Information Extraction and Homogeneity Validation of an Identity Document

Fernando Broncano Morgado, Marcos Jesús Sequera Fernández, Sergio Guijarro Domínguez, José Carlos Sancho Núñez

{fbroncano, marcosjesus, sergiosgd, jcsanchon}@unex.es

Covilhã, July 9, 2025

Financiado por la Unión Europea NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

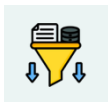
Table of Contents

Introduction

Identity Management

Physical Identity Media

Identity Violation



Digital Identity Mechanisms

Our Proposal:

Create a system that not only reads information but also **verifies its coherence** to ensure document authenticity.

Project Objectives

Main Objective

To build a set of algorithms that allow the **reading of identification elements** from the National Identity Document, as well as the **verification of their homogeneity**

Specific Objectives

Reading of Identification Elements

Homogeneity of Information Elements

Methodology

National Identity Document

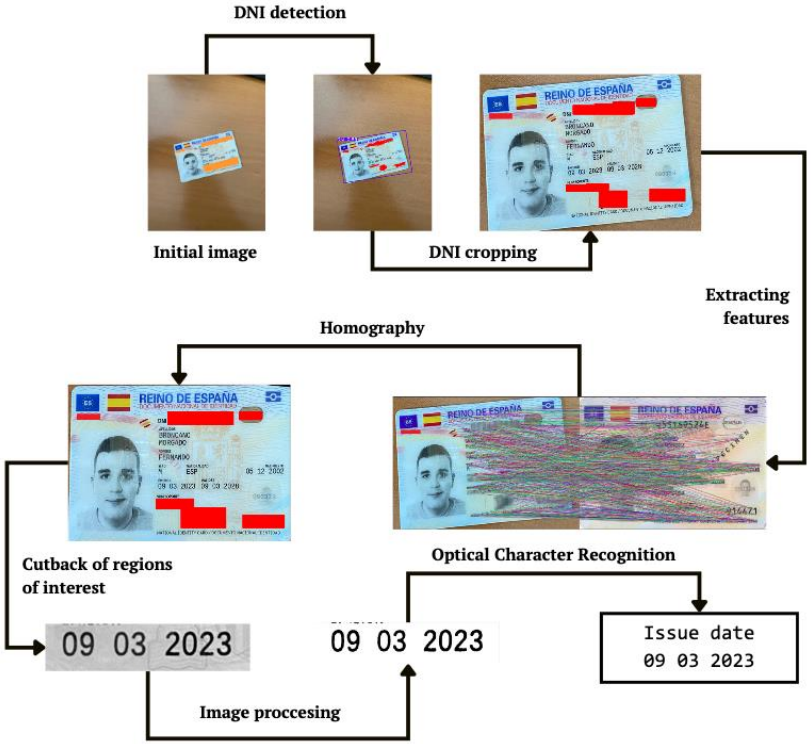
Visual Inspection Zone (VIZ)

Front

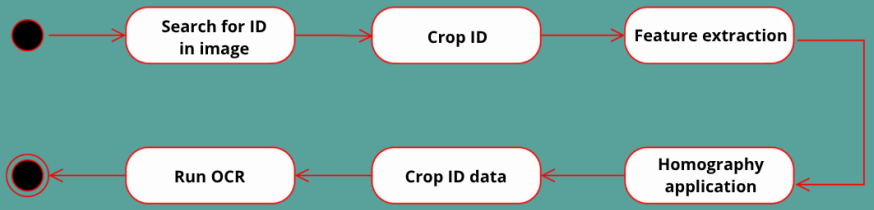
Reverse

Machine Readable Zone (MRZ)

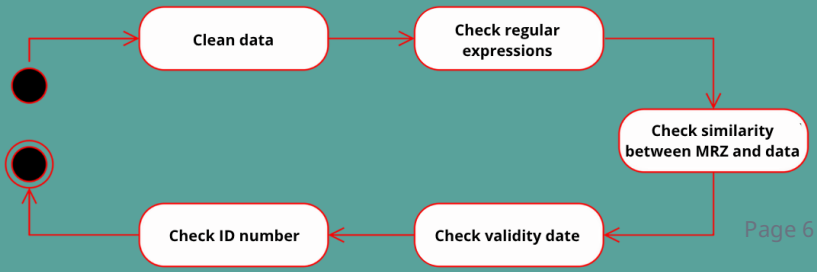
Methodology



ID Detection and Reading



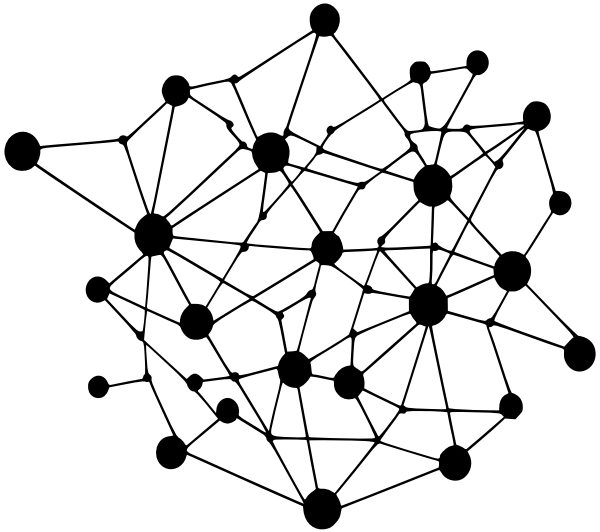
Data Normalization



Methodology

ID Processing Steps – ID detection

img

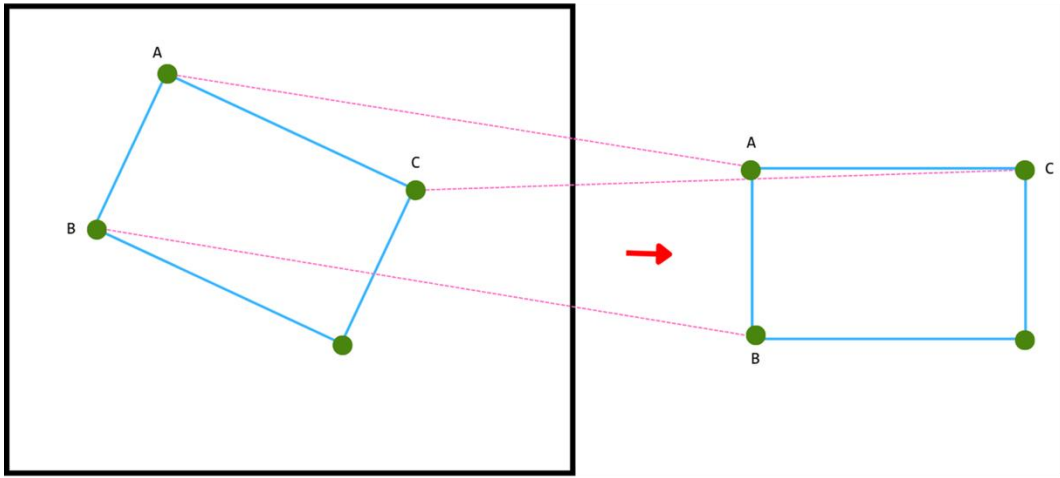


result = model(img)

result

Methodology

ID Processing Steps – Affine Cropping



result = cv2.warpAffine(img, matrix, A)

img

result

Methodology

ID Processing Steps – Homography / Perspective


```
matrix = cv2.findHomography(dni_p, temp_p)
result = cv2.warpPerspective(img, matrix, center)
```

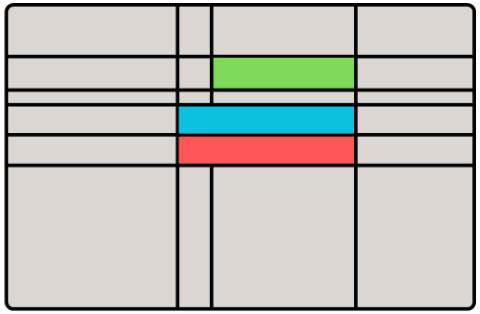
img

result

Methodology

ID Processing Steps – Cropping

Identification Element	Source Point	Target Point
ID Number	(0.35, 0.10)	(0.70, 0.20)
Surnames	(0.30, 0.25)	(0.70, 0.35)
Name	(0.30, 0.35)	(0.70, 0.45)



Date of Issue
09 03 2023

```
element_image = adapt_image(img, element)
element_text = pass_ocr(element_image)
```

img

template

Methodology

ID Processing Steps – Normalization

Character Substitution

- /
- |
- -
- ,
- .
- +
- "
- *
- ’
- |
- !
- ? -> 2
- O -> 0

Identification Element	Regular Expression
Dates	[0-9]{2} [0-9]{2} [0-9]{4}
ID Number	[0-9]{8}[A-HJ-NP-TV-Z]
Sex	[MF]
Nationality	[A-Z]{3}
Support Number	[A-Z]{3}[0-9]{6}
Card Access Number	[0-9]{6}
Issuing Authority / Device	(0[1-9]|[1-4][0-9]|5[0-2])[0-9]{3}[A-Z][0-9A-Z][A-Z][0-9A-Z]

Name	Date of Birth
PaBL0!	11 11 200!
PaBL0	11 11 200
PaBLO	11 11 200
PABLO	11 11 200

Methodology

ID Processing Steps – Truthfulness and Homogeneity

Do MRZ and VIZ match?

Are the dates correct?

Is the ID number correct?

Validity, Birth and Issue Dates

- 0 – 5 years: 2 years of validity
- 5 – 30 years: 5 years of validity
- 30 – 70 years: 10 years of validity
- In case of theft or loss, the validity may not match these timeframes

Control Character of the ID Number

Letter linked to the modulo 23 of the ID number

MRZ Check Digit

Calculated using the MRZ characters by assigning weights 7, 3, 1 to the numeric value of each character

Results

Set of Real ID Cards

Nine ID cards cropped to content
Three with low sharpness, five under optimal conditions

Detected Issues

Cropping precision varies with lighting conditions
OCR engine confuses letters and numbers

The overall accuracy rate reaches 90.47%

Recognized Region	Matches
Validity Date	14
Issue Date	14
Support Number	11
Card Access Number	11
First Parent	10
Providence of Residence	5
Address	5
ID Number	5
Issuing Autority	4
MRZ	1

The five least recognized categories are excluded.
ID cards with fewer than seven recognized elements are excluded.

Conclusions

- Reliable reading & alignment
- Challenges with low-resolution images
- High performance on good-quality images
- Effective homogeneity validation

Future Work

- Image –to-text recognition model
- Customized synthetic generation
- Deppen in truthfulness recognition
- Original image manipulation through impersonation

Protect people in both the physical and digital worlds