A Framework of Online Handwritten Gurmukhi Script Recognition

[1]Gurpreet Singh, [2]Manoj Sachan
[1]SLIET Longowal, Sangrur, Punjab, India
[2]Associate Prof., SLIET Longowal, Sangrur, Punjab, India

Abstract

The field of Pattern Recognition contributed up to a great extent in the Computer Vision and Machine Vision applications. Handwriting Recognition is one of the area under Pattern recognition. The current topic under handwriting recognition is Online mode of handwriting recognition. In this case user input the handwritten characters with the help of electronic devices like Digitizer Tablets and simultaneously a software system convert these handwritten characters to digital form. This recognition process totally depends on the complexity associated with the language in which text is written. This paper elaborates a system for Online handwriting recognition for Indian language "Punjabi". Worldwide Punjabi is the 10th most popular spoken language with around 102 million native speakers. This paper covers the overall Online recognition process used for handwritten text in Punjabi language using Gurmukhi script and also explain the complexities associated with Gurmukhi script.

Keywords

Recognition, Computer Vision, Machine Vision, Gurmukhi Script, Handwriting Recognition.

I. Introduction

The futuristic approach of Humans, open number of areas where research can be done. These researches are carried out to achieve the main goals like comfort, speed and accuracy. Handwriting recognition is also one of these research areas. It is the technique by which, the computer system can recognize characters and other symbols written by hand using natural handwriting. All this should be done to keep the records in the computer system in digital form (Singh & Sachan, 2014). These digital records can be used for future references. So, basically to interact with computer system or to exchange information with the computer, users have to input the required data into the system. For this particular input purpose, the devices like keyboard, mouse etc. are used. But these devices have some limitations also. Limitations are like, the slowness in typing speed of the user, user may not be familiar with the keyboard typing, Data may be required in different languages from the same keyboard etc. Input through natural handwriting is the alternative and fastest way to input data by any user to the system. The concept of using natural handwriting as an input technique comes under Online handwriting recognition process. Natural handwriting can be given as input to the computer system with the help of devices like Digitizer tablets, PDA (Personal Digital Assistants), Cross pads etc. These devices capture the information as; the number of strokes, direction of writing of each stroke, speed of writing etc. Where a stroke is the collection of points or (x, y) coordinates information covered by the stylus between Pen_Down event to Pen_Up events on the surface of digitizer (Sachan, Lehal & Jain, 2011). While making a system for Online handwriting recognition, the main consideration is always given to the concerned language. Every language has its own issues related to writing styles. This paper focuses on an Indian language "Punjabi". It is on 10th position in the list of most spoken languages across world. Its native speakers are around 102 million, which covers total 1.44% of the world population. Punjabi language is mainly spoken in India (Punjab region), Pakistan, United States, Canada, United Kingdom etc. In India, Punjabi languages is written with the help of Gurmukhi script and in Pakistan region, Punjabi language is written by using Shahamukhi script. This paper mainly deals with the Online handwriting recognition aspects of Gurmukhi script. This script has a character set of 41 consonants, 12 Vowels and 3 half characters. Its character set is shown in the fig. 1. In this paper a complete framework of Online Gurmukhi script is presented and this discussion also covers the issues related to each and every phase of the system under consideration.

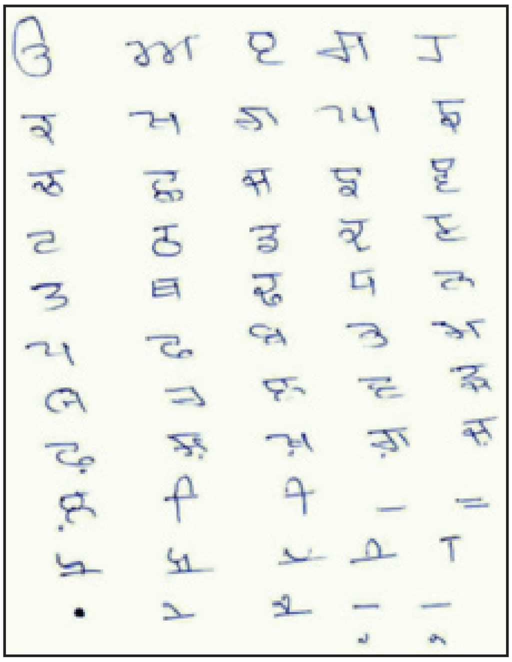

Fig. 1: Handwritten Character set of Gurmukhi Script

II. Related Work

A. Sharma et al 2008, presents the implementation of elastic matching technique to recognize online handwritten Gurmukhi characters. They recognize characters in two stages. First stage recognizes strokes and Second stage recognizes the characters. The database of strokes stores stroke number, script number etc. Authors obtain recognition rate of 90.08% by considering 60 writers and a set of 41Gurmukhi characters.

A. Sharma, R. Kumar & R. K. Sharma 2009, presents a system to recognize Online handwritten Gurmukhi words. Authors proposed a new step as rearrangement of recognized strokes. It includes stroke's identification as dependent and major dependent strokes, rearrangement of strokes with respect to the position, the combination of strokes to recognize characters. They achieve 81.02 % recognition rate for a set of 2576 Gurmukhi dictionary words.

A. Sharma et al 2010, presents a Hidden Markov Model (HMM) based Online handwritten characters recognition system for Gurmukhi Script. They test 60 handwritten samples, each sample includes 41 characters of Gurmukhi script. They observed the recognition rate of 91.95% and average recognition speed of 0.112 seconds per stroke.

M. Sachan et al 2011, presented a system for Online recognition of Gurmukhi script. The input of the user's handwriting is taken as a sequence of packets captured through the movement of stylus or pen on the surface of the tablet. The user's writing is segmented into meaningful shapes. Then segmented shapes are processed to extract the features. They used Nearest Neighbour classifier for recognition purpose. The average recognition accuracy achieved by the authors was 86.9% for complete Gurmukhi words.

A. Sharma & K. Dahiya 2012, presented a system to recognize Gurmukhi and Devanagiri characters in touch screen based mobile phones. For recognition, they use small line segments. Authors achieve overall recognition rate of 94.69% for Gurmukhi characters. For training set 5330 characters were included and for test set 1640 characters were included. They achieve 86.90% recognition rate for Devanagiri Script characters by considering training set with 1050 and test set with 504 characters.

R. Kumar & R.K. Sharma 2013, developed a post-processor for increasing the accuracy of character recognition of real time Gurmukhi script. They use the dataset consisting of 184 samples of 45 characters of Gurmukhi script collected from 4 different categories of writers. Authors proposed an algorithm to achieve the recognition accuracy of 95.60%.

M. Gupta, N. Gupta & R. Agarwal 2013, presented an implementation using SVM to recognize Online Gurmukhi handwriting. The pre-processing phase of the proposed system consists of 5 basic algorithms. A basic step of stroke capturing was done to sample data points along the trajectory of the input device. K-Fold technique was used for recognition phase. They considered 100 words from three different writers. Each word was subdivided into strokes and each stroke was given a unique id. Then these strokes were further divided into three zones. First writer wrote 810 strokes, second writer used 747 strokes for writing and third used 875 strokes. Experiment concluded that third writer's handwriting was better for the recognition purpose which consist more strokes.

III. Framework for Gurmukhi Script Recognition

In case of Online handwriting recognition systems, handwritten documents are recognized while being written. These systems are either writer dependent or writer independent. In case of writer dependent recognition systems, some information about the writer's handwriting is available with the system being used for matching or recognition purpose. But in case of writer independent systems, no such prior information about the writer's handwriting is given to the system. So it is bit harder system for recognition point of view. In this paper an Online handwriting recognition system is explained, which follow the steps like, Data collection, Pre-processing, Segmentation & Feature extraction, Classification and Post-processing. The next part of the paper will explain these steps in a sequential manner as present in fig. 2.

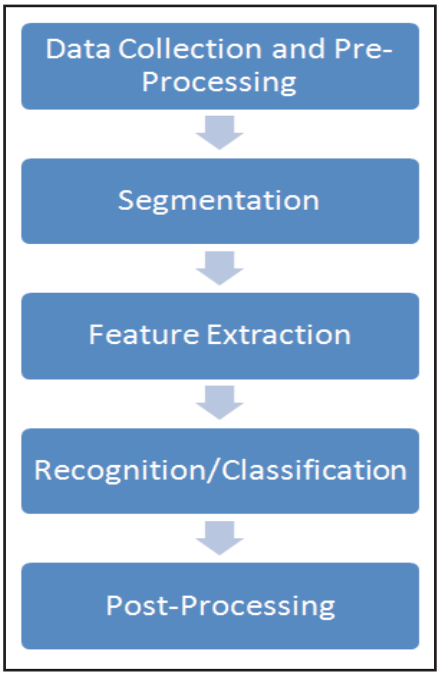
Data Collection and Pre-Processing
Segmentation
Feature Extraction
Recognition/Classification
Post-Processing

Fig. 2: Online Handwriting Recognition System

A. Data Collection

In Online handwriting recognition systems, data or handwriting can be captured with the help of the devices like Digitizer tablets, Tablet PCs, Personal Data Assistants (PDA) etc. Stylus is used by the user to write handwriting samples on the surface of digitizer. Fig. 3. shows some of the handwriting capturing devices. During data collection phase, sequence of coordinate points are recorded for further recognition of handwritten symbols. The set of these coordinate points create a Stroke. The collection of these coordinate points started from the Pen_Down event of the stylus on the surface of digitizer and ended when Pen_Up event generated. In between, the recorded sequence of points by the moving pen form a single Stroke. A single character or symbol, belongs to the language which is used by the user for writing, may consists of one or more than one strokes. Fig. 4. shows handwritten character "Jajha" of Gurmukhi script. This character is written with the help of three different strokes as mentioned in fig. 4.

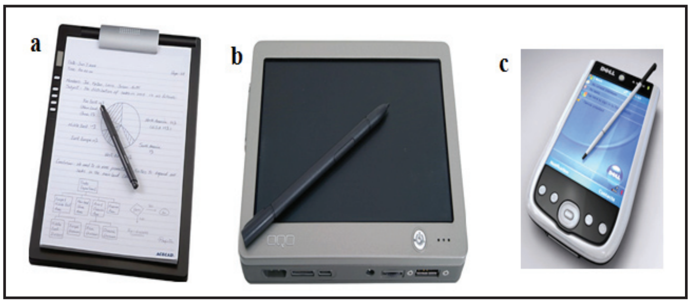

Fig. 3: Devices to Capture Online handwriting; (a) Digimemo; (b) Tablet PC; (c) PDA

B. Pre-processing Phase

In case of Online handwriting capturing, hardware devices as well as software interface are used to make an effective system. But these devices and interfaces also have some limitations, which produce some noise in the captured data (Pesch, Hamdani & Forster (2012). The main reasons behind this noise may be the speed of writing of the user which cause some missing points, when the speed is very high; sharp edges etc. The presence of noise may affect the recognition rate. To deal with the noise factor or to improve the recognition rate, pre processing steps are necessary. In case of Online Gurmukhi script, recognition following preprocessing steps may results in improved recognition rate.

- Normalization of Size and Centering of strokes
- Identification of missing points
- Strokes smoothing
- Re-sampling of points

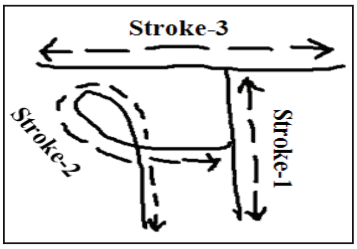

Fig. 4: Handwritten Character “Jajha” of Gurmukhi Script with Three Strokes

1. Normalization of size and Centering of strokes

Size of the stroke depends on the handwriting style of the user. It means totally depends on the movement of stylus on the surface of digitizer by the user. This step is necessary because the pen movement on the border of digitizer’s surface are not captured by the device, this results in loss of information. Following algorithm is used for the above said purpose of centering of strokes.

Algorithm 1: Centering of strokes

```
Step-1: Set $X_{len}$=256 and $Y_{len}$=256
Step-2: $P_{nx=}P_{nx} * \frac{x_l}{X_{len}}$ , $P_{ny=}P_{ny} * \frac{Y_l}{Y_{len}}$
Step 3: $P_{nx} = P_{nx} \pm x_0$ , $P_{ny} = P_{ny} \pm y_0$
```

Here (x_0, y_0), is the origin of frame of reference, x_l and y_l are length in x and y directions respectively. Pnx and Pny are new positions of the stroke points present on the border of the digitizer’s surface.

2. Identification of missing points

This method of pre-processing is used to find those points on the surface of digitizer, which the user missed while written in a high speed. These missing points can be captured with the help of Bezier curves or B-Spline techniques. This information also improves the recognition rate. For locating these missing points consecutive set of four points have to be considered for obtaining Bezier Curve. These points are considered in the algorithm as C_1, C_2, C_3 and C_4. Following algorithm describe the procedure to locate missing points with the help of Bezier curve:

Algorithm 2: Missing points identification

```
Step-1: Assume S as a variable.
Step-2: Set S=0.2 and $\Delta S = 0.2$
Step-3: Repeat Step 4 and 5 until $S \leq 1$
Step-4: $X_{new} = C_{1_x} \times (1-S)^3 + C_{2_x} \times 3 \times S \times (1-S)^2 + C_{3_x} \times 3 \times S^2 \times (1-S) + C_{4_x} \times S^3$
        $Y_{new} = C_{1_y} \times (1-S)^3 + C_{2_y} \times 3 \times S \times (1-S)^2 + C_{3_y} \times 3 \times S^2 \times (1-S) + C_{4_y} \times S^3$
Step-5: Set $S = S + \Delta S$
Step-6:Return
```

3. Stroke Smoothing

To avoid the flickers exists in individual handwriting style, K-neighbors technique is used at pre-processing stage. This also improves the overall recognition rate. This algorithm considered the new identified missing point with the help of other points like C_{i-2}, C_{i-1}, C_{i+1} and C_{i+2} for smoothing purpose. Fig. 5 shows the identification of this new point.

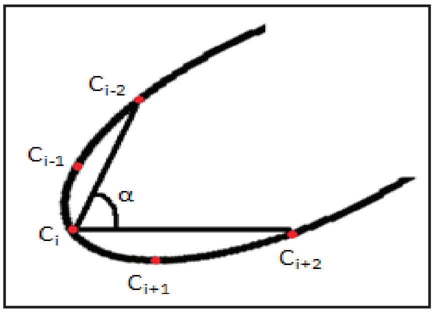

Fig. 5: Formation of Angle at point C_i

Algorithm 3: Stroke Smoothing

```
Step-1: Calculate K as total number of points in current stroke
Step-2: Repeat Steps 3 and 4 for $C_i$, i=3,4,_ _ _, m-2.
Step-3: Calculate $\alpha = \text{Angle}(C_{i-2}, C_i, C_{i+2})$
Step-4:Set $C_{ix} = \left(C_{(i-2)x} + C_{(i-1)x} + (\alpha \times C_{ix}) + C_{(i+1)x} + C_{(i+2)x}/2 \times 2 + \alpha\right)$
          $C_{iy} = \left(C_{(i-2)y} + C_{(i-1)y} + \left(\alpha \times C_{iy}\right) + C_{(i+1)y} + C_{(i+2)y}/2 \times 2 + \alpha\right)$
```

4. Re-sampling of Points

To place the points in a stroke at equal distance to each other for further recognition, Re-sampling of points technique is used. This algorithm focuses on the issue that if in between two points, there is a possibility to fill some more points for proper recognition then those points can be filled or considered with the help of the same Bezier function procedure which is used in case of missing points algorithm.

C. Segmentation

After the data input process of data collection phase of Online handwriting recognition, the next important process is to identify the data or understand the data at character level or at stroke level. The segmentation algorithms contain the logic related to the complexity of language in which the handwriting sample is given by the user. Segmentation process is divided into two tasks, Internal segmentation and External segmentation. The segmentation process performed before the recognition process is known as internal segmentation and segmentation done during recognition process is called external segmentation. As mentioned earlier part of this paper, the large character set of Gurmukhi language shows its complex behaviour in concern to the segmentation. The words in Gurmukhi script are written from left to right direction and a complete word of Gurmukhi script may contain more than one character. These characters are joined together with a line at the top of the characters. This line is called “Headline”. So the presence of headline again increases the complexity of the segmentation process as, at the time of segmentation we have to separate characters from each other for further recognition. Another aspect or the difficulty which is related to segmentation process in case of Gurmukhi script recognition is that, a complete word of Gurmukhi script is divided into three zones, First is the upper zone, where those symbols appeared which are present above the headline, mostly these are from vowels. The second zone is the middle zone, mostly the consonants and some vowels appeared in this zone and Third is the lower zone, which appeared at the bottom of the characters. It usually contain half characters or some vowels. So the segmentation process is done in case of Gurmukhi script recognition by considering the following points:

- Identification of the location of headline
- Identification of Upper, Middle and Lower zones
- Identification of different strokes belong to a single character

D. Feature Extraction

To analyze the input data or to identify the different meaningful patterns, feature extraction is used. Features are classified as Local features or Low-level features and Global features or High-level features. Local features focuses on direction, position, slope, area etc. In case of Global features, we consider Headline, Straight line, Dots etc. The broader categories of features are considered as Statistical features, Series expansions coefficients and Structural features (Chowdhury, Garain & Chottopadhyay, 2011). Statistical features deals with Zoning, Direction code histograms etc. Series expansion coefficients deals with coefficients of infinite series. The structural features try to find out about the information or idea about the shape of the pattern for further classification.

E. Classification

The main aim of classification phase is to identify the class of pattern extracted earlier. This can be done by comparing the extracted patterns with the already existing pattern in the database, according to some classification model. The different classification approaches are:

- Structural and Rule based methods
- Statistical classification methods

1. Structural and Rule Based Methods

In rule based methods, a test pattern is matched with the structural model of each candidate class on the basis of minimum matching distance. Structural techniques consists, Decision tree matching, Dynamic programming, Elastic matching etc. The main advantage of these methods are the fastest classification time, less training data etc.

2. Statistical Classification Methods

This approach consists of representation of a pattern as an ordered fixed length list of numerical values. The major techniques used in this category of recognition approach are Artificial Neural Network (ANN), Hidden Markov Model (HMM), Support Vector Machine (SVM), Convolution Time Delay Neural Network (TDNN) etc. The main advantage of these models are, they model the temporal relationship well and classification time is fast as compared to structural algorithms. Table 1. shows some work done in the field of Online Gurmukhi script recognition in recent times.

IV. Conclusion

The main goal of this paper was to elaborate the Online handwriting recognition procedure for Gurmukhi script. The complexities related to Gurmukhi script made the recognition task challenging. Different phases of online recognition process were explained like, Data collection in the form of strokes, Pre-processing stages to improve the quality of input data for further recognition. The algorithms like, Size normalization & centering of strokes, Missing points identification, Stoke smoothing, Re-sampling of points were discussed for Pre-processing phase. Segmentation and Feature extraction phases were discussed with complexity issues like, identification of the location of Headline, identification of different zones etc. In the classification phase the pros and cons of Structural & rule based methods and Statistical classification methods like TDNN, HMM and SVM were discussed. Then the emphasis was also given to Post-processing phase, which helps to recognize the misclassified results by using the linguistic knowledge. The future aspects in this field like, the use of hybrid of classifiers such as TDNN, HMM and SVM etc. can improve the performance. Also some factors related to the hardware devices like, Pressure of strokes while writing on the surface of digitizer, Temporal information etc. also helps to increase the overall recognition rates. At complete word level recognition, maximum of about 87% recognition rate shows the scope for researchers in this area.

Table 1: Current Status of Recognition of Online Handwritten Gurmukhi Script

S.No.	Authors	Method	Script Used	Recognition % age
1.	A. Sharma et al. (2008)	Elastic string matching	Gurmukhi/ Character level	90.08%
2.	A. Sharma et al. (2009)	Rearrangement of Strokes	Gurmukhi/ Word level	81.02%
3.	A. Sharma et al. (2010)	HMM	Gurmukhi/ Character level	91.95%
4.	M. Sachan et al. (2011)	Nearest Neighbour /SVM	Gurmukhi/ Word level	86.90%
5.	K. Dahiya et al. (2012)	Smallest Line Segment	Gurmukhi/ Character level	94.69%
6.	M. Gupta et al. (2013)	SVM	Gurmukhi/ Word level	---------

References

[1] Sharma A, Kumar R, Sharma RK.,"Recognizing Online Handwritten Gurmukhi Characters using Elastic Matching", IEEE proceedings of International Congress on Image and Signal Processing.2008, Vol. 2, pp. 391-396.

[2] Sharma A, Kumar R, Sharma, R K.,"Rearrangement of Strokes in Recognition of Online Handwritten Gurmukhi Words", IEEE Proceedings of 10th International Conference on Document Analysis and Recognition, Barcelona, Spain (ICDAR). 2009, pp. 1241-1245.

[3] Sachan MK, Lehal GS, Jain VK,"A novel method to segment online gurmukhi script", Proceedings of International Conference on Information Systems for Indian Languages ICISIL, 2011. Vol. 139, pp. 1-8.

[4] Sachan MK, Lehal GS, Jain VK, A System for Online Gurmukhi Script Recognition", Proceedings of International Conference on Information Systems for Indian Languages ICISIL. 2011, Vol. 139, pp. 294-295.

[5] Sharma A, Dahiya K.,"Online handwriting recognition of gurmukhi and devanagri characters in mobile phone devices", IJCA proceedings of International conference of recent advances and future trends in information technology. 2012, pp. 201-205.

[6] Sharma A, Kumar R, Sharma RK,"HMM based online handwritten gurmukhi character recognition. ACM digital library machine graphics and vision international journal. 2010,Vol.19, pp. 439-449.

[7]. Gupta M, Gupta N, Aggarwal R.,"Recognition of online gurmukhi handwriting using SVM approach", International conference of Bio-inspired computing theories and applications. 2013, pp. 495-506.

[8] Kumar R, Sharma RK,"An efficient post-processing algorithm for online handwritten gurmukhi character recognition using set theory", International journal of pattern recognition and artificial intelligence. 2013, Vol. 27, pp. 270-275.

[9] Chowdhury S, Garain U, Chottopadhyay T,"A weighted finite-state transducer (WFST) based language model for online Indic script handwriting recognition", IEEE International conference on document analysis and recognition. 2011, pp. 599-602.

[10] Pesch H, Hamdani M, Forster J, Ney H,"Analysis of pre-processing techniques for latin handwriting recognition", IEEE international conference on frontiers in handwriting recognition, 201, pp. 280-284.

[11] Singh G, Sachan M, Multi-layer perceptorn (MLP) neural network technique for offline handwritten gurmukhi character recognition, IEEE International conference on computational intelligence and computing research. 2014, pp. 221-225.

Gurpreet Singh received B.Tech. Degree in Computer Science & Engineering from Punjab Technical University, Jalandhar in 2007 and M.Tech Degree in Computer Science & Engineering from BBSBEC in 2011. He worked as Lecturer in the Department of Computer Science & Engineering, IET, Bhaddal from 2007 to 2010. He worked as Assistant Professor in the Department of Computer Science & Engineering, IET, Bhaddal from 2010 to 2014. Currently, he is a Research Scholar in CSE Department at SLIET, Longowal. His research areas include Digital Image Processing and Handwriting recognition. He is a lifetime member of Indian Society for Technical Education.

Manoj Sachan received B.Tech. degree in Computer Science from TIET, Patiala; ME degree from Punjabi University Patiala and Ph.D. degree in Computer Science & Engineering from Punjab Technical University, Jalandhar. He worked as Lecturer in the Department of Computer Science & Engineering, TIET, Patiala from 1992 to 1999. Then he worked as Assistant Professor in the Department of Computer Science & Engineering, SLIET, Longowal from 1999 to 2006. Presently he is working as Associate Professor in Computer Science & Engineering department, SLIET, Longowal since 2006. His research areas include Pattern recognition, Neural Networks and Online Gurmukhi Script Recognition. He is a lifetime member of Indian Society for Technical Education (ISTE).