Arabic Font Recognition System

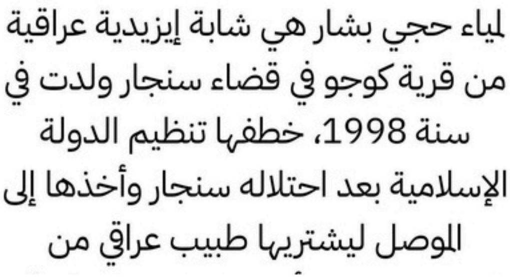
Spring 2024

Overview

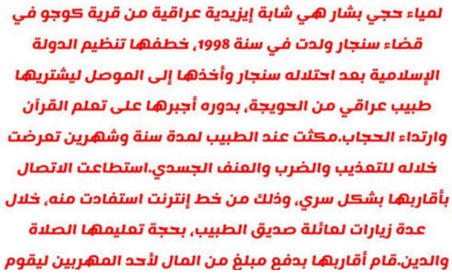
In this project, we implemented an Arabic Font Recognition System.
Given an image containing a paragraph written in Arabic, the system can classify the paragraph into one of four fonts.

Name	Font Code
Scheherazade New	0
Marhey	1
Lemonada	2
IBM Plex Sans Arabic	3

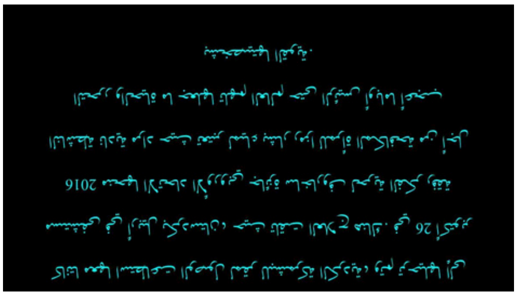
Our system can handle variations like Text size, Image blur, Brightness variation, Text color, Text rotation, Salt and pepper noise, like the images below

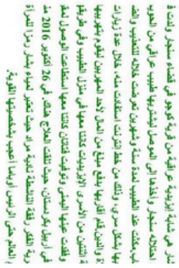


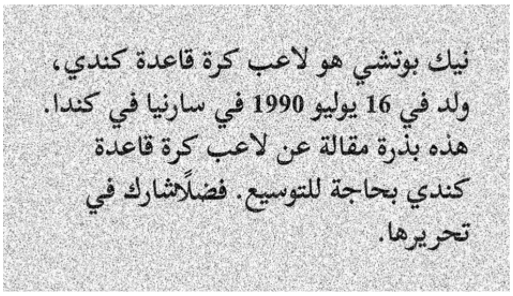
لياء حجي بشار هي شابة إيزيدية عراقية
من قرية كوجو في قضاء سنجار ولدت في
سنة 1998، خطفها تنظيم الدولة
الإسلامية بعد احتلاله سنجار وأخذها إلى
الموصل ليشتريها طبيب عراقي من



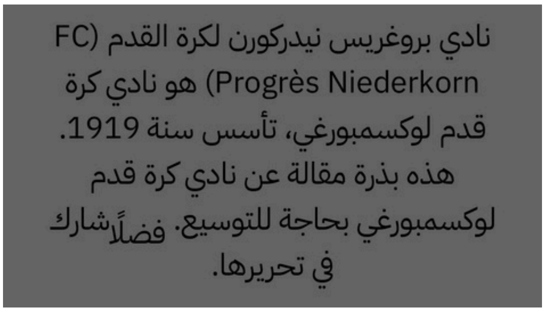
لمياء حجي بشار هي شابة إيزيدية عراقية من قرية كوجو في
قضاء سنجار ولدت في سنة 1998، خطفها تنظيم الدولة
الإسلامية بعد احتلاله سنجار وأخذها إلى الموصل ليشتريها
طبيب عراقي من الحويجة، بدوره أجبرها على تعلم القرآن
وارتداء الحجاب.مكثت عند الطبيب لمدة سنة وشهرين تعرضت
خلاله للتعذيب والضرب والعنف الجسدي.استطاعت الاتصال
بأقاربها بشكل سري، وذلك من خط إنترنت استفادت منه، خلال
عدة زيارات لعائلة صديق الطبيب، بحجة تعليمها الصلاة
والدين.قام أقاربها بدفع مبلغ من المال لأحد المهربين ليقوم







نيك بوتشي هو لاعب كرة قاعدة كندي،
ولد في 16 يوليو 1990 في سارنيا في كندا.
هذه بذرة مقالة عن لاعب كرة قاعدة
كندي بحاجة للتوسيع. فضلاًشارك في
تحريرها.



نادي بروغريس نيدركورن لكرة القدم FC)
(Progrès Niederkorn هو نادي كرة
قدم لوكسمبورغي، تأسس سنة 1919.
هذه بذرة مقالة عن نادي كرة قدم
لوكسمبورغي بحاجة للتوسيع. فضلاًشارك
في تحريرها.

Project Pipeline

Reading data

- This module is designed to read image data from a specified folder path and organize it along with corresponding labels into provided lists.

Preprocessing

- **salt_paper(image):**
  - This function is used to remove salt-and-pepper noise from the input image using median filter.

- **text_binary(image):**
  - This function applies Gaussian blur to smooth the image and then performs Otsu's thresholding to create a binary image. If specified, it checks whether the white pixel count is greater than the black pixel count. If true, it inverts the binary image. Finally, it applies Gaussian blur again to further smooth the binary image and returns it.

- **text_rotation(image):**
  - This function corrects the rotation of text in the input image. It applies Canny edge detection to detect edges in the image. Then, it performs morphological closing to close any gaps in the edges. Using the Hough transform, it detects the base lines in the image. It calculates the median angle of these lines and rotates the image by that angle to correct the text orientation.

- **preprocess(data):**
  - This function applies a preprocessing pipeline to a list of input images, typically used before feeding them into a text recognition model. It iterates over each image in the input list, then applies salt-and-pepper noise removal, binarization and rotation correction using the previously mentioned functions

Feature Extraction

- **adaptive_line_segmentation(preprocessed_image):**
    - This function performs adaptive line segmentation on the preprocessed image. It computes the histogram of pixel intensities along the vertical axis. Based on the histogram, it determines local thresholds to identify potential line boundaries. Zero crossings in the histogram below the global threshold indicate line boundaries. Lines are extracted based on identified boundaries, and a list of line images is returned.

- **extract_lpq_feature(image):**
    - This function computes Local Phase Quantization (LPQ) features from the input image (line). It convolves the image with sinusoidal filters and calculates the frequency response. The responses are used to compute the LPQ descriptors. Finally, it returns the normalized histogram of the LPQ descriptors.

- **extract_angles_feature(image):**
    - This function extracts angles feature from the input image using the Hough Line Transform. It applies Canny edge detection to the image and detects lines using HoughLinesP. It calculates the angle of each detected line and filters lines based on angle and length criteria. The mean angle of filtered lines is computed as the feature.

- **feature_extraction(font, whole_image, data, labels):**
    - This function performs the feature extraction process for text lines within the whole image. It first performs adaptive line segmentation to extract individual text lines. For each text line, it extracts LPQ features and angle features using the previously defined functions. If angle features are available, they are concatenated with LPQ features. The concatenated feature vector is appended to the data list, and the font label is appended to the labels list.

Model Training

- **train_data(data, datalabels):**
    - This function is used to train a Random Forest classifier using the provided training data and their corresponding labels where each point corresponds to a line in image. It initializes a Random Forest classifier using Random Forest Classifier from scikit-learn. After training, the trained classifier is saved to a pkl file.

Model Testing

- **test_data(test, testlabels):**
    - This function is used to test the trained Random Forest classifier using the provided test data and their corresponding labels. It loads the trained classifier from the pkl file then use it to predict labels for the test data. Then it calculates the majority class of the image lines and returns it.

Performance

Accuracy

- The accuracy achieved using random sample of training images and test images was:
    - 97.5% to 99.3% for 400 training images and 40 test images for each font
    - 98.5% to 99.4% for 700 training images and 200 test images for each font

Time

- Time taken for was:
    - 9.8 mins for reading data and model training of 700 training images for each font
    - 3.2 min for reading data and model testing of 200 test images for each font

Future Work

Add more fonts

- Diversifying the dataset with additional fonts to boost the model's ability to handle different text styles. We are planning to collect samples from various fonts, including serif, sans-serif and decorative styles, to ensure model robustness.

Improve Time

- Enhancing the pipeline's speed is essential for real-time applications. We are planning to optimize algorithms, using parallel processing and reducing feature complexity to speed up both training and inference.

Work Load

Name	ID	Task
Menna Mohammed AbdelBaset	9211242	LPQ feature & Text Binarization & Server
Rawan Mostafa Mahmoud	9210423	Model Training & Testing
Fatma Ebrahim Sobhy	9210799	Angle feature & Text Rotation & Noise Removal
Sara Bisheer Fikry	9210453	Image Segmentation