ሀለ Amharic Letters Handwriting Recognition

----- RAD Mini Project ----

Submitted to: Desta Z. (MSc.)

ጥር 20 ፣ 2008 ዓ/ም

Table of Contents

1.0 Introduction

Handwriting is a personal biometric that is considered to be unique to an individual. I.e. People have different hand writing styles. In order to recognize these styles in computer systems, it is impossible to use predefined structures and identify them using simple if … this, else … this clauses. Hence, the solution to do handwriting recognition systems is using machine learning where the system learns from the provided learning dataset, either at runtime or design time, and makes a best guess on new writing styles.

2.0 Handwritten Recognition

Handwritten Recognition refers to the process of translating images of hand-written, typewritten, or printed digits into a format understood by user for the purpose of editing, indexing/searching, and a reduction in storage size.

Handwritten recognition system is having its own importance and it is adoptable in various fields such as online handwriting recognition on computer tablets, recognize zip codes on mail for postal mail sorting, processing bank check amounts, numeric entries in forms filled up by hand and so on.

There are two distinct handwriting recognition domains; online and offline, which are differentiated by the nature of their input signals. In offline system, static representation of a digitized document is used in applications such as check, form, mail or document processing. On the other hand, online handwriting recognition (OHR) systems rely on information acquired during the production of the handwriting. They require specific equipment that allows the capture of the trajectory of the writing tool. Mobile communication systems such as Personal Digital Assistant (PDA), electronic pad and smart-phone have online handwriting recognition interface integrated in them.

3.0 Amharic Handwritten Recognition

In Africa more than 2,500 languages, including regional dialects, are spoken. Some are indigenous languages, while others are installed by conquerors of the past. English, French, Portuguese, Spanish and Arabic are official languages of many of the African countries. As a result, most African languages with a writing system use a modification of the Latin and Arabic scripts. There are also many languages with their own indigenous scripts and writing systems. Some of these scripts include Amharic script (Ethiopia), Vai script (West Africa), Hieroglyphic script (Egypt), Bassa script (Liberia), Mende script (Sierra Leone), Nsibidi/Nsibiri script (Nigeria and Cameroon) and Meroitic script (Sudan)

	Ge'ez ä	Ka'eb u	Salis ï	Rab'e a	Hamis é	Sadis i	Sab'e o
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
ḥ	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
s	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
h	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
a	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
ẹ	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
ṭ	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
p	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
ts	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
ts	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

Figure 1: Amharic Alphabets

Amharic, which belongs to the Semitic language, became a dominant language in Ethiopia back in history. It is the official and working language of Ethiopia and the most commonly learnt language next to English throughout the country.

Accordingly, there is a bulk of information available in printed form that needs to be converted into electronic form for easy searching and retrieval as per users' need. Suffice is to mention the huge amount of documents piled high in information centers, libraries, museums and government and private offices in the form of correspondence letters, magazines ,newspapers, pamphlets, books, etc. Converting these documents into electronic format is a must in order to (i) preserve historical documents, (ii) save storage space, and (iii) enhance retrieval of relevant information via the Internet. This enables to harness existing information technologies to local information needs and developments. (Meshesha, Million;C.V Jawahar)

In order perfectly recognize Amharic texts in documents mentioned above, the first step is perfect recognition of letters in the Amharic alphabet. This can be done using different algorithms such as KDA, OCR and SVM.

4.0 Algorithms for Handwritten Recognition

Many types of classifier are applicable to the handwritten recognition system. Recognition of a pattern can be done using a template matching, statistical, syntactic (structural) and neural network approach. SVM classification is selected in this project because it gives a better recognition result than compared to other classifiers.

4.1 Support Vector Machine

In machine learning, **support vector machines** (**SVMs**, also **support vector networks**) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training

examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier.

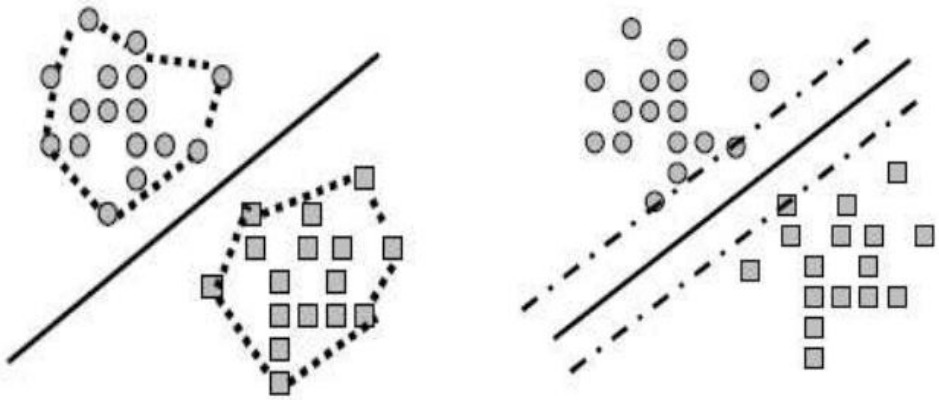
An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

SVM in its basic form implement two class classifications. It has been used in recent years as an alternative to popular methods such as neural network. The advantage of SVM is that it takes into account both experimental data and structural behavior for better generalization capability based on the principle of structural risk minimization (SRM). Its formulation approximates SRM principle by maximizing the margin of class separation, the reason for it to be known also as large margin classifier.

SVM classifier is an algorithm which maximizes the margin between the classes and minimizes the classification error. SVM used to identify a set of linearly separable hyper planes which are linear functions of the feature space. The hyper planes are placed such that there is a maximum distance between the classes.



4.2 Multi class Support Vector Machine

Binary classifiers like SVM are basically designed for two class classification problems. However, because of the existence of a number of characters in any script, optical character recognition problem is inherently multi-class in nature. The field of binary classification is mature, and provides a variety of approaches to solve the problem of multi-class classification.

Most of the existing multi-class algorithms address the problem by first dividing it into smaller sets of a pair of classes and then combine the results of binary classifiers using suitable voting methods such as majority or weighted majority approaches

Multi-class SVMs are usually implemented as combinations of two-class solution using majority voting methods. It has been shown that the integration of pairwise classifiers using decision directed acyclic graph (DDAG) results in better performance as compared to other popular techniques such as decision tree, Bayesian, etc.

It can be observed that the number of binary classifiers built for an N class classification problem is (N) (N-1)/2.

5.0 Accord.Net – C# Machine Learning Library

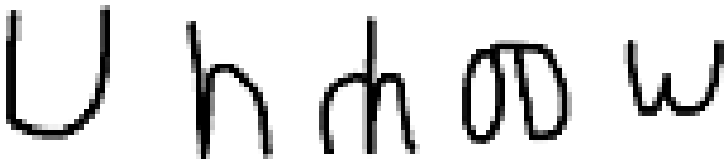
The Accord.NET Framework is a .NET machine learning framework combined with audio and image processing libraries completely written in C#. It is a complete framework for building production-grade computer vision, computer audition, and signal processing and statistics applications even for commercial use.

6.0 Our Application

6.1 Dataset Collection

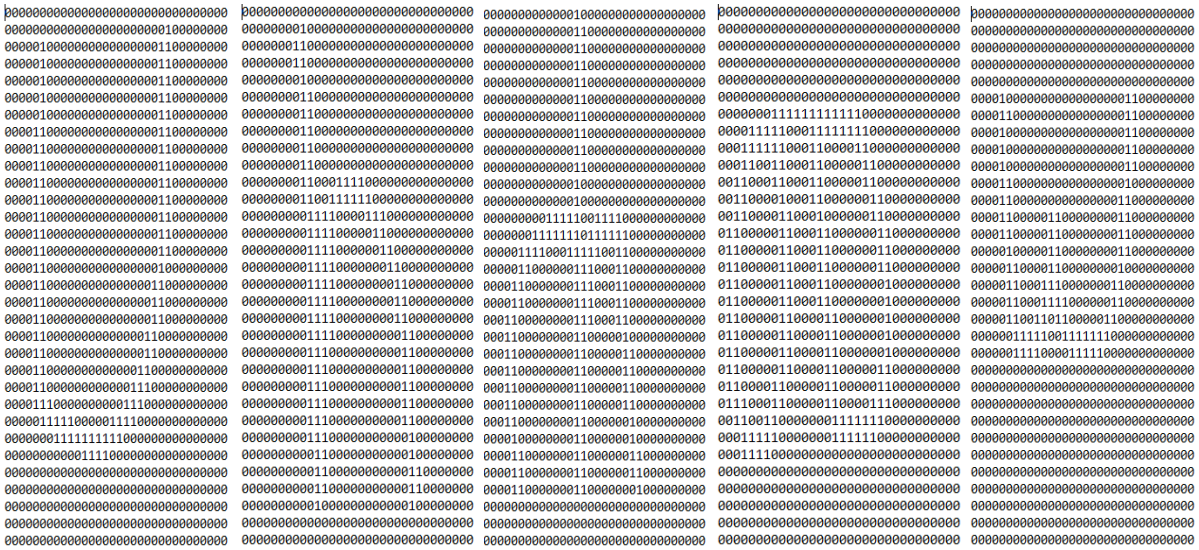
Since we couldn't find any available dataset, we were supposed to prepare our own ones. The dataset we prepared is collected from 12 writers. Each writer contributed 50 characters, 10 letters from each of the letters selected to be dealt in this mini project. Hence, the total number of samples in the dataset is 600 (12 * 50).

The handwriting samples were collected using paint program installed in Windows machines. We then resized the sample sizes to be 32 * 32 pixels. The dataset contains 12 folders, each representing the data from a singular writer. Every folder contains another 5 folders in it, which are created for storing 10 samples of each selected letter (ሀ፣ለ፣ሐ፣መ፣ሠ)



6.2 Preprocessing

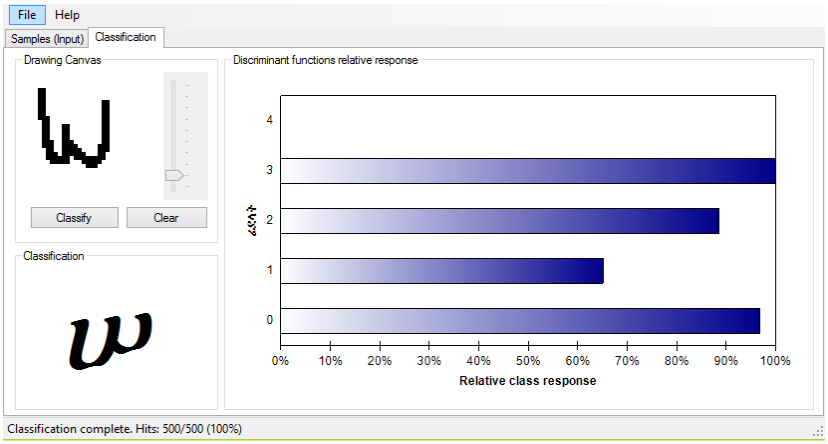
After collecting the datasets, we then used **MatLab** to convert the letters to **binary images**; images represented as black and white with each pixel holding one of the binary values, and represent the letters as 32x32 matrices.



6.3 Experimental Application

In order to use the experimental application that we prepared, first press load data button. This will load entries from the Amharic letters' dataset into the application.

The next step is to perform the training. To do so, press start Training button. This will train the system with the Amharic letters' training dataset. Then the Multiclass Support Vector Machine will create definitive class boundaries between the selected five Amharic letters.



Once the training is complete, the next thing to do is to test the system against Amharic letters' testing dataset by pressing start testing button. This will show the test results of every feature; compare the result with the expected result and finally show the total efficiency.

Training and testing are the two basic phases of any pattern classification problem. During training phase, the classifier learns the association between samples and their labels from labeled samples. The testing phase involves analysis of errors in the classification of unlabeled samples in order to evaluate classifier's performance. In general it is desirable to have a classifier with minimal test error.

6.4 Results

We can see that our system correctly identifies 94% of the testing data. Notice that the testing and training datasets are independent and disjoint.

References

(1) *Handwritten Digit Recognition using Support Vector Machine*, Anshuman Sharma
(2) *Handwritten English Character And Digit Recognition Using Multiclass SVM Classifier And Using Structural Micro Features*, Shubhangi D. C, Prof. P. S. Hiremath
(3) *Optical Character Recognition of Amharic Documents*, Million Meshesha, C. V. Hawahar
(4) *The study of Handwriting Character Recognition (HCR) and Support Vector Machine*, Dewi Nasien, Habibollah Haron, Siti Sophiayati Yuhaniz