Offline Handwritten Amharic Word Recognition Using HMMs

Yaregal Assabie and Josef Bigun
School of Information Science, Computer and Electrical Engineering
Halmstad University, Halmstad, Sweden
{yaregal.assabie, josef.bigun}@hh.se

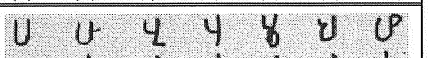
Abstract

This paper describes two approaches for Amharic word recognition in unconstrained handwritten text using HMMs. The first approach builds word models from concatenated features of constituent characters and in the second method HMMs of constituent characters are concatenated to form word model. In both cases, the features used for training and recognition are primitive strokes and their spatial relationships. The recognition system does not require segmentation of characters but requires text line detection and extraction of structural features, which is done by making use of direction field tensor. The performance of the recognition system is tested by DEHR dataset of unconstrained handwritten documents collected from various sources.

1. Introduction

Amharic is the official language of Ethiopia which has a population of over 80 million at present. It belongs to Afro-Asiatic language family, and today it has become the second most widely spoken Semitic language in the world, next to Arabic. Along with several other Ethiopian languages, Amharic uses Ethiopic script for writing. The Ethiopic script used by Amharic has 265 characters including 27 labialized (characters mostly representing two sounds, e.g. ቧ for ሸዋ) and 34 base characters with six orders representing derived vocal sounds of the base character. The alphabet is written in a tabular format of seven columns where the first column represents the base characters and others represent their derived vocal sounds. Part of the alphabet is shown in **Table 1**. Automatic recognition of unconstrained handwritten text is one of the most challenging pattern recognition problems due to the varying nature of the data. With respect to this inherent nature, various recognition methods have been proposed. Offline recognition of Latin, Chinese, Indian, and Arabic handwritten text has long been an area of active research and development [4], [5]. However, Ethiopic handwriting recognition in general, and Amharic word recognition in particular, is one of the least investigated problems. The difficulty in automatic recognition of Ethiopic script arises from the relatively large number of characters, their interclass similarity and structural complexity. In this paper, we present Amharic word recognition in unconstrained handwritten text using hidden Markov model (HMM).

Table 1. A sample of handwritten Ethiopic characters.

	Base Sound	Orders						
		1st (ä)	2nd (u)	3rd (i)	4th (a)	5th (e)	6th (ə)	7th (o)
1	h							
2	l							
3	h							
4	m							
5	s							
.	.							
.	.							
.	.							
32	f							
33	p							
34	v							

2. Description of the Recognition System

HMMs are doubly stochastic processes which model time varying dynamic patterns. The system being modeled is assumed to be a Markov process that is hidden (not observable), but can be observed through another stochastic process that produces the sequence of observations. The hidden process consists of a set of states connected to each other by transitions with probabilities, while the observed process consists of a set of outputs or observations, each of which may be emitted by states according to some output probability density function [6]. HMMs have emerged as a powerful paradigm for modeling pattern sequences in different areas such as speech recognition and online handwriting recognition.