# Information Extraction and Homogeneity Validation of an Identity Document

Fernando Broncano Morgado
*Grupo de Ingeniería de Medios*
*Universidad de Extremadura*
fbroncano@unex.es

Marcos Jesús Sequera Fernández
*Grupo de Ingeniería de Medios*
*Universidad de Extremadura*
marcosjesus@unex.es

Sergio Guijarro Domínguez
*Grupo de Ingeniería de Medios*
*Universidad de Extremadura*
sergiosgd@unex.es

José Carlos Sancho Núñez
*Grupo de Ingeniería de Medios*
*Universidad de Extremadura*
jcsanchon@unex.es

***Abstract*—The digitisation of identity documents has evolved from manual processes to automated solutions, driven by advancements in computer vision. This study introduces a methodology for extracting and validating information from Spanish identity documents. The approach leverages YOLO for region detection, ORB for feature extraction, homography for image alignment with a template, and OCR for data interpretation. A validation system is also implemented to ensure consistency between the visual inspection zone –VIZ– and the machine readable zone –MRZ–, alongside check digits to confirm redundancies. Document homogeneity serves as a critical first step in verifying document authenticity.**

***Index Terms*—Identity document, digital image, image recognition, text recognition, authentication**

## I. Introduction

The digitisation of physical documents into electronic information systems has always been a challenging task. Initially, document digitisation relied on manual processes involving the entry of information into the system. However, with advancements in computer vision, new mechanisms have emerged, enabling this manual task to transition into an automated process.

The development of automated information systems aims to facilitate the querying and storage of data. These systems are built on data models that enable structured information storage. Furthermore, they prevent issues such as duplication and verify data redundancy. Identity documents are a clear example of physical systems that require digitisation.

Current identity documents are still issued as physical documents, despite recent efforts to develop a digital version [1]. These documents contain relevant personal information that often needs to be processed mechanically. In response, the International Civil Aviation Organization proposed the creation of a standard document featuring a machine-readable zone [2]. Identity documents are structured into a visual inspection zone –VIZ–, which contains the document's information, and a machine readable zone –MRZ–, consisting of three lines of 30 OCR-B characters designed to be easily interpreted by machines.

Various organizations require systems for the mechanization of identity documents. In this regard, several studies [3], [4] have been proposed in the literature that address automatic reading and information recognition in identity documents to meet this need.

Tools for the mechanization of physical information must extract and process data from digital images. Traditionally, computer vision systems have relied on algorithms such as SURF [5], SIFT [6], and ORB [7] to study key points in images, or optical character recognition –OCR– platforms for character extraction. Additionally, with the rise of artificial intelligence and the introduction of more advanced models, solutions like *You Only Look Once* –YOLO– [8] have been proposed, enabling the detection of objects within an image.

This work proposes the development of a methodology for extracting information from a Spanish identity document. Additionally, it introduces a process for verifying the homogeneity of a document by comparing the machine readable zone –MRZ– with the visual inspection zone –VIZ–.

## II. Data Extraction from an Identity Document

Traditionally, reading a document has involved identifying the region of interest within an image and applying a character recognition algorithm to interpret its meaning. Based on this approach, a methodology for extracting information from an identity document is proposed, leveraging these characteristics.

The proposed methodology for extracting data from an identity document consists of the following steps: (1) identifying the region of interest containing the identity document within an image; (2) extracting features using ORB from that region of interest; (3) applying a homography process to map key points between the region of interest and the base template; (4) cropping the regions of interest according to a table of positions and sizes; (5) applying thresholding and optical character recognition to extract the data associated with that

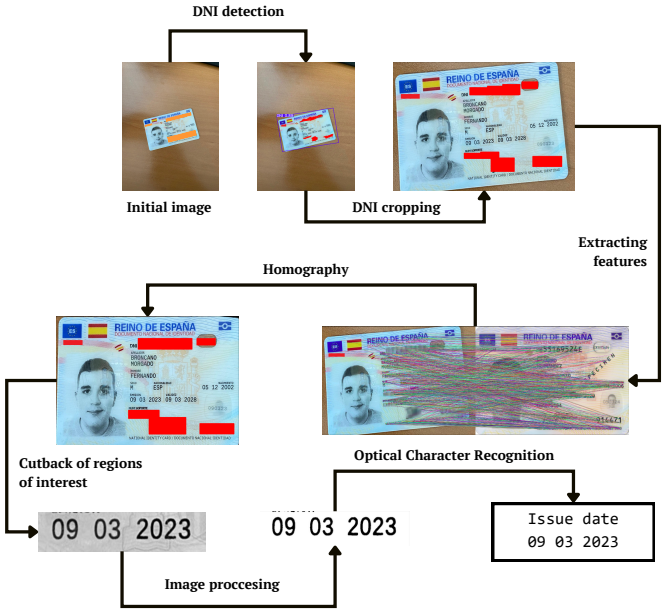


Figure 1. Information extraction using the proposed methodology

region of interest. This methodology is illustrated graphically in Figure 1.

Prior to applying this methodology, a template of an identity document must be selected. A table of regions of interest should then be created for this template and linked to the corresponding data. This table is used to perform cropping operations on the provided image. Additionally, a *You Only Look Once* –YOLO– model with oriented bounding boxes capabilities has been trained to identify the region containing the document with maximum precision.

## III. Homogeneity of Information in an Identity Document

Once the information from an identity document has been extracted, the existing redundancy can be verified using regular expressions and information redundancy checks. The primary redundancy feature in an identity document is the cross-checking of information between the visual inspection zone –VIZ– and the machine readable zone –MRZ–. An example of a machine-readable zone can be seen in Figure 2.

The verification of this information is performed once the data has been extracted and structured according to its corresponding feature. The information from both zones must match. Additionally, the Spanish National Identity Document –DNI– includes a redundancy mechanism through a security letter in the DNI number, as well as a consistency check between the issuance date, date of birth, and validity period.

Within the machine readable zone –MRZ–, several check


Figure 2. Example of a machine readable zone

digits –CD– are present. These check digits are calculated and compared with those present in the MRZ. In this way, redundancy within the machine-readable zone is verified.

## IV. Conclusions

This work proposes a methodology for extracting information from an identity document using YOLO, ORB, and OCR. Additionally, the validity of the information is verified by cross-checking redundancy and ensuring homogeneity between zones. As future work, advancements should focus on expanding the dataset used for region-of-interest recognition, as well as improving base templates and interest zone tables.

## Acknowledgements

This initiative is carried out within the framework of the funds of the Recovery, Transformation and Resilience Plan, financed by the European Union –Next Generation– and National Cybersecurity Institute –INCIBE– in the Project C108/23 “Detección de Falsificación de Documentos de Identidad mediante Técnicas de Visión por Computador e Inteligencia Artificial”.

## References

[1] *Real Decreto 255/2025, de 1 de abril, por el que se regula el Documento Nacional de Identidad.*, Boletín Oficial del Estado, 2025. [Online]. Available: https://www.boe.es/eli/es/rd/2025/04/01/255/

[2] *Doc 9303 Documentos de viaje de lectura mecánica*, Organización de la Aviación Civil Internacional, 2021. [Online]. Available: https://www.icao.int/publications/pages/publication.aspx?docnum=9303

[3] S. Carta, A. Giuliani, L. Piano, and S. G. Tiddia, “An end-to-end ocr-free solution for identity document information extraction,” in *Procedia Computer Science*, vol. 246, 2024, p. 453 – 462.

[4] M. K. Gupta, R. Shah, J. Rathod, and A. Kumar, “Smartidocr: Automatic detection and recognition of identity card number using deep networks,” in *Proceedings of the IEEE International Conference Image Information Processing*, vol. 2021-November, 2021, p. 267 – 272.

[5] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” in *Computer Vision – ECCV 2006*, 2006, pp. 404–417.

[6] D. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2.

[7] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: An efficient alternative to sift or surf,” in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.