Manuscript Recognition and Detection project

Graduation project 2022

Faculty of computers & artificial intelligence - Beni Suef university

Overview

In the early centuries and before the existence of modern machines such as the typewriter or the computer, people used to write their words on paper by hand, and since science is successive pieces and each generation completes what the previous generation started, we always need to know those writings that were written in the past.

Many researchers in the field of religious sciences, Arabic language, heritage, and Arab history deal directly with manuscripts to know the heritage and know the history of science, as well as pay attention to what previous people left and try to understand, explain and increase.

The graduation project always aims to help people and save time, effort, and money, so the idea that we wanted to work on was to try to reach a software way to identify and discover these manuscripts and help researchers by providing this software solution.

Keywords and Abbreviations

- **DL:** Deep learning
- **ANN**: Artificial neural network

Section

Technically, a manuscript is written by hand, and the American Encyclopedia defines a manuscript as being written by hand in any kind of literature, whether on paper or on any other material such as leather, ancient clay tablets, stones, and others.

Arabic manuscripts are considered the oldest human intellectual heritage that arrived safely in this era. These manuscripts were so numerous that they exceeded in number and variety of subjects any world intellectual heritage, and the number of Arabic manuscripts reached nearly three million manuscripts scattered in the libraries of the Arab and Islamic world and other parts of the world.

These Arabic and Islamic manuscripts are considered a human heritage and a global repository, whether they are in the sciences, literature, or general knowledge. Manuscripts were not limited to a specific type of science. Rather, it extended to the physical sciences, optics, mathematical, engineering, and medical sciences. The Arabic and Islamic manuscripts have gone beyond these topics to geography, travels, geographical discoveries, poetry, arts, especially music.

The oldest manuscripts in the world date back to (3500) years of manufacture and were rolls of papyrus.

As for the parchment sizes, they were not as fixed as in thousands of printed books, and were sometimes not equal to the sizes of the sheets of a single manuscript. There are two manuscript sizes (18 cm x 12 cm) and (25 cm x 18 cm).

The scribes who transcribe Arabic manuscripts often use the ruq'ah script and the Naskh script in the Naskh script, and the titles are in Kufic script.

There are two types of manuscripts that differ among themselves in the form of drawing letters and words:

- The first type is the oriental manuscripts: that is, those written by the people of the East from the Arab world and have a certain drawing and a certain way of the letters
- The second type is the Western manuscripts: that is, those written by the people of the East from the Arab world and have a specific drawing and a certain way of letters that differ from those that were written in the Arab Mashreq.

Goals

Our goal is to build a system that is able to use real images of Arabic manuscripts and work on identifying, recognizing, and detecting them in a programmatic way, as well as addressing problems that may be encountered in the image such as lack of clarity or worn-out paper and others.

Problem statement

Many researchers have problems identifying words in these manuscripts, and these problems may be time, financial, and physical health costs. Therefore, when we looked at the problems faced by these researchers that make them spend all this effort and exhaustion in interpreting and understanding the manuscripts, we found that the manuscripts contain a number of problems that make the task difficult for these researchers, including:

- handwritten
- There are marginal notes
- The lines are not straight
- Font sizes are different
- Different types of fonts
- faded ink
- decorations
- The presence of non-rectangular areas
- Paper wear problems because of its age
- .....

In this project, we will try to find software solutions to all these problems

Solution statement

The solution we offer is in the use of artificial intelligence techniques, machine learning, image analysis and processing, and computer vision to find a software solution to the problems we previously presented, by creating a mobile phone application and a website that the researcher can use to upload the manuscript's image and the system works to

identify it, identify the words in it, address the existing problems, and extract The words contained in this image are from the manuscript in computer typing form.

Specifications

Suggestion Approaches

Our goal is to build a system that is able to use real images of manuscripts and work on identifying, recognizing, and detecting them in a programmatic way, as well as addressing problems that may be encountered in the image such as lack of clarity or worn-out paper and others.

Datasets

From the research on the idea, we did not find a clear data set dedicated to Arabic manuscripts in particular, but we came up with some solutions to find data sets that we can use in the project. There are three ways so far, and perhaps in the course of the research in the future there will be more than one, and the three ways are:

- The First way:
  Web scraping from Qatar Digital Library (QDL):
  https://ediscovery.qnl.qa/ar/islandora/object/QNL%3AMANUSCRIPTS

- The Second way:
  KERTAS dataset
  We found some scientific papers talking about it but unfortunately we couldn't get it (Link to the paper you mentioned is attached)
  https://link.springer.com/article/10.1007/s10032-018-0312-3

- The Third way:
  Request some pictures of some manuscripts from well-known governmental cultural places such as the Egyptian National Library and Documentation House, the Al-Azhar Library, the Library of Alexandria and other places that contain a large number of rare and useful manuscripts

Project steps

Linkes

https://ar.wikisource.org/wiki/%D9%85%D8%B3%D8%AA%D8%AE%D8%AF%D9%85:%D8%B9%D8%A8%D9%8A%D8%AF/%D9%85%D8%AE%D8%B7%D9%88%D8%B7%D8%A7%D8%AA

https://wadod.net/bookshelf/category/2
http://alnadeem-bks.malecso.org/cgi-bin/koha/opac-main.pl
https://alkitabdar.com/manuscripts/

Timeline and phases

References

Project Team

- Mohamed Abdelrahman
- Mohamed Gamal
- Ola Abdallah
- Nourhan Mahmoud
- Shaimaa Mostafa