**Machine-Printed and Handwritten Ethiopic Script Recognition**

YAREGAL ASSABIE LAKE

School of Information Science, Computer and Electrical Engineering,
Halmstad University

Department of Signals and Systems, Chalmers University of Technology

# ABSTRACT

A written language is represented by using machine-printed or handwritten symbols called characters. For automatic recognition of written languages, handwritten script can be captured offline (by a scanner) and online (by electronic digital devices), whereas machine-printed text is captured offline. In line with the method used to capture texts, automatic recognition can be made offline (after writing is completed) or online (at the time of writing). In this thesis, recognition systems for machine-printed and handwritten Ethiopic script are presented. While the main focus of this work is on offline recognition of machine-printed and handwritten Ethiopic script, it is also extended to online recognition of handwritten characters. The offline recognition system presented in this work treats recognition of machine-printed and handwritten characters, and handwritten Amharic words. Lexicons and hidden Markov models (HMMs) are used for recognition of Amharic words in unconstrained handwritten text.

In both machine-printed and handwritten script recognition systems, a similar set of features called *primitive structural features* and their *spatial relationships* are suggested as basic units of recognition. The idea behind these features is to represent graphically complex characters by less complex primitive structures and their spatial interrelationships. The advantage is that these features are easier to extract and process for recognition than complex-shaped characters. The features are extracted by making use of the 2D direction field tensor for offline recognition of machine-printed and handwritten script. The resultant direction field image is also used for text line detection, character segmentation, and word segmentation. For online recognition, a 2D direction field tensor parameterized by time is used to extract the features.

The thesis also presents datasets for testing the performance of recognition systems. The datasets, referred to as EDIDB for the machine-printed and DEHR for the handwriting, were collected from real-life sources and various writers. Test results reported in the thesis are based on these datasets unless specified otherwise.

**Keywords**: Ethiopic Character Recognition, Amharic Word Recognition, OCR, Handwriting Recognition, HMM, Direction Field Tensor, Structure Tensor.