2009 10th International Conference on Document Analysis and Recognition

Rearrangement of Recognized Strokes in Online Handwritten Gurmukhi Words Recognition

Anuj Sharma
Department of Mathematics, Panjab University, Chandigarh (INDIA)
anujs@pu.ac.in

Rajesh Kumar and R.K. Sharma
School of Mathematics and Computer Applications, Thapar University, Chandigarh (INDIA)

Abstract

This paper presents a system to recognize online handwritten Gurmukhi words. We have proposed a new step as rearrangement of recognized strokes in online handwriting recognition procedure. The rearrangement of recognized strokes includes: strokes identification as dependent and major dependent strokes; the rearrangement of strokes with respect to their positions; the combination of strokes to recognize character. We have achieved an overall recognition rate as 81.02% in online handwritten cursive handwriting for a set of 2576 Gurmukhi dictionary words.

Keywords: Online handwriting recognition, post-processing, words recognition.

1. Introduction

Online handwritten words recognition for Indian scripts have made impressive improvements in recent years [1-11]. The present work has been done for Gurmukhi script. Gurmukhi is the script of Punjabi language which is widely spoken across the globe. Gurmukhi shares many similarities to other asian scripts such as Devanagiri and Bangla. Gurmukhi script is written in left-to-right direction and in top-down approach. Most of the characters have a horizontal line at upper part. The characters of words are connected mostly by this line called head line. A word in Gurmukhi script can be partitioned into three horizontal zones, namely, upper zone, middle zone and lower zone. The upper zone denotes the region above the head line, where some of the vowels and sub-parts of some other vowels reside, while the middle zone represents the area below the head line where the consonants and some sub-parts of vowels are present. The middle zone is the busiest zone. The lower zone represents the area below middle zone where some vowels and certain half characters lie in the foot of consonants. These zones are shown in Fig. 1.

This paper is an attempt towards the development of a system that could recognize online handwritten Gurmukhi words. In the next section, system design has been discussed. Section 3 includes development of online handwritten Gurmukhi words recognizer. Experimental results are discussed in Section 4.

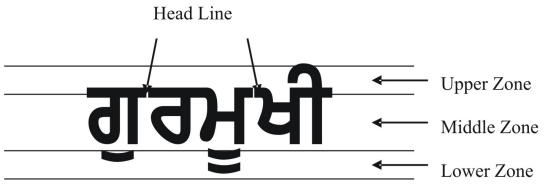

Fig. 1: Upper, Middle and Lower zones in Gurmukhi script.

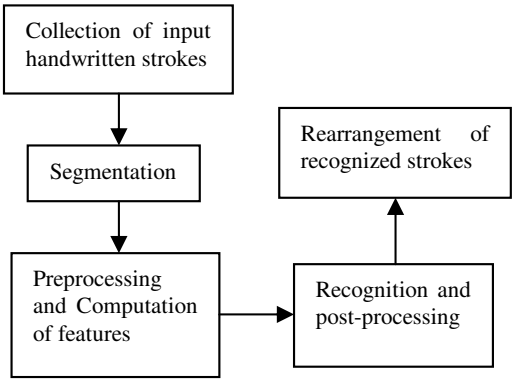

Fig. 2: Recognition stages in online handwritten Gurmukhi words.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.36

2. System Design

The understanding of recognition of Gurmukhi words based on input handwritten strokes has been presented in Fig. 3. In Fig. 3, n is the total number of input handwritten strokes and S_i represent the input handwritten stroke, where $1 \leq i \leq n$. The recognized stroke of corresponding input handwritten stroke is represented by RS_i. The characters recognized through combination of recognized strokes are represented by C_j, where $1 \leq j \leq m$, and $m \leq n$ as one or more strokes recognizes a character. The dotted lines between recognized strokes and recognized characters show other possibilities of various combinations through recognized strokes. A word is mixture of recognized characters in sequence.

After using K-means clustering technique, it is observed that the recognition of Gurmukhi strokes may result in one of the 40 different types of strokes [11]. These 40 strokes for Gurmukhi script are presented in Table 1. The recognition of characters in word is mainly based upon position and ordering of input handwritten strokes other than recognized strokes only. The process flow of online handwritten Gurmukhi words is presented in Fig. 4. The different steps of recognition procedure are discussed in next section.

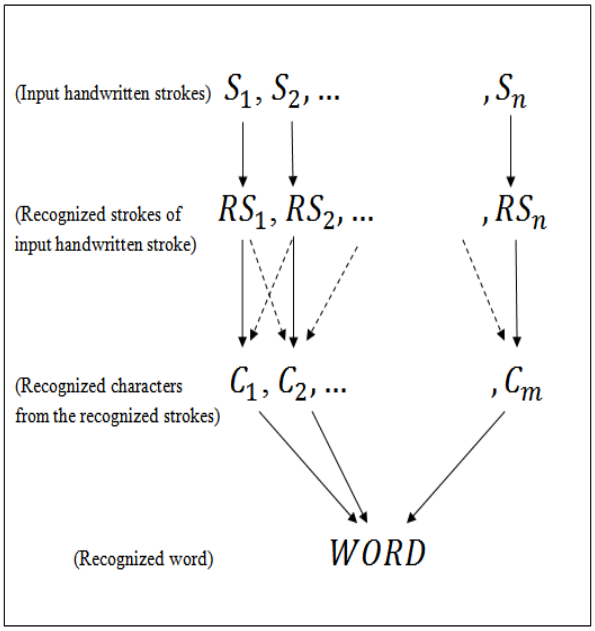

Fig. 3: Recognition of word from input handwritten strokes.

Table 1. Unique strokes and their ids.

Stroke id	Stroke shape	Stroke id	Stroke shape	Stroke id	Stroke shape	Stroke id	Stroke shape
11		21		31		41	
12		22		32		42	
13		23		33		43	
14		24		34		44	
15		25		35		45	
16		26		36		46	
17		27		37		47	
18		28		38		48	
19		29		39		49	
20		30		40		50	

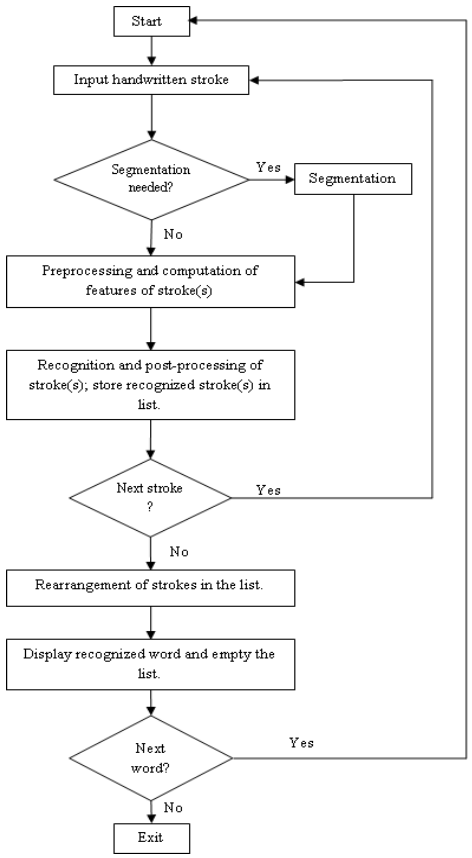

Fig. 4: Online handwritten Gurmukhi word recognition process.

3. Recognition of Online Handwritten Gurmukhi Words

The established procedure to recognize online handwriting includes: collection of input handwritten stroke; segmentation; preprocessing; computation of features; recognition; post-processing. We have introduced a new step as rearrangement of strokes after post-processing step where recognition of characters from recognized strokes has been performed. The above mentioned steps are discussed from subsections 3.1 to 3.4.

3.1 Collection of input handwritten stroke and its segmentation

A typical format of online handwriting data is a sequence of coordinate points of the moving pen point. Connected parts of the pen trace, in which the pen point is touching the writing surface, are called strokes. The collected stroke passes through the decision process of segmentation. If segmentation is required, it goes to segmentation phase, otherwise, stroke is sent to preprocessing phase. It has been noted that a good number of large strokes appear in online cursive word handwriting. In most of instances, these large strokes are mixture of one or more strokes. Segmentation is needed to segment large strokes into sub strokes so that suitable recognition of sub strokes can be performed. We have implemented a point based segmentation procedure that segments the large strokes into sub strokes on the basis of average number of points. Using K means clustering technique, the average number of points has been observed as '300' in our experiments for any stroke. A stroke is segmented at a point where number of points from start of stroke has crossed above 300 points and the movement at this point is less than 90 degrees. It has been noted that movement less than 90 degrees results in sharp angular behavior of strokes. The new stroke(s) obtained after segmentation are sent to preprocessing phase.

3.2 Preprocessing and Computation of Features

Preprocessing and computation of features are performed to the stroke(s) obtained after collecting input handwritten stroke and if any, segmentation is done. The preprocessing stages have been used as size normalization and centering of stroke, interpolating missing points, smoothing, slant correction and resampling of points. The features are computed after preprocessing of input handwritten stroke. The high level features are computed on the basis of low level features. The high level features include loop, crossings, straight line, headline and dots. Some of the common low level features are position of stroke, area, length, curliness and slope. The use of preprocessing stages and computation of features help in recognition stages [12].

3.3 Recognition and Post-processing

We have used elastic matching as the recognition method in recognition of input handwritten strokes [11]. Post-processing is applied after recognition process in order to refine the recognition results. We have applied trade-off of computed features with respect to recognition results. If recognition of stroke demands presence of particular features and such features exists, then only stroke is recognized correctly. For examples, ਹ and ਰ; ਤ and ਭ; ਤ and ਡ; ਟ and ਫ; ਦ and ਢ are distinguished on the basis of loop available in the major dependent stroke. ਮ and ਸ; ਪ and ਧ; ਖ and ਥ are distinguished on the basis of headline and its position in the character. ਅ; ਚ; ਦ; ਢ; ਬ; ਲ; ਛ; ਣ; ਛ; ਝ; ੜ; ੲ are distinguished on the basis of crossings inside a stroke. The strokes with stroke ids, 11, 12, 13 and 14 are distinguished on the basis of curliness and linearity of the strokes, respectively. ਸ਼; ਖ਼; ਗ਼; ਜ਼; ਫ਼; ਲ਼ are distinguished on the basis of dot feature of the stroke in the character. (ਹ or ਰ) and ਗ; ਵ and ਵ; ਤ and ੜ are distinguished on the basis of straight line or extra stroke in the character.

3.4 Rearrangement of Strokes

The nature of this proposed step in recognition procedure is such that it can not be included in post-processing step. We have classified recognized strokes as dependent and major dependent strokes. A character is combination of one or more strokes. A stroke that is mandatory to recognize a character is called major dependent stroke and other strokes are dependent strokes. It is worth mentioning here that the recognition of a word can not happen without its major dependent strokes and dependent strokes. It is because missing any stroke could result in incorrect word. In order to make suitable combination of dependent and major dependent strokes before recognition of particular word, we have rearranged these strokes. The process of this rearrangement of strokes include following steps.

1) The strokes identification as dependent and major dependent strokes.

2) The rearrangement of strokes with respect to their positions from y-axis after implementing step 1.
3) The combination of strokes to recognize character after implementing step 2.

The strokes identification as dependent strokes and major dependent strokes is the first step of rearrangement of strokes. The eight dependent strokes are considered and other strokes are considered as major dependent strokes. In step 2, the strokes are rearranged with respect to their positions from y-axis. Position of stroke has been considered as a low-level feature. This position includes four corners points of the strokes evaluated on the basis of their minimum and maximum values on the respective *x* and *y* axis. The points of the stroke with minimum and maximum values on *x* axis are referred as *xMin* and *xMax*. Similarly, the points of the stroke with minimum and maximum values on *y* axis are referred as *yMin* and *yMax*. These points are illustrated in Fig. 5. The strokes of the list are rearranged with respect to value of *xMin* of each stroke in ascending order. Therefore, the stroke with minimum value of *xMin* becomes the first stroke of list. The next stroke of list is considered with value of *xMin* more than first stroke and less than all other strokes of the list. Similarly, other strokes of the list are rearranged with their values of *xMin*. We have considered value of *xMin* as *xMax* in case of vowels or few characters that reside in upper or lower zones because of their nature of writing and positions with respect to neighboring strokes.

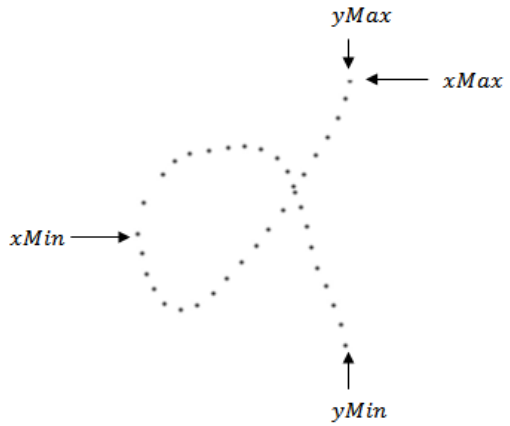

Fig. 5: A stroke with four corner points.

The characters are recognized in the sequential order of the strokes present in the list. As discussed, a character is combination of one or more strokes, therefore we have discussed following cases for characters recognition in word.

(i) A major dependent stroke with no suitable combination from neighboring stroke(s) directly recognizes a character.

(ii) Some of the characters can be recognized from the stroke list with neighboring stroke id as 12. This implies that if current stroke id is 12 and previous stroke is major dependent stroke with stroke id as 16 or 18 or 23 or 28 or 34 or 41 or 45, then it leads to recognition of various characters as: Stroke id 16 recognizes **ਅ**; stroke id 18 recognizes **ਸ, ਮ** or **ਸ਼**; Stroke id 23 recognizes **ਘ**; Stroke id 28 recognizes **ਜ** or **ਜ਼**; Stroke id 34 recognizes **ਠ**; Stroke id 41 recognizes **ਨ**; Stroke id 45 recognizes **ਯ**.

(iii) The characters as **ਪ, ਧ, ਖ, ਥ** and **ਖ਼** are recognized on the basis of information about headline (stroke id 11). This information includes position of headline, number of headline and the neighboring stroke(s).

(iv) The character **ਗ** is recognized with major dependent stroke id 22 and stroke id 12. It is worth mentioning here that length of stroke id 12 has been considered to recognize **ਗ**. If length of stroke id 12 is ignored, it could lead to recognize **ਰਾ**. The **ਗ** and **ਰਾ** are two different characters.

(v) The character **ੜ** is recognized with neighboring stroke id as 13 and major dependent stroke id is 38. The characters with similar shape to **ੜ** do not require neighboring stroke id as 13. These characters are **ਭ, ਤ** or **ਡ**.

(vi) The characters as **ਸ਼, ਮ, ਗ਼, ਜ਼, ਫ਼** or **ਲ਼** are recognized if the neighboring stroke is dot feature and form suitable combination with major dependent strokes.

(vii) The special characters are directly recognizes through their major dependent stroke(s). The special characters as ੱ and ੁ include the repetition of major dependent strokes.

(viii) The positions of the strokes with respect to neighboring strokes have been considered in recognition of characters.

(ix) The positions of major dependent strokes of special characters are considered with respect to neighboring strokes.

(x) The presence of more than one headline in the list has been considered separately with respect to position of each stroke.

The nature of above mentioned stroke ids are presented in Table 1.

4. Results and Discussion

Our focus is to test whether some target word shape can be recognized correctly or not. Therefore, current bench marking of our system is based on exact word recognition irrespective of number of characters in a

word recognized. An application in VC++ is developed that implement recognition procedure discussed in Section 3. The evaluation of proposed Gurmukhi word recognition system has been performed on 2576 dictionary words in terms of recognition accuracy. These 2576 words have been written by 11 writers and their results are presented in Table 2. We have noted that a total number of 2087 words have been recognized correctly and the overall recognition rate has been achieved as 81.02% for online cursive handwriting. The elastic matching method has been used as recognition technique in this experiment [11]. The present results can not be compared due to following reasons: database is new; unavailability of results for online handwritten Gurmukhi words in literature.

Gurmukhi is a popular script in northern India and shares many similarities to other asian scripts such as Devanagiri and Bangla. We believe that the present study will be helpful in recognition of such similar scripts. The present results motivate to study the present system with other recognition methods with more number of writers and dictionary words in near future. In addition, more number of cases can be proposed as discussed in Section 3.

Table 2. Recognition results of the system.

Writer ID	Number of Words Contributed	Number of Words Recognized
1	307	251
2	279	236
3	223	174
4	265	210
5	179	143
6	178	144
7	256	205
8	241	192
9	213	168
10	252	214
11	183	150

5. References

[1] Babu, V., Prasanth, L., Sharma, R., Rao, G. V., and Bharath, A. 2007. HMM-Based online handwriting recognition system for Telugu symbols. Proceedings of the ninth international conference on document analysis and recognition, vol.1, pp. 63-67.

[2] Bharat A. and Madhvanath, S., 2007. Hidden markov models for online handwritten tamil word recognition, Proceedings of international conference on document analysis and recognition, vol. 1, pp. 506-510.

[3] Bhattacharya, U., Gupta, B.K., Parui, S., 2007. Direction code based features for recognition of online handwritten characters of Bangla. Proceedings of international conference on document analysis and recognition, vol. 1, pp.58-62.

[4] Bhattacharya, U., Purui, S.K., Shaw, B. and Bhattacharya, K., 2006. Neural combination of ANN and HMM for handwritten Devanagari numeral recognition. Proceedings of ICFHR, 2006, pp. 613-618.

[5] Connell, S.D., Sinha, R.M.K. and Jain, A.K., 2000. Recognition of unconstrained on-line devanagari characters. Proceedings of international conference on pattern recognition, vol. 2, pp. 368-371.

[6] Elanwar, R.I., Rashwan, M.A., Mashali, S.A., 2007. Simultaneous segmentation and recognition of arabic characters in an unconstrained on-line cursive handwritten document. Proceedings of world academy of science, engineering and technology, vol. 23, pp. 288-291.

[7] Joshi, N., Sita, G., Ramakrishnan, A. G., Deepu, V., and Madhvanath, S. 2005. Machine recognition of online handwritten Devanagari characters. Proceedings of the eighth international conference on document analysis and recognition, pp. 1156-1160.

[8] Kumar, A., Balasubramanian, A., Namboodiri, A.M. and Jawahar, C.V., 2006. Model-Based annotation of online handwritten datasets. Proceedings of the international workshop on frontiers in handwriting recognition.

[9] Namboodiri, A.M., Narayanan, P.J. and Jawahar, C.V., 2007. On using classical poetry structure for Indian language post-processing. Proceedings of international conference on document analysis and recognition, pp. 1238-1242.

[10] Prasanth, L., Babu, V., Sharma, R., Rao, G. V., and M., D., 2007. Elastic matching of online handwritten tamil and telugu scripts using local features. Proceedings of the ninth international conference on document analysis and recognition, vol. 2, pp. 1028-1032.

[11] Sharma Anuj, R.K. Sharma and R. Kumar, 2008, "Recognizing Online Handwritten Gurmukhi Characters using Elastic Matching", IEEE proceedings of International Congress on Image and Signal Processing (CISP-2008), vol. 2, pp. 391-396.

[12] Sharma Anuj, R.K. Sharma and R. Kumar, 2009. "Online Handwritten Gurmukhi Strokes Preprocessing", International Journal of Machine Graphics and Vision, vol. 18, no. 1.