Technology Profile Fact Sheet

Title: High Speed Model-Free Optical Character Recognition (OCR) Technique For Arabic Script

Aliases: None

Technical Challenge: Current OCR techniques require an expert to "train" the OCR device via software models. These models (often numbering in the thousands) are typically created from numerous samples taken from the intended source material. This process is inherently slow and once the device has been trained, its accuracy is strongly tied to the data sample that was used for training. These models are also vulnerable to changes in font typeface, which requires retraining. Their implementations are generally slow particularly when accuracy demands a large model set.

Description: This technique is not affected by font and does not require models for training. The software's user is not required to provide any information other than the original optical data in any of the common graphics formats (e.g., .tif or .gif). The current demonstration software (KOARS) reads and transcribes Arabic at the speed of approximately one page per second with an accuracy ranging from 90% (for cases in which the data contains dust, scratches, streaks, etc.) to 98%. The technique is applicable to a variety of other languages especially South Asian/Indian vocabularies.

Demonstration Capability: Current software implementation can be run on UNIX and LINUX platforms.

Potential Commercial Application(s): The most obvious applications are the filtering/selection of extremely high volume data (phonebooks) for which human readers cannot be spared, and the general processing of Arabic language data.

Patent Status: A patent application has been filed with USPTO.

Reference Number: 1389