RESEARCH ARTICLE OPEN ACCESS

Performance Comparison between Different Feature Extraction Techniques with SVM Using Gurumukhi Script

Sandeep Dangi, Ashish Oberoi, Nishi Goel
Department of Computer Science & Engineering Maharishi Markandeshwar University, Mullana India

ABSTRACT
This paper represent the offline handwritten character recognition for Gurumukhi script. It is a major script of india. Many work has been done in many languages such as English , Chinese , Devanagri , Tamil etc. Gurumukhi is a script of Punjabi Language which is widely spoken across the globe. In this paper focus on better character recognition accuracy. The dataset include 7000 samples collected in different writing styles. These dataset divided in two set Training and Test. For Training set collect 5600 samples and 1400 as test set. The evaluated feature extraction include: Distance Profile, Diagonal feature and BDD(Background Direction Distribution). These features were classified by using SVM classifier. The Performance comparison have been made using one classifier with different feature extraction techniques. The experiment show that Diagonal feature extraction method has achieved highest recognition accuracy 95.39% than other features extraction method.

Keywords: OCR, Isolated Handwritten character recognition , Handwritten Gurumukhi script, Background Directional Distribution, Diagonal feature, Distance Profile, SVM classifier.

I. INTRODUCTION

Optical character recognition (OCR) is mechanism which document is scanned or image format convert into editable text format. The input document consist of printed or type written and handwritten text of any language. Handwritten text is difficult than printed text for character recognition Because handwritten text consist of different writing styles of different people at same time with different variations. The input document image can be taken as offline i.e process will work after writing down completely then scanned it. The input for character recognition is also taken as online in which characters are recognized as soon as these are written down or within a fraction of time. For offline character recognition take basic mechanism. This mechanism consists of following phases: Pre processing, Segmentation, Feature extraction, Classification and Post processing.

TRADITIONAL METHODOLGY

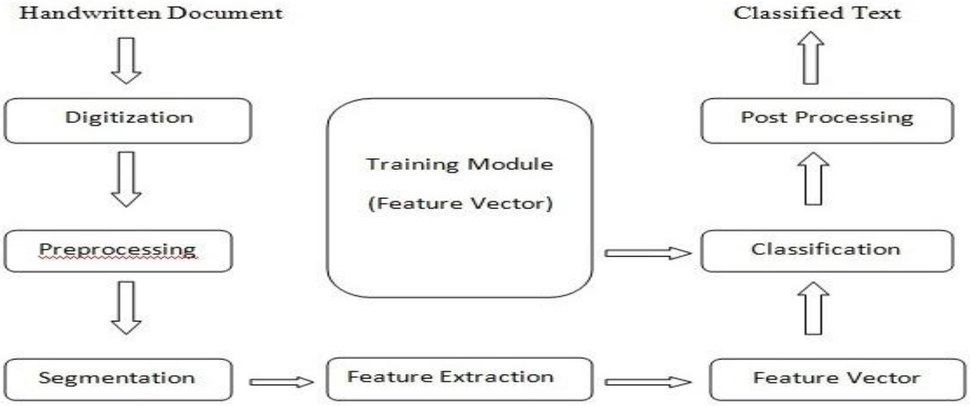
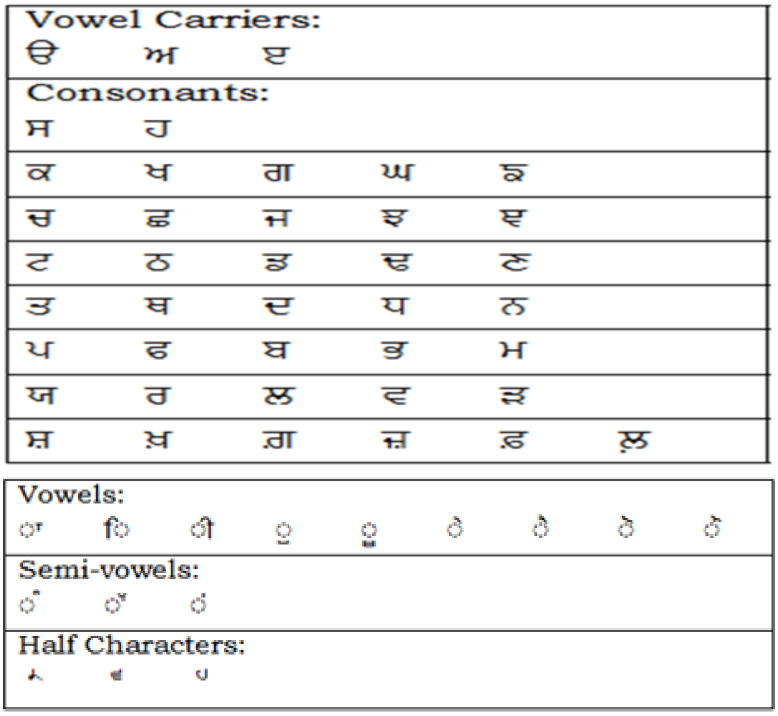


Fig.1. Basic steps of character recognition

The pre processing stage is a collection of operations that apply successive transformations on an image. This stage in handwriting recognition system is applied to reduces noise and distortion present in input stroke. The segmentation stage takes in a page image. This stage will separates the different logical parts such as text from graphics, lines of a paragraph, and characters of a word. The feature extraction stage analyzes a text segment and selects a set of features that can be used to uniquely identify the text segment. Among the different design issues involved in building an OCR system, perhaps the most consequential one is the selection of the type and set of features. The classification stage is the main stage for decision making of an OCR system and uses the features extracted in the previous stage to identify the text segment according to preset rules. The post-processing stage is the final stage. This stage will improves recognition by refining the decisions taken by the previous stage and using context for recognizes words.

I(a) INTRODUCTION TO GURUMUKHI

Gurumukhi script is used primarily for Punjabi language. It is the world's 14th most widely spoken language. The name Gurumukhi derived from the old Punjabi. In Punjabi term "Gurumukhi", mean from mouth of the Guru. We will write Gurumukhi script from left-to-right direction and in top-down approach. Properties of Gurumukhi character. Most of the characters have a horizontal line on the upper part by which characters of the word are connected. This is called as the headline. Gurumukhi has 41 consonants, 9 vowel, and 3 sound modifiers(semi-vowels) and 3 half of the character.



Vowel Carriers:								
ੳ	ਅ	ੲ						
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				
ਸ਼	ਖ਼	ਗ਼	ਜ਼	ਫ਼	ਲ਼			
Vowels:								
◌ਾ	ਿ◌	◌ੀ	◌ੁ	◌ੂ	◌ੇ	◌ੈ	◌ੋ	◌ੌ
Semi-vowels:								
◌ਂ	◌ੰ	◌ੱ						
Half Characters:								
◌੍ਹ	◌੍ਰ	◌੍ਵ						

Fig.2. Character set of Gurumukhi Script

Each word divided in three zones: upper zones , Middle zones and Lower Zones.

Upper Zone: This zone represent above the headline.

Middle Zone: This zone represents the area where the consonants and some other parts of vowels reside, *i.e.*, the area below the headline but above the lower zone.

Lower Zone: This zone represents the area below the middle zone where some vowels, certain half character lie at the foot of the consonants.

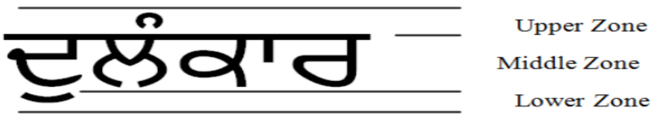


Fig.3. Three Zones and headline in Gurumukhi word

II. LITERATURE SURVEY

This survey show basic idea of doing this survey is to find that how many different types of techniques have been applied for the recognition and to see the recognition accuracy in percentage of their respective methods.

Gita Sinha et al.[1] Zone-Based Feature Extraction Techniques and SVM for Handwritten Gurumukhi Character Recognition obtained recognition accuracy of 95.11 for Gurumukhi character recognition.

Kartar singh siddharth et al.[2] Handwritten Gurumukhi Character Recognition Using Zoning Density and Background Directional Distribution Features obtained recognition accuracy of 94.23% for Gurumukhi character recognition.

Anuj Sharma et al. [3], [4] have presented the implementation of three approaches: elastic matching technique, small line segments and HMM based technique, to recognize online handwritten Gurmukhi characters and reported 90.08%, 94.59% and 91.95% recognition accuracies respectively.

Ubeeka Jain et.al. [5] has recognition system of isolated handwritten characters by using Neocognitron overall recognition accuracy for both learned and unlearned Gurumukhi characters are 92.78 %.

Puneet Jhajj et.al. [6]. used a 48*48 pixels normalized image and created 64 (8*8) zones and used zoning densities of these zones as features. They used SVM and K-NN classifiers and compared the results and observed 72.83% highest accuracy with SVM kernel with RBF kernel.

Munish Kumar [7] presents an efficient offline handwritten Gurumukhi 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 neighbours. It achieves a maximum recognition accuracy of 94.12% using diagonal features and *k*-NN classifier.

Munish Kumar et al. [8] K-Nearest neighbour based offline handwritten Gurumukhi character recognition and obtained 94.12% average recognition accuracy.

III. PROPOSED METHODOLOGY

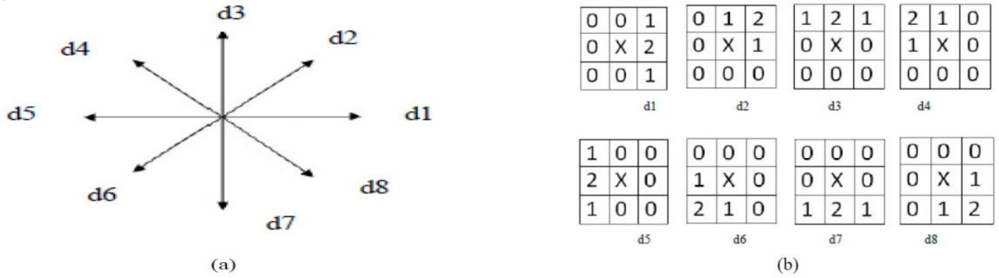
In our methodology we have considered 35 basic character of Gurumukhi alphabet our experiment . 20 writers of different ages, profiles have written this sample on A-4 size paper. Total 7000 samples of our database . 10 samples for each character by each writer. We have SVM classifier for classification. We obtained 95.39% accuracy by 5 fold cross validation.

FEATURE EXTRACTION

We have use three types of features for our experiment-Distance Profile , Diagonal feature and Background Directional Distribution(BDD).

Background Directional Distribution(BDD):

For BDD Feature we have considered the directional distribution background pixels to foreground pixels. We have computed 8 directional distribution features. Directional distribution feature help to calculate directional distribution values of background pixels for each foreground pixel, For each direction used masks shown in fig.4. The pixel at centre 'X' is foreground pixel under consideration to calculate directional distribution values of background.



0	0	1
0	X	2
0	0	1

d1

0	1	2
0	X	1
0	0	0

d2

1	2	1
0	X	0
0	0	0

d3

2	1	0
1	X	0
0	0	0

d4

1	0	0
2	X	0
1	0	0

d5

0	0	0
1	X	0
2	1	0

d6

0	0	0
0	X	0
1	2	1

d7

0	0	0
0	X	1
0	1	2

d8

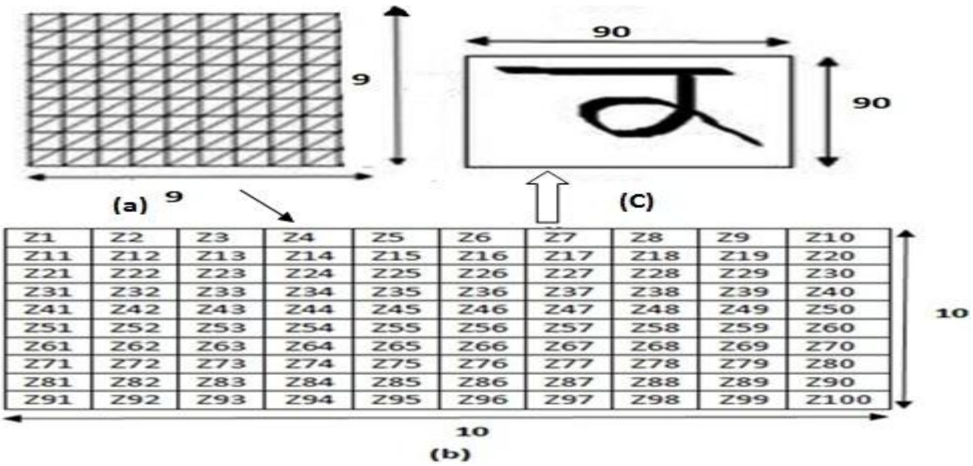
(b)

Fig.4. (a) 8 directions used to compute directional distribution

(b) Masks used to compute directional distribution in different directions

Diagonal Feature:

Diagonal Feature are good feature for achieve highest character recognition accuracy and this feature also reduce the misclassification . Diagonal Feature are extracted form the pixels of each zone by moving along it's diagonal . it is shown in Fig 5. We have used character image size 90*90 pixels having 9*9 zones for describe the diagonal computation and each zone having 9*9 Pixel size. Each zone are present 17 diagonal.



Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
Z11	Z12	Z13	Z14	Z15	Z16	Z17	Z18	Z19	Z20
Z21	Z22	Z23	Z24	Z25	Z26	Z27	Z28	Z29	Z30
Z31	Z32	Z33	Z34	Z35	Z36	Z37	Z38	Z39	Z40
Z41	Z42	Z43	Z44	Z45	Z46	Z47	Z48	Z49	Z50
Z51	Z52	Z53	Z54	Z55	Z56	Z57	Z58	Z59	Z60
Z61	Z62	Z63	Z64	Z65	Z66	Z67	Z68	Z69	Z70
Z71	Z72	Z73	Z74	Z75	Z76	Z77	Z78	Z79	Z80
Z81	Z82	Z83	Z84	Z85	Z86	Z87	Z88	Z89	Z90
Z91	Z92	Z93	Z94	Z95	Z96	Z97	Z98	Z99	Z100

Fig.5. a) Normalized character image b) Image divided in to 100 zones c) Diagonal Feature extraction in a zone of each size 9*9 pixels.

Distance Profile:

In our feature we have used distance profile for simplicity and different variation like feature left , right, top and bottom profile are being used. Profile count number of pixels form bounding box of character image to outer edge of character . Left and right profiles show horizontal transverse. We have used horizontal transverse of distance from left bounding box in forward direction and from right bounding box in backward direction due to outer edge of character. Top and bottom profiles show vertical transverse. We have used vertical transverse of distance from top bounding box in downward direction and from bottom bounding box in upward direction due to outer edge of character. Total 128 feature for all types of profiles. Experiment show four profiles forming 128 pixels(32*4) to define feature vector.

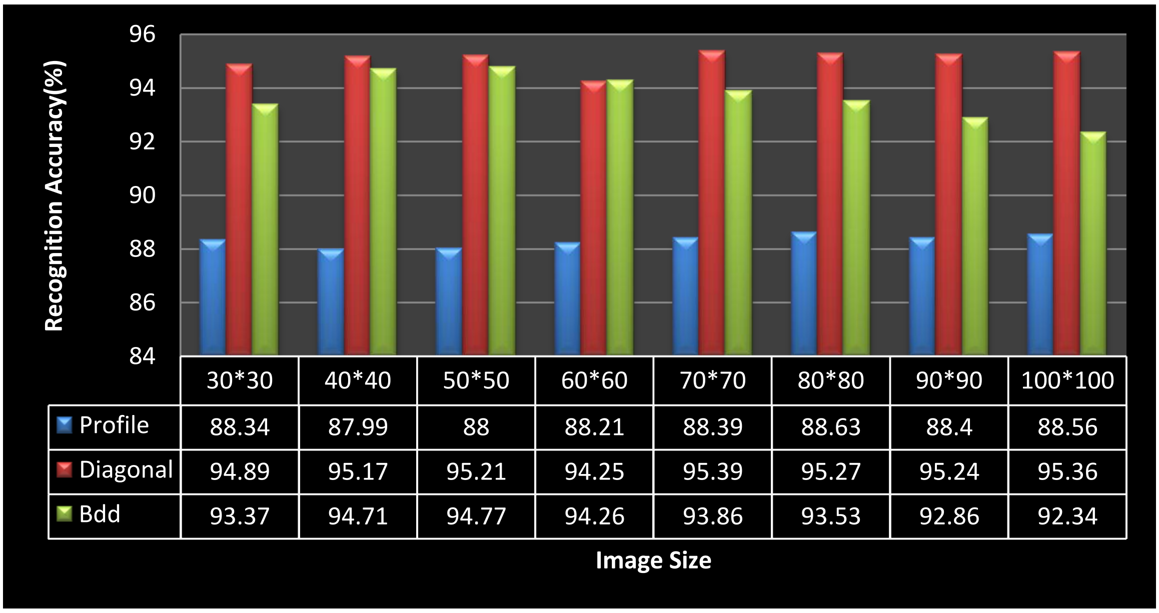
IV. CLASSIFICATION

We have used SVM(Support Vector Machine) classifier for classification and recognition . SVM is very Popular classification tool used for pattern recognition and other classification purpose. SVM is learning machine with good generalization ability. The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes. SVM classifier is trained by a given set of training data and a model is prepared to classify test data based upon this model. we decompose multiclass problem into multiple binary class problems for multiclass classification problem. Different types of kernel used in SVM classifier: Linear kernel, polynomial kernel, Gaussian *Radial Basis Function* (RBF) and *Sigmoid* (*hyperbolic tangent*). Kernel parameters and penalty parameter C. RBF kernel have some choices which single parameter have used gamma (g or γ). Then one subset is used to test by classifier trained by other remaining V-1 subsets. Cross validation through train data is predicated with each sample and it will give the result in percentage of correctly recognized dataset.

V. RESULTS AND CONCLUSION

We have used isolated handwritten Gurumukhi character for experiment . The Dataset consist of 7000 samples in our experiment. This samples written by 20 writers and each writer contribute to write 10 samples of each character out of 35 characters. We have used three different techniques : Distance profiles, Diagonal Feature and Background Direction Distribution with SVM classifier practiced on this samples.

We have used different size of image and feature vector also obtained from the image. Diagonal feature show maximum accuracy 95.39 % from image 70*70 compare to other two techniques The accuracy depend on different size of image.



	30*30	40*40	50*50	60*60	70*70	80*80	90*90	100*100
🟦 Profile	88.34	87.99	88	88.21	88.39	88.63	88.4	88.56
🟥 Diagonal	94.89	95.17	95.21	94.25	95.39	95.27	95.24	95.36
🟩 Bdd	93.37	94.71	94.77	94.26	93.86	93.53	92.86	92.34

Fig.6 : Graph represent recognition accuracy with different size of image

REFERENCES

[1] Gita Sinha, Anita Rani, "*Zone-Based Feature Extraction Techniques and SVM for Handwritten Gurmukhi Character Recognition*" International Journal of Advanced Research in Computer Science and Software Engineering , Vol 2,pp 106-111, June 2012.

[2] Kartar Singh Siddharth, "*Handwritten Gurumukhi Character Recognition Using Zoning Density and Background Directional Distribution Features"*, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (3) pp 1036-1041, 2011.

[3] Anuj Sharma, Rajesh Kumar, R. K. Sharma, "*Online Handwritten Gurmukhi Character Recognition Using Elastic Matching,*" *Image and Signal Processing, 2008. CISP'08. Congress on,* vol.2, pp.391-396, 27-30 May 2008

[4] Anuj Sharma, R.K. Sharma, Rajesh Kumar, "*Online Handwritten Gurmukhi Character Recognition*", Ph.D. Thesis, Thapar University, 2009[Online].

[5] Ubeeka Jain, D. Sharma, "*Recognition of Isolated Handwritten Characters of Gurumukhi Script using Neocognitron*", International Journal of Computer Applications (IJCA),Vol. 4, No. 8,pp 10-16, 2010.

[6] Dharamveer Sharma, Puneet Jhajj "*Recognition of Isolated Handwritten Characters in Gurmukhi Script***",** International Journal of Computer Applications (0975 – 8887) Volume 4– No.8,pp 106-113, August 2010.

[7] Munish Kumar, M. K. Jindal, R. K. Sharma, "*Classification of Characters and Grading Writers in Offline Handwritten Gurmukhi Script* ", International Conference on Image Information Processing (ICIIP 2011) Volume 6-No.10,pp 368-375, 2011.

[8] Munish Kumar, M. K. Jindal, R. K. Sharma, "*K-Nearnest neighbour based offfline handwritten Gurumukhi character reocognition*" International Conference on Image Information Processing (ICIIP 2011) Volume 6-No.10,pp 245-252, 2011.

[9] Rajiv Kumar, Kiran Kumar, "*On the Performance of* Devnagari Handwritten Character Recognition" IDSOI Publication Volume 6, pp 1012-1019 , 2014.

[10] Gurpreet Singh Chandan Jyoti Kumar, " *Feature Extraction of Gurmukhi Script and Numerals: A Review of Offline Techniques*" International Journal of Advanced Research inComputer Science and Software Engineering Volume 3, pp 257-263, June 2013.

[11] Satish Kumar, "*Performance Comparison of Features on Devanagari Hand-printed Dataset*" International Journal of Recent Trends in Engineering, Vol. 1, pp 33-37, may 2009.

[12] Pritpal singh*, sumit budhiraja, "*Feature Extraction and Classification Techniques in O.C.R. Systems for handwritten Gurmukhi Script – A Survey* " , International Journal of Engineering Research and Applications (IJERA) Vol. 1, pp. 1736-1739 , 2006.

[13] Dungre, V. J. " A review of Research on Devnagari Character Recognition", International Journal of Computer Applications, vol. 12, No, 2, pp. 8-15.

[14] Kumar, R., Sharma, R. K. " An Efficient Post-processing Algorithm for Online Handwritten Gurmukhi Character Recognition using Set Theory", International Journal for Pattern Recognition and Artificial Intelligence. Vol. 27, No. 4. Pp. 23-28.

[15] Pal, U., Wakabayashi, Kimura." Comparative Study of Devnagari Handwritten Character Recognition using Different Feature and Classifier", 10th International Conference on Document Analysis and Recognition, pp. 1111-1115, 2009 .

[16] Arora, S., Bhattacharjee, D., Nasipuri, M., Basu, D. K. and Kundu, M., "Combinig Multiple Feature Extraction Techniques for Handwriting Devnagari Character Recognition", Industrial and Information Systems, IEEE Region 10 Colloqium and the Third ICIIS, Vol. 3, pp. 1-6, 2008.

[17] Aparna, K.H., Subramanian, V., Kasirajan, M., Prakash, G. V., Chakravarthy, V.S., and Madhvanath, S., "Online handwriting recognition for Tamil", Ninth International Workshop on Frontiers in Handwriting Recognition, pp. 438-443.

[18] V. K. Govindan and A. P. Shivaprasad, "Character recognition – A survey ", Pattern Recognition, Vol. 23, pp 671-683,1990.