Algerian Democratic and Popular

Ministry of Higher Education and Scientific Research

University of Constantine 2 – AbdelHamidMehri

Department: TLSI

Theme: Car Number Plate Detection and Recognition System

Realized by: Acil Ykhlefoune
Wail Benoulha
Wissem Debbibi

Supervised by: Dr. Wafa Ghamaz
Dr. Naila Marir

Project Statement

Introduction

In the modern domain of traffic control and surveillance systems, the imperative of Automated Car Number Plate Detection (NPD) and Recognition cannot be emphasized enough. As the number of vehicles continues to rise, the capability to accurately detect and identify license plates via automated means has emerged as a cornerstone of efficient traffic management. This technology assumes paramount significance in facilitating streamlined toll collection processes, optimizing parking management, enforcing regulatory measures, and fortifying security protocols. Consequently, the development of precise and reliable Car Number Plate Detection and Recognition Systems using computer vision and deep learning techniques has become crucial to enhance various aspects of transportation and security.

Figure 1:

Problem Statement

In the world of traffic control and security, making Car Number Plate Detection and Recognition Systems work well depends on having high-quality data. This project focuses on preparing and enhancing the quality of data collected to improve the performance of machine learning models.

Approach

The project execution followed a structured iterative approach based on the principles of the Data Science Process involves several key steps:

- **Data collection:** We've compiled datasets utilizing images and corresponding text files available at the following links: https://github.com/LicensePlatesofAlgeriaDataset https://public.roboflow.com/object-detection/license-plates-us-eu
- **Exploratory Data Analysis (EDA):** Involves employing fundamental statistical techniques and visual representations to analyze the datasetby checking missing values, null values and duplicate , take a look at the number of rows and columns ,show show the first and last rows of the data.
- **Preprocessing:** Contains the following many steps like data cleaning ,enhacing the quality of images ,use the ocr model to extract the text from the plates and others .
- **Modeling:** Describes the methodology we employed to produce the best model for this dataset.
- **Prediction:** Employing learned patterns to make estimations on new data.

Data Collection

In the context of image processing, the first big step is collecting data , and it's super important For our project.
in our project we get the datasets from the links mentioned earlier,the first dataset is the "License Plates of Algeria" and the second is the "License Plates of the World".
These datasets are the starting point for our project, giving us the pictures and related text we need to train and make our Car Number Plate Detection and Recognition System better. These datasets are the starting point of our whole project. They're the big sets of pictures and extra details about license plates.

Exploratory Data Analysis (EDA)

IS an analysis approach that identifies general patterns in the data. EDA is an important first step in any data analysis. We employed several commands for this section, including df.info(), df.count(), df, df.mean(), and others Which is mentioned on our Jupiter.
Our current focus involves the analysis of two datasets: the "License Plates of the World" and the "License Plates of Algeria." the first dataset we combine multiple CSV files into a single DataFrame using the pandas library in Python.

The first dataaset has 1123 rows and 8 columns.the second dataset has 236 rows and 3 columns We observe that there are neither duplications nor null values.

Visualisation

Visualisation of world dataset

This chart shows us how many things, like cars and license plates, are usually in each picture in our dataset. It helps us see how crowded the pictures are with objects. This information is handy for figuring out how busy the dataset is in terms of how many things are in each picture.

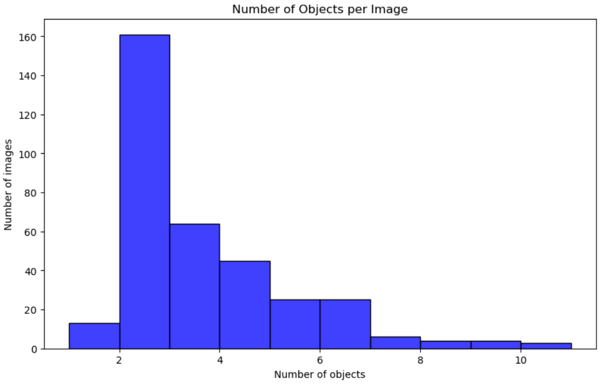


Figure 2: Number of Objects per Image

The bar chart we created gives a quick look at how often each category appears in our dataset. This helps us see if our dataset is balanced or if some categories are more common than others. We're focusing on license plates and treating vehicles as outliers since our main interest is in working with license plates.

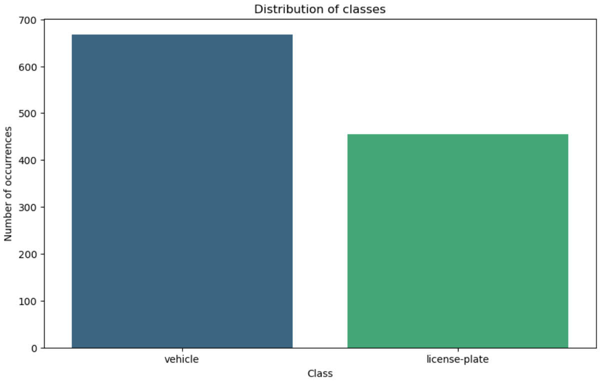


Figure 3: Distribution of classes

Visualisation of algerian dataset

In the Jupiter notebook, we have a specific format that license plates should follow. If the content of a plate matches this specific format, we say it's true, meaning it has the correct syntax. If it doesn't match the format, we say it's false, indicating it has an incorrect format. The plot illustrates which plates in our dataset have the correct syntax (true) and which ones don't (false)

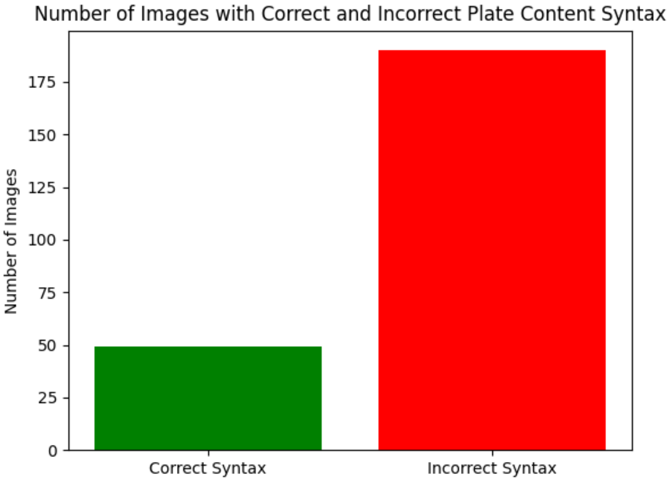


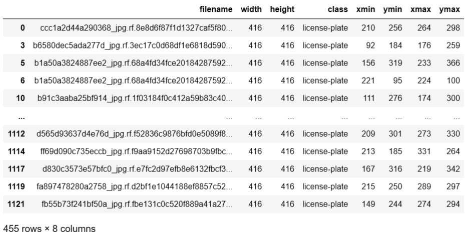
Figure 4: Numder of images with correct and incorrct plate content

Data Preprocessing

Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

In our project, we're treating vehicles as outliers because we're specifically interested in license plates. So, in this step, we want to simplify things by removing these outliers. What we end up with is a neat DataFrame that only keeps the rows where the category is marked as 'License-plate'. This helps us focus on what matters most for our project



	filename	width	height	class	xmin	ymin	xmax	ymax
0	ccc1a2d44a290368_jpg.rf.8e8d6f87f1d1327caf5f80...	416	416	license-plate	210	256	264	298
3	b6580dec5ada277d_jpg.rf.3ec17c0d68df1e6818d590...	416	416	license-plate	92	184	176	259
5	b1a50a3824887ee2_jpg.rf.68a4fd34fce20184287592...	416	416	license-plate	156	319	233	366
6	b1a50a3824887ee2_jpg.rf.68a4fd34fce20184287592...	416	416	license-plate	221	95	224	100
10	b91c3aaba25bf914_jpg.rf.1f03184f0c412a59b83c40...	416	416	license-plate	111	276	174	300
...	...	...	...	...	...	...	...	...
1112	d565d93637d4e76d_jpg.rf.f52836c9876bfd0e5089f8...	416	416	license-plate	209	301	273	330
1114	ff69d090c735eccb_jpg.rf.f9aa9152d27698703b9fbc...	416	416	license-plate	213	185	331	264
1117	d830c3573e57bfc0_jpg.rf.e7fc2d97efb8e6132fbcf3...	416	416	license-plate	167	316	219	342
1119	fa897478280a2758_jpg.rf.d2bf1e1044188ef8857c52...	416	416	license-plate	215	250	289	297
1121	fb55b73f241bf50a_jpg.rf.fbe131c0c520f889a41a27...	416	416	license-plate	149	244	274	294

455 rows × 8 columns

Figure 5: Dataset after removing outliers

Enhacing quality of images

We aim to enhance image quality by transforming them into red, blue, green, and black-and-white color channels. However, based on the Kernel Density Estimate for Color Channels, it was observed that black and white gives superior results like the figure 6 shows.
So we improved the quality of images by changing them to black and white . This helps make the images clearer and highlights important details, all to assist in reading the license plates more effectively for our Car Number Plate Detection and Recognition System.the function how we convert the images to black and white is mentioned on our Jupiter

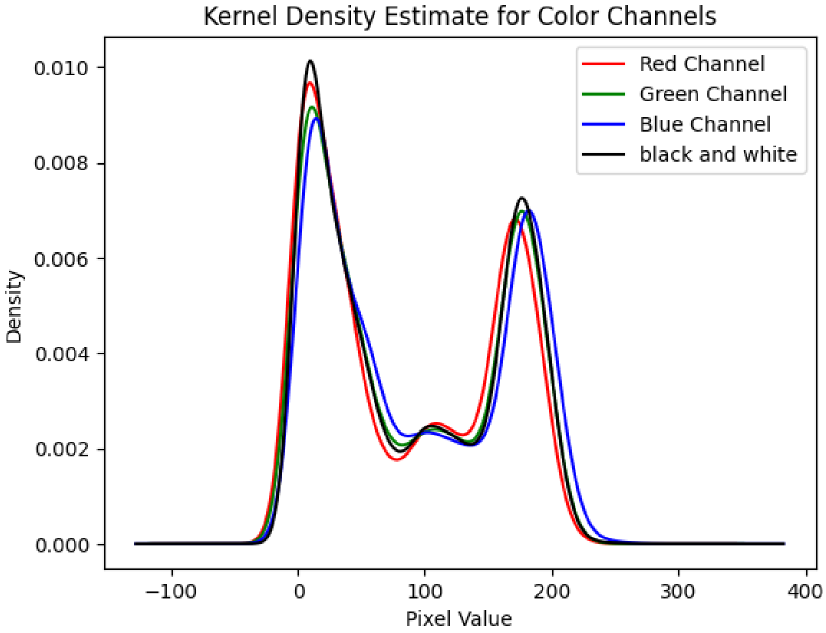
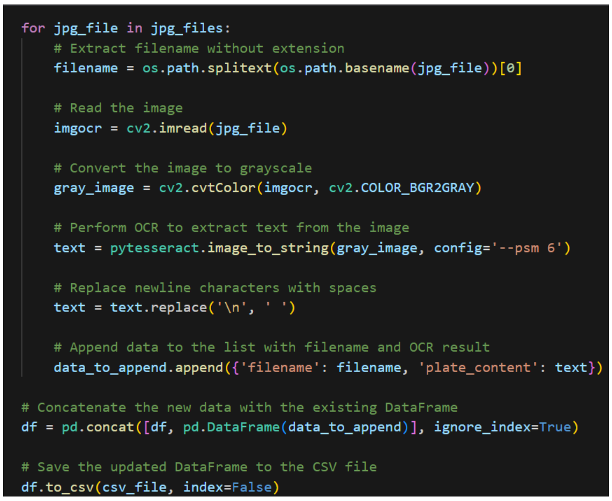


Figure 6: Kernel Density Estimate for Color Channels

Apply the OCR model

”OCR (Optical Character Recognition) is a technology that extracts text from images, is like a smart tool for computers that reads text from pictures”. after we enhace the quality of images we apply the OCR model to extract the content of the plates .

We employ two libraries, OpenCV and PyTesseract, in our project. OpenCV is utilized for image processing tasks, while PyTesseract is employed for optical character recognition (OCR), allowing us to extract text from images.the steps are montioned in the code below .



```
for jpg_file in jpg_files:
    # Extract filename without extension
    filename = os.path.splitext(os.path.basename(jpg_file))[0]

    # Read the image
    imgocr = cv2.imread(jpg_file)

    # Convert the image to grayscale
    gray_image = cv2.cvtColor(imgocr, cv2.COLOR_BGR2GRAY)

    # Perform OCR to extract text from the image
    text = pytesseract.image_to_string(gray_image, config='--psm 6')

    # Replace newline characters with spaces
    text = text.replace('\n', ' ')

    # Append data to the list with filename and OCR result
    data_to_append.append({'filename': filename, 'plate_content': text})

# Concatenate the new data with the existing DataFrame
df = pd.concat([df, pd.DataFrame(data_to_append)], ignore_index=True)

# Save the updated DataFrame to the CSV file
df.to_csv(csv_file, index=False)
```

Figure 7: code of applying ocr model

Scaling

Data scaling is the process of transforming the values of the features of a dataset till they are within a specific range.

In image preprocessing, scaling involves adjusting the size or dimensions of an image. It's a technique used to standardize or resize images to a consistent format, doing this helps us compare and work with the pictures more easily.

Conclusion

In conclusion, our project aimed to develop an efficient Car Number Plate Detection and Recognition System, focusing on the Algerian and world datasets. Through careful data collection, preprocessing, and advanced techniques such as optical character recognition (OCR), we successfully built a system capable of accurately identifying license plates. By addressing challenges like outliers, enhancing image quality, and utilizing OCR .This project highlights the impactful use of technology in solving real-world issue.

References

Siddhartha Choubey Associate Professor (Dept. of CSE) Shri Shankaracharya College of Engg. Tech., Bhilai C.G..India.

https://github.com/lmiguelmh/car-license-plate-preprocessing?fbclid=IwAR2$_LU7rbAaZ_GgqUh99I4BANUz_{KpTD5t8Alwv7VYubx9Gs_gKRYYnswY}$

link to our portfolio

https://sites.google.com/d/1U8-A2DlEIfMzHXteNwK9mWdEyYAY8Zhv/p/18fL1F33Wnm-I8rqlvvsvTYsAcpNmwbYa/edit