Online Handwriting Recognition of Ethiopic Script

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

Abstract

Online recognition of handwritten characters is gaining a renewed interest as it provides a natural way of data entry for a wide variety of handheld devices. In this paper, we present online handwriting recognition system for Ethiopic script based on the structural and syntactical analysis of the strokes forming characters. The complex structures of characters are represented by the spatio-temporal relationships of simple-shaped strokes called primitives. A special tree structure is used to model spatio-temporal relationships of the strokes. The tree generates a unique set of primitive stroke sequences for each character, and for recognition each stroke sequence is matched against a stored knowledge base. Characters are also classified based on their structural similarity to select a plausible set of characters for un unknown input, which improves recognition and processing time. We also present a dataset collected for training and testing online recognition systems for Ethiopic script. The dataset is prepared in accordance with the international standard UNIPEN format. The recognition system is tested with the collected dataset and experimental results are reported.

Keywords: Ethiopic, Handwritten, Online Recognition.

1. Introduction

Handwriting can be recognized online or offline. In online handwriting recognition, character symbols are converted into their equivalent text at the time of writing, as opposed to offline recognition which processes a scanned image of characters after writing is completed. Online recognition is gaining a renewed interest because of its applications in the digital world. A wide variety of handheld electronic devices such as TabletPCs, PDAs, Digimemos, and mobile phones have now become common tools for day-to-day personal use. However, their smaller-sized keyboard poses a challenge for data entry as compared to desktop computers. This problem has brought a growing need of online handwriting recognition as an alternative means for data entry. Its merit also comes from the fact that handwriting is the most natural method of text entry for many users.

Ethiopic script is used as a writing system for many languages spoken in Ethiopia. Many languages in the country use the script for writing, but it has been largely used by Geez and Amharic, which are the liturgical and official languages of the country, respectively. Ethiopic is a modification-based script where the modifiers ('vowels') are usually added to the base character to give a derived vocal sound. Sometimes, the modification can also be achieved by slightly deforming the shape of the base character. Although Ethiopic alphabet has recently been standardized to have 435 characters, roughly half of them are used practically in daily communications by the official language Amharic and other major languages. The alphabet is conveniently written in a tabular format of seven columns (orders) where the first column represents the base character and other columns represent derived vocal sound of the base character. Part of the alphabet is shown in Table 1.

Table 1. Part of the Ethiopic alphabet

	Base Sound	Orders						
		1st (ä)	2nd (u)	3rd (i)	4th (a)	5th (e)	6th (ə)	7th (o)
1	h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
2	l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
3	h	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
4	m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
5	s	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
6	r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
7	s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
8	š	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ
9	*q*	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
10	*b*	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
31	p'	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
32	f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
33	p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ
34	v	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ

Research and development on online handwriting recognition of scripts like Latin, Chinese, and Japanese has reached a level of maturity where the applications of online recognition are now available for market. However, only little has been studied on handwriting recognition of Ethiopic text in general, and to the knowledge of authors, there is no published work on this research area. Owing to the large number of characters, a combination of up to three keys (in Latin-based keyboard) is required to write a single Ethiopic character. Because of this, writing Ethiopic text into a computer is largely limited to trained typists even in desktop computers. Thus, the potential application of online handwriting recognition system for Ethiopic is enormous because it would help decrease the need for special training, that is acquired with difficulty currently, to write Ethiopic text into computers efficiently.

Several techniques have been employed for recognition of online handwritten characters. Time delay neural networks, elastic matching (deformable template or dynamic time warping), Hidden Markov Model (HMM), structural (stroke) analysis, and a combination of multiple classifiers are some of the commonly used techniques [5], [6], [7], [8]. In this paper, we present online recognition system for handwritten Ethiopic script which is based on the spatio-temporal relationships of primitive-shaped strokes whose combination forms complex structures of characters. The sequences of primitive strokes and their relationships of the unknown character symbols are matched against a knowledge base for recognition. We classified characters based on their structures so that better prediction is made for the unknown input and the time for pattern matching is minimized. We also collected a dataset of online handwritten characters. The dataset is collected using digital pen and ordinary paper placed on a Digimemo writing pad, which produces digital pages. Character symbols in digital pages are then converted to UNIPEN format. The dataset is divided into training and test sets.

2. Dataset Collection

To compare the results of character recognition systems for a specific script, there has to be a standard dataset which is used for training and/or testing. In the case of online character recognition, an internationally accepted UNIPEN format has been set to standardize the format in which the trajectories of pen-tip movements are to be stored [4]. Scripts such as Latin, Chinese, Arabic, and Indic now have datasets with the international standard format. However, we have assessed and found out that no dataset for online Ethiopic characters is collected so far in UNIPEN format. Therefore, as a benchmark for the study of online Ethiopic character recognition, we collected dataset in accordance with the UNIPEN format, the specification of which is presented below.

2.1. Data collection tool

There are different devices used to capture and record the trajectories of pen-tip movements in online handwriting. Most of them, e.g TabletPC and PDAs capture the digital ink directly on their screen and other devices like Digimemos capture the digital ink using ordinary paper placed on their writing pad. There is an advantage of using Digimemos in that it creates a natural feeling of writing on a paper in which native writers are already used to[1]. It is also cost-effective for large scale data collection and therefore it is affordable as compared to TabletPCs and PDAs. Data collection forms are prepared in A5 and A4 papers. The forms contain Ethiopic characters with rectangular boxes beneath each of them, in which the writers would be writing the corresponding character. The writers were provided with the forms clipped to the Digimemo pad and a digital pen is used for writing. A sample data collection form filled by a writer is shown in Fig. 1.

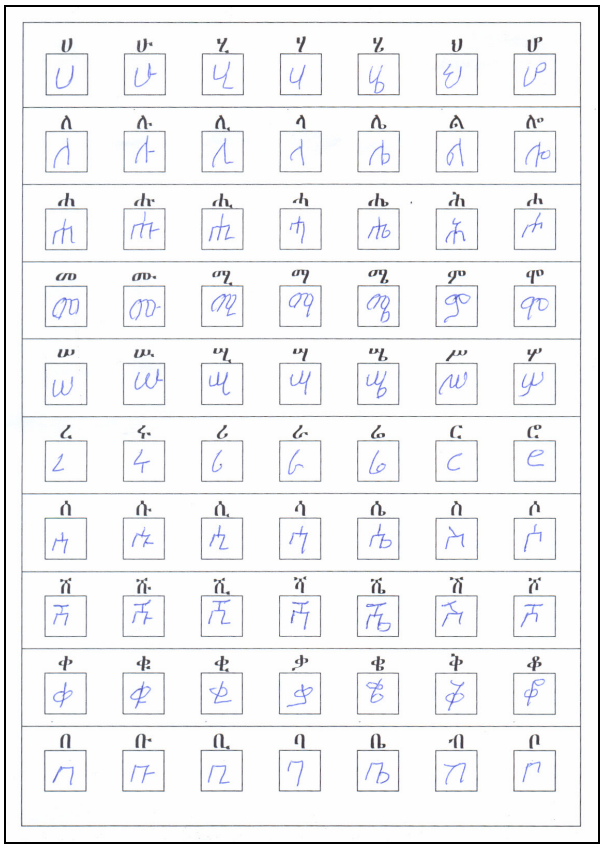

Figure 1. sample form filled by a writer

1 We used ACECAD® Digimemo A502 (for A5 size) and A402 (for A4 size) devices for our data collection task. These have a resolution of 1000 points per inch and record 125 points per second.

The devices capture digital inks written on the forms and store them as digital pages. For each character in the forms, the dimension and absolute location of its corresponding box is recorded and also the corresponding area in the digital page is extracted to be stored as its equivalent online data. A correspondence problem between digital pages and the forms arises if the user is tilting the pen during the writing process. To circumvent this, we used larger virtual boxes in the digital pages so that no digital ink remains with out being extracted. The extracted character symbols in the digital pages are stored in UNIPEN format. We used the Open Source Lipi Toolkit to convert digital inks of each character symbol into a file in the UNIPEN format [4].

2.2. The dataset

The dataset is roughly divided into two groups based on the paper sizes used (A4 or A5). In the case of A5 size, the boxes in the form have a uniform size of 10mm-by-10mm, where as for the A4 size, the boxes are designed uniformly as 15mm-by-15mm. The purpose is to include different sizes of characters in the dataset. Apart from box sizes, no other limitation was imposed on writers. The writers were oriented to feel as if they were writing naturally on a paper. Although there are currently 435 characters in the Ethiopic alphabet, many of them are designed recently for minority languages and they are not used practically in daily communications by the official language Amharic and other major languages. Therefore, the dataset includes the 34 base characters and their derivatives (i.e., 34x7=238 characters), which are commonly used by most Ethiopian languages including Amharic. Despite their rare appearance in texts, the dataset also includes the so called labialized characters (most of which are used to represent two vocal sounds, e.g. ቋ for ቁዋ) listed as the following: ኋ ሟ ሯ ሷ ሿ ቋ ቧ ቷ ቿ ኋ ኗ ኟ ኳ ዟ ዃ ዷ ጇ ጎ ጧ ፏ ፗ ቈ ኰ ጐ ጕ and ኧ.

Thus, a total of 265 Ethiopic character symbols are included in the dataset and 204 writers have participated in the data collection task where each writer writes all the 265 character symbols. Therefore, the dataset consists of a total of 54,060 samples belonging to different age groups, gender, educational background and region. Each sample data is evenly distributed and the dataset is also divided as training and test data. The training set consists of 16,165 (30%) samples from 61 writers and the test set consists of 37,895 (70%) samples collected from 143 writers. Both the training and test data contain all sample characters from the two dataset groups, i.e., from the A5 and A4 datasets.

3. The Online Recognition System

The online recognition system for Ethiopic script is developed based on the offline handwriting recognition presented in [2]. The underlying techniques are also applied for machine printed character recognition and are also further explained in [1]. To make it applicable for online handwriting, however, we modified and extended the system. Below we briefly describe the techniques applied in handwriting recognition in general and online recognition in particular.

Structural and syntactic techniques are applied to model the spatio-temporal relationships of the primitive strokes. The structural and syntactic model encodes the orientation, structure, relative spatial position, and relative length of primitive strokes, to reduce the impact of their absolute size on recognition. It means that the recognition system does not depend on the size of character symbols. Primitives are extracted from the 1D signal, i.e. UNIPEN data file, and classified based on their orientation. A special tree structure is used to handle the spatio-temporal relationship and the tree is traversed to generate a set of unique sequence of primitive strokes for each character. The generated sequence of strokes is matched against a knowledge base built for Ethiopic alphabet. To improve the recognition process, we also classify characters based on their primitive characteristics. The flowchart of the recognition system is shown in Fig. 2 and the details are presented below.

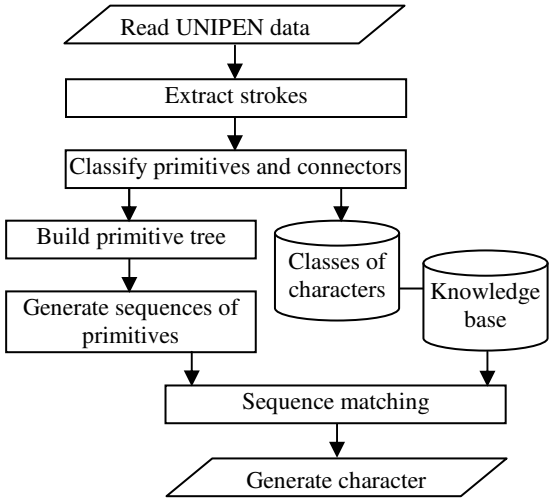

Figure 2. The online recognition system

3.1. Primitives

In our proposed technique, vertical and diagonal lines in characters are defined as primitives and horizontal lines are considered as connectors of primitive structures. Primitives are then hierarchically classified based on their orientation/structure type, relative length with in the character, and relative spatial position. For the purpose of computation, each classification level is assigned with numbers, and the classification is given as follows.

i. **Orientation/structure type**: There are three groups of orientations for primitive strokes namely, *forward slash* (9), *vertical* (8), and *backslash* (7). *Appendages* (6) do not fit to a specific orientation. Rather, they are recognized by their structure type in the case of machine printed text, e.g. in ፡. In handwritten text, *appendages* are usually not marked well. Because of this, we had to assume that they are always present at the end points of horizontal lines as in ት.

ii. **Relative length**: The orientation of primitives is further classified based on their relative length as *long* (9), *medium* (8), and *short* (7). Long is defined as a primitive that runs from the top to the bottom of the character, whereas short is a primitive that touches neither the top nor the bottom of the character. *Medium* refers to a primitive that touches either the top or the bottom (but not both) of the character. Due to their small size, *appendages* are always considered as *short*.

iii. **Relative spatial position**: At this level of classification hierarchy, primitives are further classified according to their spatial position with in the character as *top* (9), *top-to-bottom* (8), *bottom* (7), and *middle* (6). *Short* primitives can only have a relative spatial position of *middle*. *Top-to-bottom* position applies to *long* primitives which run from the top to the bottom of the character. Primitives with *medium* relative size can have either *top* or *bottom* spatial position. *Appendages* may appear at the *top*, *middle*, or *bottom* of the character.

Thus, a total of 15 primitive types are defined to represent all the 435 Ethiopic characters, and Table 2 summarizes the list of primitives and their numerical codes.

Table 2. Hierarchical classification of primitives

Orientation/ Structure	Length	Position	Code	Example Character
Vertical	Long	Top-to-bottom	898	
	Medium	Top	889	
		Bottom	887	
	Short	Middle	876	
Forward Slash	Long	Top-to-bottom	998	
	Medium	Top	989	
		Bottom	987	
	Short	Middle	976	
Backslash	Long	Top-to-bottom	798	
	Medium	Top	789	
		Bottom	787	
	Short	Middle	776	
Appendage	Short	Top	679	
		Middle	676	
		Bottom	677	

3.2. Spatio-temporal relationships of primitives

Spatial relationships of primitives refer to the way primitives are connected to each other. In the case of offline text the spatial position of primitives can be estimated from the (x,y) coordinates of image pixels forming primitives. In the case of online data, the spatial positions are also computed from the (x,y) data points extracted from trajectories of pen-tip movements. Since the (x,y) data points are arranged in chronological order, we are able to compute the direction of primitives. Thus, in addition to the spatial information, we also use temporal information to build the relationship between primitives. This temporal information is used to clear redundant parts of primitives which occur when users are over-writing on the lines. As shown in Fig. 3, redundant lines are spatially close to each other and usually have opposite directions.

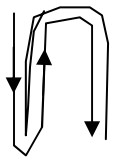

Figure 3. Redundant lines in the character ቦ.

A primitive can be connected to another at one or more of the following regions of the strokes: *top*, *middle*, and *bottom*. Connection regions between primitives are also assigned with numbers as: *top* (1), *middle* (2), and *bottom* (3). The number 4 is used for cases where there is no connection. A connection between two primitives is represented by ***xy*** where ***x*** and ***y*** are numbers representing connection regions for the *left* and *right* primitives, respectively. Between two primitives of Ethiopic characters, there can also be two or three connections, and a total of 18 connection types are identified, which are listed as: 11 (ቨ), 12 (ቶ), 13 (ሯ), 21 (ኩ), 22 (ዘ), 23 (ዦ), 31 (ኒ), 32 (ዛ), 33 (ዟ), 1123 (ዥ), 1132 (ዣ), 1133 (ዧ), 1232 (ዻ), 2123 (ዤ), 2132 (ዣ), 2133 (ዤ), 112232 (ዣ), and 112233 (ዧ).

We model the spatio-temporal relationships of primitives of a character in a special tree structure known as *primitive tree*, which is designed based on the interconnection analysis of primitives in Ethiopic characters. The primitive tree inherits the properties of binary search trees by arranging the primitives as left and right according to their relative spatial positions. Figure 4 shows the general structure of primitive tree and child nodes represent primitives connected to the parent primitive at different connection regions. The two middle nodes in the right nodes is due to connection of two primitives at the middle of the right of their parent primitives for some characters like ት and ፑ.

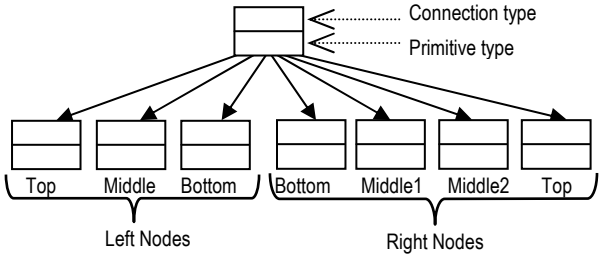

Figure 4. General structure of primitive tree

A primitive stroke spatially located at the left top position of the character is selected as a root node. Other primitives are recursively built into the tree based on their spatial position and connection types. Figure 5 shows an example of primitive tree for the character 千.

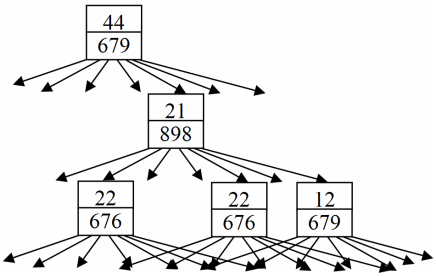

Figure 5. Primitive tree for the handwritten character 千.

3.3. Extracting primitives and connectors

Strokes are a series of time-ordered (x,y) coordinates (online data points) of the trajectories of pen-tip movements collected from pen-down to pen-up. A stroke can represent a single primitive, a connector, a combination of primitives and connectors, or the whole character. On the other hand, two or more strokes can make a primitive or a connector. Data points are grouped either as parts of primitives or connectors based on their orientation (degree). To compute the orientation at every data point, we first smoothed the 1D discrete (x(t), y(t)) signals with a Gaussian to straighten high curvature points in the strokes, introduced usually when users write slowly. The orientation of a point is then computed by averaging the orientation of the lines formed with its neighborhood points. It means that for the data point $\boldsymbol{b}$, the orientation $\boldsymbol{\theta}$ of $\boldsymbol{b}$ with coordinates (b_x,b_y) is computed as:

$$\theta(b) = \sum_{t} \left(x'(t) + iy'(t)\right)^2 g(t) \qquad \textbf{(1)}$$

where $x'(t) = \frac{dx(t)}{dt}$ and $g(t)$ is a Gaussian. The orientation is represented in double angle and opposite directions have the same orientation. By converting into a simple angle representation in the range of [0..180) degrees, primitives (diagonal and vertical lines) are then extracted as consecutive data points which have orientations of [30,150] degrees. The remaining data points are built to form connectors (horizontal lines). Primitives are further classified based on their relative length, orientation and spatial position. The orientation $\boldsymbol{\alpha}$ of a primitive is computed as:

$$\alpha = \frac{1}{n} * \left(\sum_{i=1}^{n} \theta_i \right) \qquad \textbf{(2)}$$

where $\boldsymbol{\theta_i}$ denotes the orientation of data points that form the primitive and $\boldsymbol{n}$ is the total points.

3.4. Matching sequences

Because of the computational cost of trees, the information stored in primitive trees is converted to a string data structure using in-order traversal, i.e., in the order of {**left**{top, middle, bottom}, **parent**, **right**{bottom, middle1, middle2, top}}. This traversal generates a unique sequence of primitives and their connections in the form of string data. For example, the primitive tree in Fig. 5 is converted to: {44,679,22,676,21,898,22,676,12,679}. For each character in the Ethiopic alphabet, possibly occurring sequences of primitive strokes and their connections are stored as a knowledge base. Recognition of character symbols is achieved by matching their sequence of primitive strokes and connections against the knowledge base. During sequence matching, the similarity between the unknown input and each record in the knowledge base is computed and the best match is considered for recognition, which is decided based on a similarity threshold.

3.5. Classification of characters

Writers may not properly write connectors, which poses a problem to draw the spatial relationships and eventually it becomes difficult to recognize the unknown input. In such cases minimal information about spatial relationships is sufficient to recognize the unknown input if the characteristics of primitives is known. Thus, we classify characters using the following characteristics of primitives: *number of salient primitives (s)*, *number of long primitives (l)*, *number of salient primitives touching the top of the character (t), number of salient primitives touching the bottom of the character (b),* and *number of appendages (a)*. Salient primitives are defined as long, medium and short primitives. Each character C is represented by a feature vector as $C = (s, l, t, b, a)$ where $s \in \{1, 2, \ldots, 8\}$, $l \in \{0,1, 2, 3, 4\}$, $t \in \{1, 2, 3, 4\}$, $b \in \{1, 2, \ldots, 6\}$, and $a \in \{1, 2, \ldots, 6\}$. Then, based on their feature vector values, characters are classified and stored in to cells of a five-dimensional space of size 8x5x4x6x6. The classification stores structurally similar characters in the same cell or its neighborhood. Since, it is possible to locate more likely candidate characters for the unknown input (based its primitive characteristics), the classification of characters also so helps to minimize the time the taken by pattern matching process.

4. Experiment

One-dimensional Gaussian window of size 5 data points is used for smoothing data points along the stroke. The size is determined through experiment and the purpose is to straighten jaggy curves to reduce the risk of partitioning a single primitive or connector into two or more. The recognition system is tested on the collected dataset, both on the training and test sets for future comparisons. Because our system does not need training, as it relies on a stored knowledge base for recognition, there was no difference in the recognition rate between the two experiments, which is to be expected. The system also does not require size normalization of characters for we are extracting the relative size (not the absolute size) of primitives. However, characters with large sizes from the A4 dataset tend to be recognized slightly better than their corresponding small size characters from the A5 dataset. This is because, for large size characters, there is a better chance of correctly extracting, classifying, and building the spatio-temporal relationships of primitives.

The recognition rate varies due to the complexity of characters. An average recognition rate of 96% was achieved for simple-shaped characters like  ,  , ,  , ,  , ,  , etc. and their derivatives. Comparable result was also achieved for others if they are neatly and properly written, e.g.  ,  , ,  . However, The recognition drops down to an average of 82% for complex characters with small size primitives like  ,  , and their derivatives. Rarely used characters such as the derivatives of  and labialized characters like  ,  , , and are also among those with low recognition. The reason is that writers are not frequently using them and therefore could not write them correctly. For example, several writers wrote  , ,  ,  ,  , for the character  , where a random test on native users reveal that the samples are unrecognizable or mistaken for  ,  and  . For a few set of characters, writers also write similar shapes resulting in confusion even for human beings, unless the characters appear in a text to be recognized contextually. For example,  ,  ,  ;  ,  , ;  ,  ,  were collected in the dataset as character symbols for  ,  , and  respectively.

Additional recognition errors can arise from various sources. The major part of such errors occurred during extraction of primitives and connectors. Writers usually repeat lines, e.g. (), and if they fail to pass through the original line, it will come out as two separate primitives putting a setback in the recognition process, which is also a challenge for Latin as well as other writing systems. The extra lines from "pen-up to pen-down" drawn by some writers, e.g. () also pose a similar problem during extraction of primitives and connectors. Besides, unconnected primitives, as in (), are other sources of errors. It is even more ambiguous when there is another similar character to be easily misclassified; e.g. () is recognized as . Other errors arise from the pen-tilt during the writing process. If strokes of a character are written with different pen-tilts on a Digimemo device or its likes, they may not be captured exactly as they appear on the paper. For example, connected strokes on the paper may be captured as unconnected on the digital page.

5. Conclusion

We presented online recognition system and dataset for Ethiopic script. The dataset can be used as a benchmark for testing and comparing online recognition systems for Ethiopic script. The proposed recognition system does not need training because the knowledge base stores possibly occurring sequences of primitives and connectors for each handwritten character. The system also does not require size normalization as we encode only the relative length of primitives. Therefore, the recognition system is reasonably size-insensitive and writer-independent reaching 96% recognition rate. The structural and syntactic analysis provides efficient mechanism to handle neatly and properly written characters. The recognition is improved by employing characteristics of primitives, which helps to predict plausible set of characters for the unknown input.

References

[1] Y. Assabie and J. Bigun, "Multifont size-resilient recognition system for Ethiopic script", *IJDAR*, 10(2): pp. 85-100, 2007.

[2] Y. Assabie and J. Bigun, "Writer-independent offline recognition of handwritten Ethiopic characters", *accepted in ICFHR2008, Montreal, Canada, 2008.*

[3] J. Babu, L. Prasanth, R. Sharma, P. Rao, and A. Bharath, "HMM-based online handwriting recognition system for Telugu symbols", *Proc. 9th ICDAR'07*, Curitiba, Brazil, pp. 63-67, 2007.

[4] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet, "UNIPEN project of on-line data exchange and recognizer benchmarks", *Proc. 12th ICPR'94*, Jerusalem, Israel, pp. 29-33, 1994.

[5] S. Jaeger, S. Manke, and A. Waibel, "NPEN++: An online handwriting system", *Proc. 7th IWFHR'00* , Amsterdam, pp. 249-260, 2000.

[6] C. L. Liu, S. Jaeger, and M. Nakagawa, "Online recognition of Chinese characters: The state-of-the-art", *IEEE Trans. PAMI*, 26(2): pp. 198-213, 2004.

[7] R. Plamondon, S.N. Srihari, "On-line and off-line hand writing recognition: A comprehensive survey", *IEEE Trans. PAMI*, 22(1): pp. 63-84, 2000.

[8] S. Uchida and H. Sakoe, "A survey of elastic matching techniques for handwritten character recognition", *IEICE Trans. Inf. & Syst.*, E88-D(8): pp. 1781-1790, 2005.