Size of Training Set Vis-à-vis Recognition Accuracy of Handwritten Character Recognition System

Munish Kumar
Computer Science Department
Panjab University Constituent College
Muktsar, Punjab, India

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

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

***Abstract*—Support Vector Machines (SVMs) have successfully been used for character recognition. In the present study, we have shown how the recognition accuracy of a SVM classifier varies with variation in the training set size. The training set for this work is taken from samples of offline handwritten Gurmukhi characters. For recognition of a handwritten Gurmukhi character, we have used curvature features extracted from the skeletonized image of each Gurmukhi character. Features of a character have been computed based on statistical measures of distribution of points on the bitmap image of character. To extract these features, the image of each Gurmukhi character is first segmented into few zones and then the curvature shape is computed within each of these zones. Considering all the zones, a feature set is formed for representation of each image pattern and a database of 3500 isolated handwritten Gurmukhi characters has been used for the same. The results of investigation presented in this paper show that the size of training set has a significant effect on the accuracy of offline handwritten Gurmukhi script recognition system.**

***Index Terms*—Feature extraction; curve fitting; handwritten character recognition; SVM.**

I Introduction

Handwritten Character Recognition, usually abbreviated as HCR, is the process of converting handwritten text into machine processable format. HCR is the field of research in pattern recognition and artificial intelligence. Handwriting recognition provides a methodology for improving the interface between user and computer as it enables computers to read and process handwritten documents which are currently being processed manually. A good number of researchers have already worked on the recognition problem of offline printed characters. 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] have presented a technique for off-line Bangla handwritten compound characters recognition. They have used modified quadratic discriminant function for feature extraction. They have also used curvature features for recognizing Oriya characters. Chaudhary and Pal [7] have proposed recognition system for two Indian scripts, Bangla and Devanagari. They have used tree classifier for character recognition. Hanmandlu et al. [8] have reported grid based features for handwritten Hindi numerals recognition. They have divided the input image into 24 zones. After that, they compute the vector distance for each pixel position in the grid from the bottom left corner and normalize these distances to [0, 1] in order to obtain the features. Bansal and Sinha [9] have provided a complete OCR system for printed Devanagari script. Kumar [10] has proposed a technique for recognition of handwritten Devanagri characters. He has used an AI approach to integrate information from sources and a fuzzy logic concept to handle uncertainties and imprecise information. In order to tackle the problem related to selection of a proper dataset for training a SVM, different strategies have been considered in this work. Chaudhury et al. [11] has been presented a scheme using a syntactic method for connected Bangla handwritten numerals recognition. In 2006, Roy and Pal have presented an automatic scheme, for word-wise identification of hand-written Roman and Oriya scripts for Indian postal automation [12]. In 2008, Sharma et al. [13] have proposed a system based on elastic matching for online Gurmukhi script recognition. Here, we have analysed the recognition performance of the SVM with variations in the training set size. We have used parabola curve and power curve based features for representation of handwritten Gurmukhi characters in feature space. In doing so, a skeletonized image of handwritten Gurmukhi character is segmented into the zones of equal size and the shape of curve in each zone is determined. This shape defines the features of the zone. This paper is organized into six sections. Introduction to Gurmukhi script and data collection for this work is described in Section 2.

doi:10.4304/jetwi.5.4.380-384

Section 3 presents a method of feature extraction for handwritten character recognition system. Classification is described in Section 4. Section 5 shows some experimental results to prove the usefulness of this approach. Conclusions are finally included in Section 6.

II Gurmukhi Script and Data Collection

Gurmukhi script is the script used for writing Punjabi language. The word Gurmukhi has been derived from the Punjabi term "Guramukhi", which means "from the mouth of the Guru". Gurmukhi script is the 12th most widely used script in the world. Gurmukhi script has three vowel bearers, thirty two consonants, six additional consonants, nine vowel modifiers, three auxiliary signs and three half characters. The character set of Gurmukhi script is given in Figure 1. For the present work, we have collected data from 100 different writers. These writers were requested to write each Gurmukhi character. All these characters are scanned at 300 dpi resolution with HP-1400 scanner A sample of handwritten characters by 5 different writers (W1, W2, …, W5) is given in Figure 2.

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

ੳ ੰ ੇ ੈ ਿ ੀ ਾ ੁ ੂ

Auxiliary Signs

ੱ ੰ ਂ

The Half Characters

ੵ ੍ਰ ੍ਵ

Figure 1. Gurmukhi script character set.

Script Character	W1	W2	W3	W4	W5
ੳ					
ਅ					
ੲ					
ਸ					
ਹ					

Figure 2. Samples of handwritten Gurmukhi characters.

III Handwritten Gurmukhi Script Recognition System

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

3.1 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 document, in the form of a bitmap image, is produced. We have used HP-1400 scanner for digitization Digitization produces the digital image, which is fed to the pre-processing phase.

3.2 Preprocessing

Preprocessing is a series of operations performed on the digital image. Preprocessing 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.

Feature extraction and classification phases are discussed in next sections.

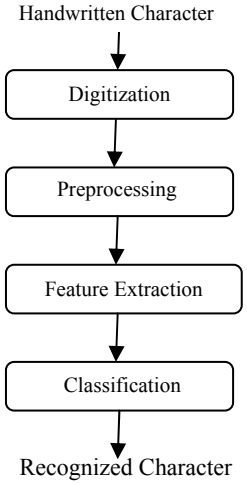

Figure 3. Block diagram of handwritten character recognition system.

IV Feature Extraction

In this phase, the features of input character are extracted. The performance of handwritten character recognition system depends on features, which are being extracted. The extracted features should be able to uniquely classify a character. The performance of recognition system greatly depends on features that are being extracted. We have used two feature extraction techniques, namely; parabola curve fitting and power curve fitting in order to find the feature sets for a given character. The skeletonized image of a handwritten Gurmukhi character is segmented into 100 zones of equal size (10×10). The shape of the curve in each zone is then

estimated by fitting a parabola and also by fitting a power curve using least square estimation. The coefficients of these curves represent the handwritten Gurmukhi character into feature space.

4.1 Parabola Curve Fitting based Feature Extraction

The skeletonized image of a character is divided into *n* (=100) zones as illustrated in Figure 4. A parabola is fitted to the series of *ON* pixels in every zone using least square method. A parabola $y = a + bx + cx^2$ is uniquely defined by three parameters: *a, b* and *c*. As such, this will give 3*n* features for a given character.

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

Step I: Divide the skeletonized image into *n* (=100) number of equal sized zones.

Step II: For each zone, fit a parabola using least square method and calculate the values of *a, b* and *c*.

Step III: Corresponding to the zones that do not have a foreground pixel, take the values of *a, b* and *c* as zero.

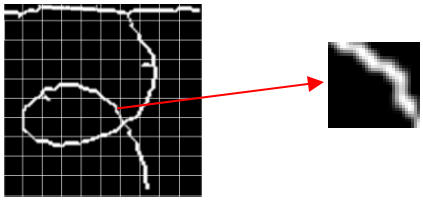

Figure 4. Parabola curve fitting based feature extraction.

4.2 Power Curve Fitting based Feature Extraction

The skeletonized image of a character is again divided into *n* (=100) zones as illustrated in Figure 3. A power curve is fitted to the series of *ON* pixels in every zone using least square method. A power curve of the form $y = ax^b$ is uniquely defined by two parameters: *a* and *b*. This will thus give 2*n* features for a given character.

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

Step I: Divide the skeletonized image into *n* (=100) number of equal sized zones.

Step II: In each zone, fit a power curve using least square method and calculate the values of *a* and *b*.

Step III: Corresponding to the zones that do not have a foreground pixel, take the value of *a* and *b* as zero.

V CLASSIFICATION

In this work, we have used Support Vector Machine (SVM) classifier for recognition. The SVM is a learning machine, which has been widely applied in pattern recognition. SVMs are based on statistical learning theory that uses supervised learning. In supervised learning, a machine is trained instead of programmed to perform a given task on a number of inputs/outputs pairs. SVM classifier has been considered with three different kernels, namely, linear kernel, polynomial kernel and RBF kernel.

VI RESULTS AND DISCUSSION

As stated earlier, we have performed experiments on different training sets sizes while using the SVM as a classifier. The total number of samples in the database is 3500. We have divided the data set using partitioning strategies as depicted in Table 1.

TABLE 1.
PARTITIONING STRATEGIES OF TRAINING AND TESTING DATA

Strategy	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*
Training Data	50 %	55 %	60 %	65 %	70 %	75 %	80 %	85 %	90 %	95 %	99 %
Testing Data	50 %	45 %	40 %	35 %	30 %	25 %	20 %	15 %	10 %	5 %	1 %

6.1 Parabola Curve Fitting based Feature Extraction

In this sub-section, we have presented recognition performance results of different training set strategies (*a, b,…, k*) based on the parabola curve fitting based features (Table 2). Using this approach, we have achieved a maximum recognition accuracy of 97.14% when we use strategy *k* and SVM with linear kernel. These results are depicted in Figure 5 graphically.

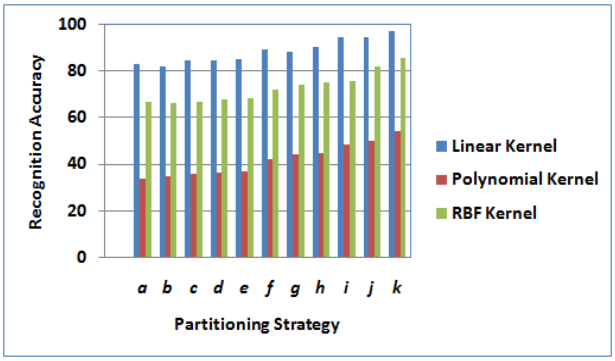

Figure 5. Effect of training set sizes on recognition performance with parabola curve fitting based features.

TABLE 2:
EFFECT OF TRAINING SET SIZES ON RECOGNITION PERFORMANCE WITH PARABOLA CURVE FITTING BASED FEATURES.

Strategy	Linear Kernel	Polynomial Kernel	RBF Kernel
a	82.86%	33.66%	66.97%
b	81.84%	34.79%	66.09%

c	84.43%	36.07%	66.79%
d	84.65%	36.16%	67.59%
e	84.86%	36.95%	68.24%
f	89.02%	42.29%	71.77%
g	88.14%	44.14%	74.20%
h	90.47%	44.57%	74.87%
i	94.57%	48.57%	75.71%
j	94.29%	49.71%	82.14%
k	97.14%	54.29%	85.71%

6.2 Parabola Curve Fitting based Feature Extraction

In this sub-section, recognition results of training set strategies (*a*, *b*,…, *k*) based on power curve fitting features using SVM with three kernels are presented (Table 3). Maximum accuracy achieved here is 89.12% when strategy *k* and SVM with linear kernel is considered. The minimum accuracy achieved is 72% when strategy *a* and SVM with linear kernel again, is considered. These results are also depicted in Figure 6.

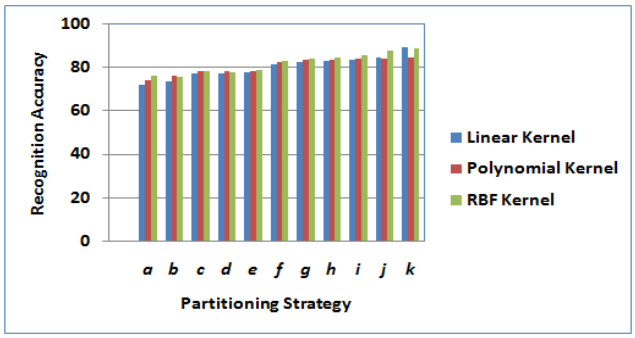

Figure 5. Effect of training set sizes on recognition performance with power curve fitting based features.

TABLE 3:

EFFECT OF TRAINING SET SIZES ON RECOGNITION PERFORMANCE WITH POWER CURVE FITTING BASED FEATURES.

Strategy	Linear Kernel	Polynomial Kernel	RBF Kernel
a	72.00%	74.24%	76.00%
b	73.36%	76.29%	75.49%
c	77.14%	78.07%	78.21%
d	77.39%	78.12%	77.79%
e	77.68%	78.14%	78.67%
f	81.28%	82.29%	83.20%
g	82.47%	83.40%	84.00%
h	83.09%	83.67%	84.38%
i	83.67%	83.92%	85.71%
j	84.28%	84.07%	87.43%
k	89.12%	84.64%	88.57%

CONCLUSION

The work presented in this paper is a study on variation of the recognition performance of the SVM classifier vis-à-vis the variation in the training set size. Eleven training strategies have been explored in this paper in order to recognize the performance of an offline handwritten Gurmukhi script recognition system. This has been noticed that irrespective of the features, the SVM classifier performs increasingly better if we increase the numbers of samples in the training data set. This claim shall be verified in our subsequent work by increasing the size of samples that as of now is 3500.

REFERENCES

[1] Bansal, V., and Sinha, R. M. K. Sinha. 2000. Integrating Knowledge Sources in Devanagari Text Recognition System. *IEEE Transactions on Systems, Man and Cybernetics- Part A: Systems and Humans*, 30, 4 (2000), 500-505.

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

[3] Chaudhary, B. B., and Pal, U. 1997. An OCR System to Read Two Indian Languages Scripts: Bangla and Devanagri (Hindi). In Proceedings of International conference on Document Analysis and Recognition (ICDAR), 2 (1997), 1011-1015.

[4] Hanmandlu, M., Grover, J., Madasu, V. K., and Vasikarla, S. 2007. Input fuzzy for the recognition of handwritten Hindi numeral, In Proceedings of ITNG, 2007, 208-213.

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

[6] Kumar, D. 1991. AI Approach to Hand Written Devnagari Script Recognition, In Proceedings of IEEE Region 10th International conference on EC3-Energy, Computer, Communication and Control Systems, 1991, 229-237.

[7] Kumar, M., Sharma, R. K., and Jindal, M. K. 2010. Lines and words segmentation of offline handwritten Gurmukhi script documents. In Proceedings of International conference on IITM, 2010, 25-28.

[8] Kumar, M., Sharma, R. K., and Jindal, M. K. 2011. SVM based offline handwritten Gurmukhi character recognition, In proceedings of SCAKD, Vol. 758 (2011), 51-62.

[9] Lehal, G. S., and Singh, C. 2000. A Gurmukhi script recognition system. In the Proceedings of 15th ICPR, 2 (2000), 557-560.

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

[11] Parui, S. K., Chaudhuri, B. B., and Majumder, D. D. 1982. A procedure for recognition of connected hand written numerals. *International Journal Systems Sciences,* 13 (1982), 1019-1029.

[12] Roy, K., and Pal, U. 2006. Word-wise Hand-written Script Separation for Indian Postal automation. In the Proceedings of 10th IWFHR, 2006, 521-526.

[13] Sharma, A., Kumar, R., and Sharma, R. K. 2008. Online handwritten Gurmukhi character recognition using elastic matching. *International Journal of Congress on Image and Signal Processing*, 2 (2008), 391-396.

[14] Pal, U., and Chaudhary, B. B. 2000. Automatic recognition of Unconstrained Offline Bangla Handwritten Numerals, In the Proceedings of Advances in Multimodal Interfaces, 2000, 371-378.

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

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

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

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

[19] Swethalakshmi, H., Jayaraman, A., Chakravarthy, V. S., and Sekhar, C. C. 2006. Online handwritten character recognition of Devanagari and Telugu characters using support vector machine. In the proceedings of 10th International workshop on Frontiers in Handwriting Recognition (IWFHR), 2006, 367-372.

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