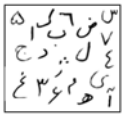

Call for participation
Handwritten Farsi/Arabic Character Recognition Competition

Motivation:

Recently many research groups around the world focused on Arabic document analysis and promising results have been reported. However, lack of communication among them caused wasteful duplication of efforts. The aim of ICDAR2009 Handwritten Farsi/Arabic Character Recognition Competition is to bring together researchers working on this field. By benchmarking the state of the art Arabic character recognition techniques on large-scale dataset, a comparative result can be obtained. Furthermore, the result of this competition would have widespread benefits to other languages such as Farsi (Persian) and Urdu which have the same characters set.

Data Sets:

Unlike previous efforts in which each group collected small sets of data from limited number of writers and implemented their recognizer on them, the ultimate goal of this competition is to evaluate how new OCR techniques deal with large databases (at least 1000 samples for each class).

In this experiment, training and test datasets are selected from large datasets that some subsets of them were published in ICFHR2006 conference (Saeed Mozaffari and et al. “A comprehensive Isolated Farsi/Arabic database for OCR research” pp.385-389).

Some samples are available in the competition homepage (http://icdar2009.kumesh.ir).

Database salient features are as follows:

1- Data was gathered from numerous writers.
2- Each character is represented by a BMP image file.
3- Images have different size.
4- Images can be binary or grayscale.
5- There are 12 classes for digits and 34 classes for characters.
6- Distributions of samples in these classes are not uniform.
7- Train and Test sets are independent.
8- Each train sample’s name starts with its class index (2 digits) and follows with 4 digits indicating its number in that class. (example: *03*1123.bmp)

Evaluation Process:

Due to similarity between some digits and characters in Farsi and Arabic, each recognizer will be tested for digit and character sets separately. Therefore, each team can take apart in digit or character recognition competition or both of them.

A recognizer may return up to 10 candidates for each classification based on a priority oriented list. These ranked results can also be used for comparison.

Two different strategies will be considered in this completion:

1- In the first experiment no reject is allowed.

2- In the second case, a recognizer can have reject strategy. Here, the recognition rate in respect to the reject rate will be compared.

Recognition time is another criterion in this competition.

Recognizer Running Format:

To have a standard approach, we will run each recognizer (called *myrec* here) by calling it from the command line as follows: "myrec address output.txt"

- *myrec*: is the name of your recognizer.
- *address*: is the test set images address. For example (F:\ICDAR 2009\Competition\Test set).
- *output.txt*: is the name of a file shows recognition results. This file should be created in the same folder indicated by *address* field. *output.txt* should have one line for each image file that was recognized. Each line must starts with image's name, followed by recognized class index. Each class index should have a confidence value. The following example shows for image tst12345b2.bmp, recognizer has produced three hypothesis classes: 02,03 and 12 with confidences of 1.0, 0.8 and 0.4. Example: tst12345b2.bmp 03 1.0 02 0.8 12 0.4.

Important date:

All teams should upload their recognizer due to: **1 February 2009**

Organizer:

Saeed Mozaffari and Hadi Soltanizadeh
Electrical and Computer Engineering Department
Semnan University
Semnan, Iran.

Contact Information:

- Email: saeed_mozaffari@ yahoo.com (mozaffari@semnan.ac.ir)
  h_soltanizadeh@kumesh.ir
- Home page: http:icdar2009.kumesh.ir