Automatic Text Recognition

Character recognition is one of the oldest fields of research since the advent of computers. However it is still an open field for researchers due to the challenging nature of perception and recognition. Human understanding of shapes and patterns is associated with a huge amount of information that we gather during our interaction with these patterns or shapes. A human can recognize a mug or its picture quite easily from any angle it will see it. We can read different fonts and styles even when characters are broken or overlapping. It is a difficult task to automatically mimic the perception of the human eye. Nevertheless, optical character recognition (OCR) has advanced a great extent for languages other than Arabic. However Arabic and languages using the Arabic script like Urdu or Farsi, have received the least attention in this field. This is due in part to the lack or little interest in the field and in part to the specificities of the Arabic language. It is written from right to left cursively most of its characters are connected even when typed or printed. Arabic characters are widely used in Arabic countries. If optical character recognition systems are available for Arabic characters, they will be very useful and will have a high commercial value**.**

In this proposal we intend to contribute in the efforts of automatic handling of Arabic in the Arab world, by developing an Arabic OCR. Our research is based on a system developed by one of the investigators of this proposal called ORAN (Offline Recognition of Arabic characters and Numerals). This system is in progress and has achieved a recognition rate of more than 97% for printed Arabic characters. The system was build and tested for the Arabic Naskh font. The recognition is achieved by matching a candidate character to a reference prototype build for this purpose. The prototypes are designed according to the description obtained for each class of a character shape. The description is obtained through a training process in which character features are extracted. The tool used for this is a method called the MCR (Minimum Covering Run) expression developed and modified to correctly describe character strokes in the Laboratory of Precision and Intelligence at Tokyo Institute of Technology.