in collaboration with InSynergy Consulting [1], a significant company in Spain devoted to developing specific software for managing documents.

The contribution of this work is, given the high number of errors, to propose a verification methodology to identify the type of identity document attached to an image and validate its visual quality. This will avoid generating validation requests that will be erroneous, either due to the poor quality of the image or because the typology of the document was mistaken. Our proposed methodology combines several techniques: computer vision, deep learning, and ontologies. Although many works can be found in the literature that addresses this issue [2] [3] [4], as far as we know, none of them approach it from this point of view, and they do not use the support of semantic technologies to improve the results.

We have applied our proposal in a real environment in collaboration with InSynergy. In this context, we have performed a set of preliminary tests with an actual document dataset, and we have integrated it with the Analysis and Semantic Interpretation (AIS) system [5] [6]. This information extraction system uses the ICIX architecture [7], showing the feasibility and the benefits of our approach. The rest of this paper is organized as follows. Section II describes the methodology used. Section III depicts the experiments, and finally, Section IV addresses the conclusions and future work.

II. Methodology

In this section, we will explain the main steps (see Fig. 1) of the proposed methodology:

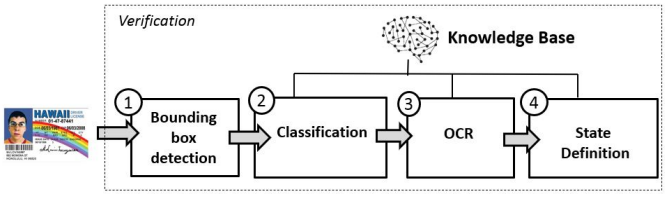

Figure 1. Overview of the proposed methodology.

1) *Bounding box detection*. In this step, the image of the identity document will be analyzed to extract only the area in which the document is located to reduce background noise and to help the system improve the classification of the document. The objective is to identify the edges from sudden changes in brightness, or limits defined through changes in reflectance or illumination in the image. To do this, an adaptation of the Canny [8] algorithm was made and it was empirically validated. This approach simplifies the image so that only the

 ISBN: 978-1-68558-089-6

outline of the objects is drawn. Since all edge detection results are easily affected by noise in the image, it is essential to filter out the noise to avoid incorrect detection.

2) *Classification*. An image classification algorithm based on deep learning techniques is responsible for classifying the typology of the attached identity document in the images uploaded by end users. A multiclass classification model has been developed. This model, which allows cataloguing the identity documents, differentiates the typology, the front and back, as well as the nationality of the document. The information to assist this process is obtained from a knowledge base of the AIS system, described in [5], which stores this data.
3) *Optical Character Recognition (OCR)*. The visual quality of the attached image will be determined when the fields of an identity document are read, and its quality will be verified. If most of these labels are read (name, surname, date of birth, gender, etc.), the visual quality of the document is considered acceptable. This process is assisted by an OCR tool and the knowledge base of AIS, which also stores the typology of the fields in each of the different identity documents.
4) *State definition*. The results of previous tasks are analyzed, and an outcome (satisfactory/failed) is determined. This step is also supported by the knowledge base, which includes a set of rules that allow you to decide if the document is suitable to be sent to the validation stage.

III. Experiments

Experiments have been executed under Linux using *Google Colaboratory (Colab)*, computer vision libraries such as *OpenCV 4.6.0*, and the Google Colab environment. On the other hand, *TensorFlow* and *Keras* libraries have been employed to develop and train neural networks. For the experiments, we used a dataset with 1,120 images of identity documents (front and back), passports, and an error class. This dataset can not be public due to data protection laws.

The results of the training are shown in Fig. 2. Then, cross-validation was performed on the entire data set, which placed the success rate at 98.77%. The final validation confusion matrix is shown in Fig. 3. It is observed how the multilayer model improves the results of the single-layer model. Finally, after implementing the entire system with its various stages, the validation error rate (errors and false positives) decreased from the initial 24% to 2%.

IV. Conclusions and Future Work

The contribution of this work is to propose a new resolution scheme for the problem of verifying the quality of identity documents, within the context of an information extraction system. In this way, we drastically improve the success rate of the validation process by avoiding carrying it out if the image quality is inadequate. To achieve this, we propose a methodology that combines computer vision techniques,

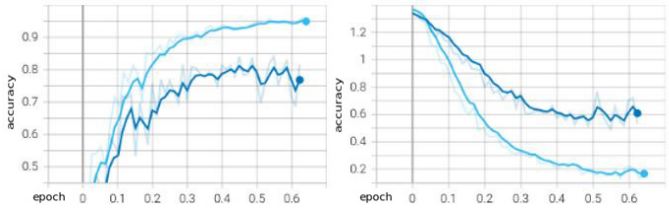

Figure 2. Training results, accuracy and loss, respectively. The top line is the multilayer model, and the bottom line is the single-layer model.

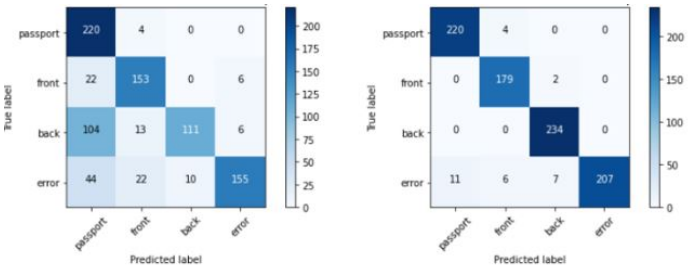

Figure 3. Confusion matrix using the single-layer or multilayer model, respectively.

classification through deep learning, and OCR checks, using an internal knowledge base to orchestrate all the steps. As future work, we are planning to use transfer learning techniques, to retrain models without losing previous training work, or without having to load the entire data set into memory.

Acknowledgment

This research work has been supported by the OTRI Project ICIX9 2019/0628-21 at the University of Zaragoza, by the Spanish project PID2020-113903RB-I00 (AEI/FEDER, UE), and by DGA/FEDER. The authors want to acknowledge Dr. Carlos Bobed for his valuable collaboration.

References

[1] http://isyc.com/, [last access Nov. 2023].
[2] M. K. Gupta, R. Shah, J. Rathod, and A. Kumar, "Smartidocr: Automatic detection and recognition of identity card number using deep networks," in *IEEE Sixth International Conference on Image Information Processing (ICIIP)*, vol. 6, 2021, pp. 267–272.
[3] L. Zhao, C. Chen, and J. Huang, "Deep learning-based forgery attack on document images," *IEEE Transactions on Image Processing*, vol. 30, pp. 7964–7979, 2021.
[4] D. P. Van Hoai, H.-T. Duong, and V. T. Hoang, "Text recognition for vietnamese identity card based on deep features network," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 24, pp. 123–131, 2021.
[5] M. G. Buey, A. L. Garrido, C. Bobed, and S. Ilarri, "The AIS project: Boosting information extraction from legal documents by using ontologies." in *ICAART (2)*, 2016, pp. 438–445.
[6] M. G. Buey, C. Roman, A. L. Garrido, C. Bobed, and E. Mena, "Automatic legal document analysis: Improving the results of information extraction processes using an ontology," *Intelligent Methods and Big Data in Industrial Applications*, pp. 333–351, 2019.
[7] A. L. Garrido, A. Peiro, C. Bobed, E. Mena, and C. Morte, "ICIX: A semantic information extraction architecture," in *Proceedings of the 25th International Database Engineering & Applications Symposium*, 2021, pp. 75–83.
[8] E. A. Sekehravani, E. Babulak, and M. Masoodi, "Implementing canny edge detection algorithm for noisy image," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 4, pp. 1404–1410, 2020.

 ISBN: 978-1-68558-089-6