k -Nearest Neighbor Based Offline Handwritten Gurmukhi Character Recognition

Munish Kumar
Computer Science Department
Panjab University Constituent College
Muktsar, Punjab, India
munishcse@gmail.com

M. K. Jindal
Department of Computer Science and Applications
Panjab University Regional Centre
Muktsar, Punjab, India
manishphd@rediffmail.com

R. K. Sharma
School of Mathematics and Computer Applications
Thapar University
Patiala, Punjab, India
rksharma@thapar.edu

Abstract— **Offline handwritten character recognition has been a frontier area of research for the last few decades under pattern recognition. Recognition of handwritten characters is a difficult task owing to various writing styles of individuals. A scheme for offline handwritten Gurmukhi character recognition based on *k*-NN classifier is presented in this paper. The system first prepares a skeleton of the character, so that feature information about the character is extracted. There is abundant literature on the handwriting recognition on non-Indian scripts, but there are very few article available related to recognition of Indian scripts such as Gurmukhi. This paper presents an efficient offline handwritten Gurmukhi character recognition system based on diagonal features and transitions features using *k*-NN classifier. Diagonal and transitions features of a character have been computed based on distribution of points on the bitmap image of character. In *k*-NN method, the Euclidean distance between testing point and reference points is calculated in order to find the *k*-nearest neighbors. In this work, we have taken the samples of offline handwritten Gurmukhi characters from one hundred different writers. The partition strategy for selecting the training and testing patterns has also been experimented in this work. We have used in all 3500 images of Gurmukhi characters for the purpose of training and testing. The proposed system achieves a maximum recognition accuracy of 94.12% using diagonal features and *k*-NN classifier.**

Keywords-Handwritten character recognition; diagonal features; transistion features; classificationl; k-NN

I. INTRODUCTION

Handwritten Character Recognition, usually abbreviated as HCR, is the process of converting handwritten text into machine processable format. Handwritten character recognition has been one of the most important and challenging research areas in the field of pattern recognition. It contributes immensely to the advancement of an automation process and can improve the interface between human and machine in various applications. Most of the published work on Indian scripts recognition deals with printed documents and very few articles deal with handwritten script problem. HCR can be online or offline. In online handwriting recognition, data are captured during the writing process with the help of a special pen and an electronic surface. Offline documents are scanned images of prewritten text, generally on a sheet of paper. Offline handwriting recognition is significantly different from online handwriting recognition, because here, stroke information is not available [1, 2]. In this work, we have proposed a recognition system for offline handwritten Gurmukhi characters. In this work, we have proposed a recognition system for offline handwritten Gurmukhi characters. A recognition system consists of the activities, namely, digitization, pre-processing, features extraction and classification. These activities in such a system have a close proximity with printed characters recognition system. For example, a printed Gurmukhi script recognition system has been proposed by Lehal and Singh [3]. Wen *et al.* [4] have proposed handwritten Bangla numerals recognition system for automatic letter sorting machine. Swethalakshmi *et al.* [5] have proposed handwritten Devanagri and Telugu character recognition system using SVM. The input to their recognition system consists of features of the stroke information in each character and SVM based stroke information module has been considered for generalization capability. Pal *et al.* [6, 7] have presented a technique for off-line Bangla handwritten compound characters recognition. They have used modified quadratic discriminant function for feature extraction. Pal *et al.* [8] have also used curvature feature for recognizing Oriya characters. Hanmandlu *et al.* [9] have reported grid based features for handwritten Hindi numerals. They have divided the input image into 24 zones. After that, they have computed the vector distance for each pixel position in the grid from the bottom left corner and normalized these distances to [0, 1] in order to obtain the features.

II. GURMUKHI SCRIPT

Gurmukhi script is the script used for writing Punjabi language and is derived from the old Punjabi term "Guramukhi", which means "from the mouth of the Guru". Gurmukhi script has three vowel bearers, thirty two consonants, six additional consonants, nine vowel modifiers, three auxiliary signs and three half characters. Gurmukhi script is 12th most widely used script in the world. Writing style of Gurmukhi script is from top to bottom and left to right. In Gurmukhi script, there is no case sensitivity. The character set of Gurmukhi script is given in Figure 1. In Gurmukhi script, most of the characters have a horizontal line at the upper part called headline and characters are connected with each other through this line. In Gurmukhi script some characters are quite similar to each other, which make their recognition a bit complex.

978-1-61284-861-7/11/$26.00 ©2011 IEEE

The Consonants

ਸ ਹ ਕ ਖ ਗ ਘ ਙ ਚ ਛ ਜ ਝ ਞ ਟ ਠ ਡ

ਢ ਣ ਤ ਥ ਦ ਧ ਨ ਪ ਫ ਬ ਭ ਮ ਯ ਰ ਲ ਵ ੜ

The Vowel Bearers

ਉ ਅ ਏ

The Additional Consonants (Multi Component Characters)

ਸ਼ ਜ਼ ਖ਼ ਫ਼ ਗ਼ ਲ਼

The Vowel Modifiers

ੇ ੈ ੋ ੌ ਿ ੀ ਾ ੁ ੂ

Auxiliary Signs

ੱ ੰ ਂ

The Half Characters

ੑ ੍ਰ ੍ਵ

Figure 1. Gurmukhi script character set.

For this work, a sample of 100 writers was selected from schools, colleges, government offices and other places. These writers were requested to write each Gurmukhi character. A sample of five handwritten Gurmukhi characters by five different writers (W1, W2, …, W5) is given in Figure 2.

Script Character	W1	W2	W3	W4	W5
ਉ					
ਅ					
ਏ					
ਸ					
ਹ					

Figure 2. Samples of handwritten Gurmukhi characters.

III. THE PROPOSED RECOGNITION SYSTEM

The proposed recognition system consists of the phases, namely, digitization, pre-processing, feature extraction and classification. The block diagram of proposed recognition system is given in Figure 3.

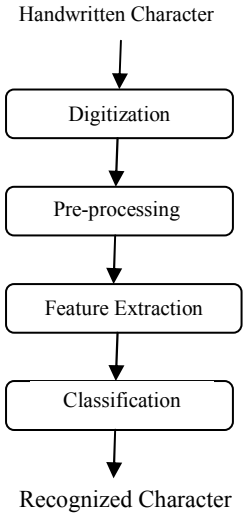


Figure 3: Block diagram of handwriting grading system.

A. Digitization

Digitization is the process of converting the paper based handwritten document into electronic form. The electronic conversion is accomplished using a process whereby a document is scanned and an electronic representation of the original, in the form of a bitmap image, is produced. Digitization produces the digital image, which is fed to the pre-processing phase.

B. Pre-processing

Pre-processing is a series of operations performed on the digital image. Pre-processing is the initial stage of character recognition. In this phase, the character image is normalized into a window of size 100×100. After normalization, we produce bitmap image of normalized image. Now, the bitmap image is transformed into a contour image. This process of pre-processing is described in Figure 4 (a) and (b) for Gurmukhi character ਕ .

Figure 4. (a) Digitized image of Gurmukhi character (ਕ) (b) Contour image of Gurmukhi character (ਕ).

C. Feature extraction

In this phase, the features of input characters are extracted. The performance of handwriting recognizer depends on features, which are being extracted. The extracted features

should be able to classify each character uniquely. We have used diagonal and transition features for recognition of offline Gurmukhi handwritten character. Here, we first transform the input character image into contour image as shown in Figure 4(a) and Figure 4(b).

1) *Diagonal feature extraction*

Diagonal features are very important features in order to achieve higher recognition accuracy and reducing misclassification. These features are extracted from the pixels of each zone by moving along its diagonals as shown in Figure 5.

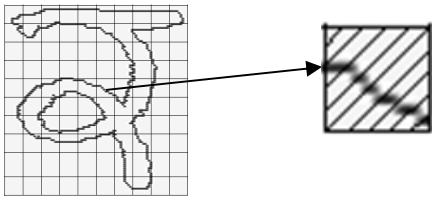


Figure 5. Diagonal feature extraction.

The steps that have been used to extract these features are given below.

Step I: Divide the input image into *n* (=100) number of zones, each of size 10×10 pixels.

Step II: The features are extracted from the pixels of each zone by moving along its diagonals.

Step III: Each zone has 19 diagonals; foreground pixels present along each diagonal is summed up in order to get a single sub-feature.

Step IV: These 19 sub-features values are averaged to form a single value and placed in corresponding zone as its feature.

Step V: Corresponding to the zones whose diagonals do not have a foreground pixel, the feature value is taken as zero.

Using this algorithm, we will obtain *n* features corresponding to every zone.

2) *Transition feature extraction*

The second feature extraction investigated in this research was based on the calculation and location of transition features from background to foreground pixels in the vertical and horizontal directions. The transition feature used here is similar to that proposed by Gader et al [13]. To calculate transition information, image is scanned from left to right and top to bottom.
Following steps have been implemented for extracting these features.

Step I: Divide the skeletonized image of a character into *n* (=100) zones, each of size 10×10.
Step II: Calculate number of transitions for each zone.
This will give us *n* features for a character image.

IV. CLASSIFICATION

Classification stage uses the features extracted in the feature extraction stage for deciding the class membership. Classification phase is the decision making phase of an HCR engine. In this work, we have used *k*-NN classifier for recognition. In the *k*-nearest neighbor classifier, Euclidean distances from the candidate vector to stored vector are computed. The Euclidean distance between a candidate vector and a stored vector is given by,

$$d = \sqrt{\sum_{k=1}^{N} (x_k - y_k)^2}$$

Here, *N* is the total number of features in feature set, x_k is the library stored feature value and y_k is the candidate feature value.

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the results of recognition system for offline handwritten Gurmukhi characters are presented. The results are based on two feature extraction techniques, namely, diagonal and transitions features. As sated earlier, we have also experimented some partitioning strategies. We have divided the data set using five partitioning strategies. In the first strategy (strategy *a*), we have taken 50% data in training set and other 50% data in the testing set. In the second strategy (strategy *b*), we have considered 60% data in training set and remaining 40% data in the testing set. Strategy *c* has 70% data in training set and 30% data in testing set. Similarly, strategy *d* has 80% data in training set and 20% in testing set. Strategy *e* is formulated by taking 90% data in training set and remaining 10% data in testing set. Feature-wise experimental results of testing are presented in the following sub-sections.

A. *Diagonal features*

In this section, the diagonal features have been considered to be taken as input to *k*-NN classifier. In this section, we have presented recognition results of five partitioning strategies *(a, b, c, d and e)* based on the diagonal features using *k*-NN classifier. Using this approach, i.e., diagonal features and *k*-NN, we have achieved an accuracy of 86.17% when we use strategy *a* and achieved an accuracy of 94.12% when we used the strategy *e*. These results are depicted in Figure 6.

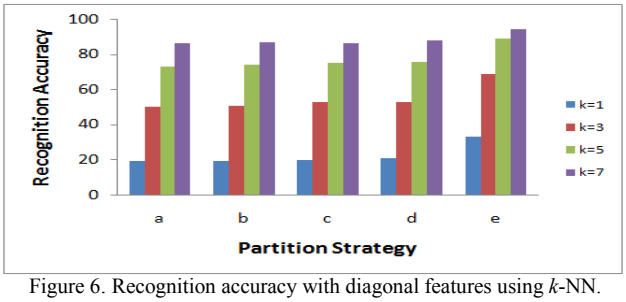
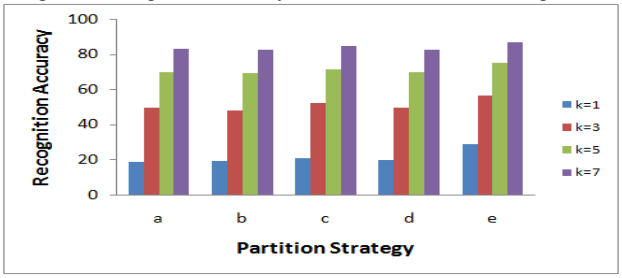


Figure 6. Recognition accuracy with diagonal features using *k*-NN.

B. *Transition features*

In this subsection, the transitions features have been considered for inputting the *k*-NN classifier. For the features under consideration and the *k*-NN classifier, the minimum accuracy achieved is 82.43% in partitioning strategy *b* and the maximum accuracy achieved is 86.57% in partitioning strategy *e*. The results for this case are depicted in Figure 7.

Figure 7. Recognition accuracy with Transition features using *k*-NN.



VI. CONCLUSION

The work presented in this paper proposes an offline handwritten Gurmukhi character recognition system. The features of a character that have been considered in this work include diagonal features and transition features. The classifier that has been employed in this work is *k*-NN. The maximum recognition accuracy of 94.12% is achieved in this work for the case when we input the diagonal features to the *k*-NN classifier. This accuracy can probably be increased by considering a larger data set while training the classifier. This work can also be extended for offline handwritten character recognition of other Indian scripts. The results of recognition accuracy with these features are depicted in Table 1.

Table 1. Performance of recognition accuracy with diagonal and transition features strategy wise.

Feature Type	Strategy	*k*=1	*k*=3	*k*=5	*k*=7
Diagonal	*a*	19.34	50.11	72.97	86.17
Diagonal	*b*	19.07	50.63	74.07	86.79
Diagonal	*c*	19.90	52.56	75.14	86.19
Diagonal	*d*	21	52.71	75.71	87.86
Diagonal	*e*	33.14	68.86	88.86	**94.12**
Transition	*a*	18.57	49.71	69.54	83.14
Transition	*b*	19	47.8	69.25	82.43
Transition	*c*	20.57	52.19	71.16	84.67
Transition	*d*	19.43	49.29	69.43	82.43
Transition	*e*	28.86	56.57	74.86	**86.57**

REFERENCES

[1] L. M. Lorigo and V. Govindaraju, "Offline Arabic handwriting recognition: a survey", *IEEE Transactions on PAMI,* Vol. 28(5), pp. 712-724, 2006.

[2] R. Plamondon and S. N. Srihari, "On-line and off- line handwritten character recognition: A comprehensive survey", *IEEE Transactions on PAMI,* Vol. 22 (1), pp. 63-84, 2000.

[3] G. S. Lehal and C. Singh, "A Gurmukhi script recognition system", in the proceedings of 15th ICPR, Vol. 2, pp. 557-560, 2000.

[4] Y. Wen, Y. Lu and P. Shi, "Handwritten Bangla numeral recognition system and its application to postal automation", *Pattern Recognition,* Vol. 40, pp. 99-107, 2007.

[5] H. Swethalakshmi, A. Jayaraman, V. S. Chakravarthy and C. Sekhar, "Online handwritten character recognition of Devanagari and Telugu characters using support vector machine", in the proceedings of 10th IWFHR, pp. 367-372, 2006.

[6] U. Pal, T. Wakabayashi and F. Kimura, "Handwritten Bangla Compound Character Recognition using Gradient Feature", in the proceedings of 10th ICIT, pp. 208- 213, 2007.

[7] U. Pal, T. Wakabayashi and F. Kimura, "Handwritten numeral recognition of six popular scripts", in the proceedings of ICDAR 07, Vol.2, pp.749-753, 2007.

[8] U. Pal, T. Wakabayashi and F. Kimura, "A system for off-line Oriya handwritten character recognition using curvature feature", in the proceedings of 10th ICIT, pp. 227-229, 2007.

[9] M. Hanmandlu, J. Grover, V. K. Madasu, and S. Vasikarla "Input fuzzy for the recognition of handwritten Hindi numeral", in the proceedings of ITNG, pp. 208-213, 2007.

[10] S. V. Rajashekararadhya and S. V. Ranjan, "Zone based Feature Extraction algorithm for Handwritten Numeral Recognition of Kannada Script", in the proceedings of IACC, pp. 525-528, 2009.

[11] J. Tripathy, "Reconstruction of Oriya alphabets using Zernike Moments", *International Journal of Computer Applications*, Vol. 8 (8), pp. 26-32, 2010.

[12] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition", in the proceedings of IEEE, Vol. 77 , pp. 257-286, 1989.

[13] P. D. Gader, M. Mohamed and J. H. Chaing, "Handwritten word recognition with character and inter-character neural networks", *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics,* Vol. 27 (1), pp. 158-164, 1997.

[14] M. K. Jindal, "Degraded Text Recognition of Gurmukhi Script", PhD Thesis, Thapar University, Patiala, India, 2008

[15] V. Shapiro, G. Gluhchev, V. Sgurev, "Handwritten document image segmentation and analysis", *Pattern Recognition,* Letters archive, Vol. 14 (1), pp. 71-78, 1993.

[16] S. Mori, K. Yamamoto, and C. Y. Suen, "Historical review of OCR research and development", in the proceedings of the IEEE, Vol. 80 (7), pp. 1029-1058, 1992.

[17] M. K. Jindal, G. S. Lehal and R. K. Sharma, "On Segmentation of touching characters and overlapping lines in degraded printed Gurmukhi script", *International Journal of Image and Graphics (IJIG)*, World Scientific Publishing Company, Vol. 9(3), pp. 321-353, 2009.

[18] K. Wong, R. Casey and F. Wahl, "Document Analysis System", *IBM Journal of Research Development*, Vol. 26(6), pp. 647-656, 1982.

[19] V. Bansal and R. M. K. Sinha, "Segmentation of touching and fused Devanagri characters", *International Journal of Pattern Recognition*, Vol. 35(4), pp. 875-893, 2002.