Communication is an essential part of relating to people, and one of the oldest and most accessible methods of communication within a given language is writing by hand. In spite of the major effort that has been expended to bring about a paper-free society, a very large number of paper-based documents are processed daily by computers all over the world in order to handle, retrieve, and store information. The problem is that the manual process used to enter the data from these documents into computers demands a great deal of time and money (Bortolozzi, de Souza Britto Jr., Oliveira and Morita, n.d.). These documents may need to be processed for a number of reasons , among them historical documentation (e.g. digitally documenting culturally and historically significant documents and scripts, which until recently were more often than not handwritten or on print paper), recognition for medical prescriptions, or for tablet soft-wares to convert user’s handwriting into digital text.

Thus, the task of handwriting recognition is the transcription of handwritten data into a digital format, and this task obviously benefits from data compression. The goal is to process handwritten data electronically with the same or nearly the same accuracy as humans (Gunter, n.d).

Basically, handwriting can be divided into two categories, cursive script and printed hand- writing. Accuracy is the main problem in handwriting recognition for both categories because of the similar strokes and shapes some letters may possess. The software may have an inaccurate recognition of the letter, considering the possibility of the handwriting being illegible or some other factors (). One notable problem that makes this task difficult especially in cursive hand- writing recognition is the fact that there may be no obvious character boundaries (the start and end of a character); compared to printed handwriting, it does not have gaps or spaces between each letter to know the start and stop of recognition per character (). This issue is compounded for languages like Arabic, where cursive is the only form of script and there exist ”shortcuts” to further simplify the cursive script and make writing more fluid (e.g. removing the ”dots”/accent marks that exist above/below certain letters).

9

Layan:

put in cita- tion later

Kai:

Here you write that these docu- ments may need to be pro- cessed for a num- ber of rea- sons, but go on

to ex-

This is where the data compression/dimension reduction and eigenface technique comes into play! In essence, if we have a large data set that consists of thousands or even millions of images of words (probably limited to one language), we want to find a way to recognize patterns in these images. From these patterns, we can then determine which ones have the most “importance” and attempt to express these images as a weighted combination of these most important patterns (and thus in a lower dimension).