UC Berkeley, Computer Science

CS289A Project Proposal

Peter Cheng, Jeff Tsui, Alice Wang

April 18, 2013

1 Overview

Our project will be to design and test an off-line classifier for gender prediction from handwritten text. The inspiration for this project is from a recent machine learning competition hosted on Kaggle, a website which often runs such contests[1]. Kaggle provides sample training and testing data, in the form of high-resolution jpg images. Each image corresponds to a writing sample, and there are 4 writing samples for each of 475 writers. The 4 samples correspond to:

1. Arabic text, different text for each writer
2. Arabic text, same text for each writer
3. English text, different text for each writer
4. English text, same text for each writer

Kaggle also provides extracted features for each writing sample, using methods described in [2]. An example of an image sample is shown in Figure 2.1.

2 Approach

The Kaggle competition page provides references to a number of relevant papers for extracting features from writing samples and constructing classifiers. From a preliminary scan through related literature, it appears that acquiring strong features is as important, if not more important than the construction of models for classification. We have also

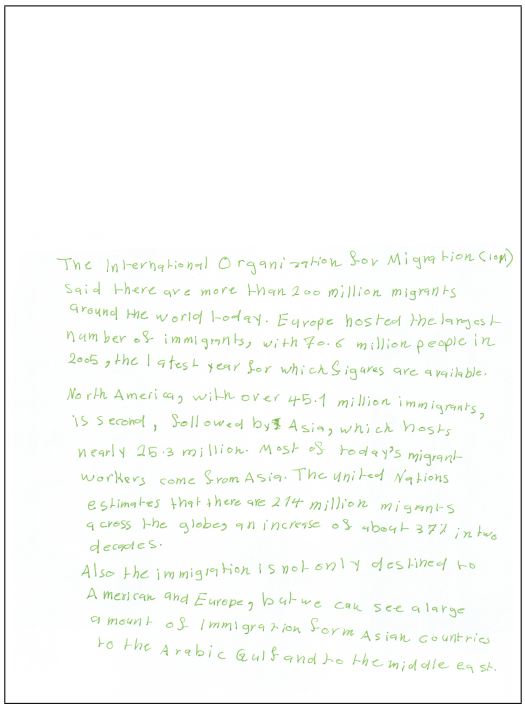

Figure 2.1: A sample training image

come across numerous papers describing different types of features, and the design of such features appears to be a very interesting problem. As a result, we are currently opting to generate features ourselves, and have begun implementing and testing methods encountered during our literature search. We also will have to perform a number of preprocessing steps, as the Kaggle dataset is highly unconstrained. The images contain color and some amount of noise, the handwriting is not centered, and much of the text is written at a curve or a slant. As a result, segmenting and extracting lines, words, and characters from each image will be a challenge in itself, and this needs to be performed before a large number of features can be acquired. Once we have successfully extracted features from our images, we should be able to implement the numerous classifiers we have learned from class, and benchmark our results against those obtained by the sources we are borrowing from.

References

[1] "Icdar2013 - gender prediction from handwriting." `http://www.kaggle.com/c/icdar2013-gender-prediction-from-handwriting`.

[2] HassaÃŕne, A., A.-M. S. and Bouridane, A., "A set of geometrical features for writer identification. neural information processing," in [*Neural Information Processing*], (2012).