Cairo University
Faculty of Engineering
Computer Engineering Department

Pattern Recognition & Neural Networks
CMP(N)450

OCR for Arabic Scripts

Fall 2019

Document Structure

- Objectives
- Problem Description
- Dataset
- Team Formation
- Project Schedule
- Milestones' Deliverables
- Final Deliverables
- Delivery Details
- Grading Criteria
- FAQ

Objectives

This Project aims to put your understanding of machine learning algorithms into practice with a real world problem. By the end of this project you should be able to:

- analyze the problem, extract features and choose the most appropriate preprocessing method (if necessary).
- assess performance of different machine learning techniques.
- design and implement your own ML pipeline.

Problem Description

Optical character recognition or optical character reader (OCR) is the recognition process of text obtained from media in the form of typed, handwritten or printed text into machine-encoded text form. The text in question may be presented in the form of a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image.

The current accuracy almost hit 100% for English language text. However, for Arabic scripts the results are still not that high. In the context of this project, we consider only scanned typed documents, i.e., not handwritten, in order to simplify the problem. The aim of this project is to design and train a model that is able to read images of scanned Arabic documents and recognize their text. The system should be able to identify the Arabic characters written in the text and finally generate the text written in those images.
You should implement a complete Machine Learning pipeline, i.e., the project should include (but not limited to) the following modules:

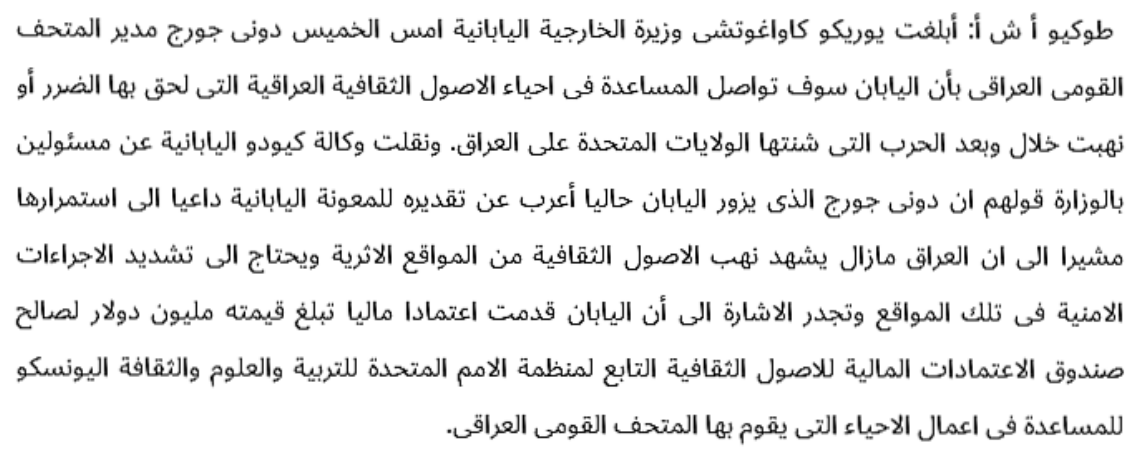
- preprocessing module
- feature extraction/selection module
- model selection and training module
- performance analysis module

Accordingly, you should review the literature about the topic, read scientific papers that tackle this problem and closely-related problems, and do a literature survey that can help you identify the best approaches/techniques that you can start with in order to improve the accuracy of your results.

Dataset

A dataset of images and its ground truth text is available for you here.

Sample Input (Image)


طوكيو أ ش أ: أبلغت يوريكو كاواغوتشى وزيرة الخارجية اليابانية امس الخميس دونى جورج مدير المتحف القومى العراقى بأن اليابان سوف تواصل المساعدة فى احياء الاصول الثقافية العراقية التى لحق بها الضرر أو نهبت خلال وبعد الحرب التى شنتها الولايات المتحدة على العراق. ونقلت وكالة كيودو اليابانية عن مسئولين بالوزارة قولهم ان دونى جورج الذى يزور اليابان حاليا أعرب عن تقديره للمعونة اليابانية داعيا الى استمرارها مشيرا الى ان العراق مازال يشهد نهب الاصول الثقافية من المواقع الاثرية ويحتاج الى تشديد الاجراءات الامنية فى تلك المواقع وتجدر الاشارة الى أن اليابان قدمت اعتمادا ماليا تبلغ قيمته مليون دولار لصالح صندوق الاعتمادات المالية للاصول الثقافية التابع لمنظمة الامم المتحدة للتربية والعلوم والثقافة اليونسكو للمساعدة فى اعمال الاحياء التى يقوم بها المتحف القومى العراقى.

Sample Output (Text)

طوكيو أ ش أ: أبلغت يوريكو كاواغوتشى وزيرة الخارجية اليابانية امس الخميس دونى جورج مدير المتحف القومى العراقى بأن اليابان سوف تواصل المساعدة فى احياء الاصول الثقافية العراقية التى لحق بها الضرر أو نهبت خلال وبعد الحرب التى شنتها الولايات المتحدة على العراق. ونقلت وكالة كيودو اليابانية عن مسئولين بالوزارة قولهم ان دونى جورج الذى يزور اليابان حاليا أعرب عن تقديره للمعونة اليابانية داعيا الى استمرارها مشيرا الى ان العراق مازال يشهد نهب الاصول الثقافية من المواقع الاثرية ويحتاج الى تشديد الاجراءات الامنية فى تلك المواقع وتجدر الاشارة الى أن اليابان قدمت اعتمادا ماليا تبلغ قيمته مليون دولار لصالح صندوق الاعتمادات المالية للاصول الثقافية التابع لمنظمة الامم المتحدة للتربية والعلوم والثقافة اليونسكو للمساعدة فى اعمال الاحياء التى يقوم بها المتحف القومى العراقى.

Team Formation

A team should be formed of **3-4 members**. Please fill in your team number in front of your name in this Google sheet. All team members should attend the project discussion, so if your team is divided between the two sections, it is your responsibility to adjust your schedule to attend discussion.

Project Schedule

Team formation Tuesday, October 29th, 11:59 PM.

Milestone 1 Saturday, November 16th, 11:59 PM.

Milestone 2 Saturday, November 30th, 11:59 PM.

Final Delivery Sunday, December 15th, 11:59 PM.

Discussion the following tutorial time.

Milestones' Deliverables

A report containing

- workload distribution, team number, and team members' names as well.
- summary of implemented preprocessing/post methods and classifiers.
- summary of the feature extraction methods used.
- summary of the other candidate methods you plan to try.
- accuracy and run-time you got so far with a summary of your experiments' setup.
- estimated task runtime (this is not for grading, but rather serves as a feedback).

Final Deliverables

1. **Report** PDF that includes the following sections fully and clearly described, while describing all the work done and approaches adopted. You should include also the unsuccessful trials.

   i. Project Pipeline.
   ii. Preprocessing Module.
   iii. Feature Extraction/Selection Module.
   iv. Model Selection and Training Modules.
   v. Post-processing Module. (if exists)
   vi. Performance Analysis Module.
   vii. *(Optional)* Any other developed modules.
   viii. Enhancements and Future work.

   The report should include also a workload distribution between team members.

2. **Code** zipped folder with the format *code_[team number].zip* containing all code developed with a *readme* file including all packages or libraries needed to run this code and how to run the code. It is highly recommended to provide an executable file beside the source code.

3. **README** text file containing the names and IDs/BNs of the team members.

Delivery Details

1. Deliverables should be sent by **email** to *submissions.pattern@gmail.com* using the subject:
   (a) *[Credit – Pattern Recognition][Team Number][Milestone Number or Final]* for credit hours system's students.
   (b) *[Semester – Pattern Recognition][Team Number][Milestone Number or Final]* for mainstream, i.e., semester students.
2. Don't print any document or submit the project on a CD. **All submissions are electronic.**
3. There will be a **late penalty** for submissions in any of the two mentioned phases.
4. Any sign of **cheating or plagiarism** will not be tolerated. If one team got caught cheating or plagiarizing from another team, both teams will receive a **ZERO** for the project grade, in addition to a penalty up to 50% of the project grade.
5. **Workload should be distributed fairly and equally.** Team members who did not contribute effectively in the project will be penalized.
6. **All team members** should attend the final discussion. A team member who fails to show up will get a **ZERO** in the discussion grade.

Grading Criteria

Report and Discussion: 20% This point includes the report submission, the quality of the report and the discussion with all team members. It also evaluates how all team members fully understand the details of the project and can elaborate on the examiner questions.

Results Accuracy: 80% Both results **accuracy (65%)** and **running time (15%)** will be taken into consideration by the given weights. However, it's not a linear formula but the ranking procedure will be based only on these two factors. *Take care that your language choice might affect the running time.*

FAQ

1. **What programming language(s) can I use?**

   You may use any programming language of your choice!

2. **Is there any restriction on the techniques or approaches to use in any phase of the project?**

   You are free to use any approach or technique you find appropriate for the problem in hand. It's actually your task to find out the best combination of techniques that will yield the best results. However, **you are limited only to classical machine learning methods** (Bayesian Classifiers, K-NN, Linear/Logistic Regression, Neural Networks, Support Vector Machines, Principal Component Analysis, etc.). You are **NOT ALLOWED** to use deep learning techniques.

3. **How to find research papers for this topic?**

   There are many resources on the web tackling this problem. You can start by creating an account on EKB (Egyptian Knowledge Bank) with your university-provided student email address, which will give you wide access to a huge number of research papers and journals.

   You can also use the Academic search engines like (Microsoft Academic or Google Scholar) a lot of articles are free for public.

4. **How will our project be tested?**

   We will test with our own test set in order to standardize the way all teams are graded and ranked. However, you should divide your dataset into Training, Validating and Testing set.

5. **How will our project be ranked?**

   Please refer to the "Grading" section for the grading criteria. The teams will be ranked according to some formula comprising the results accuracy as well as the time taken to generate the results, with the larger weight for the results accuracy. For example, Team X had results with accuracy 92.0%, with running time = 5 seconds, will be ranked above team with accuracy 80.0%, with running time = 2 seconds.