International Journal of Advances in Image Processing Techniques– IJIPT
Volume 1 : Issue 3 *[ISSN 2372 – 3998]*
Publication Date : 30 September. 2014

A New Approach for Handwritten City Name Recognition

Silky Bansal, Munish Kumar, and Mamta Garg

***Abstract*— Handwritten word recognition has been become a area of research in the field of pattern recognition. There is a rich literature available on word recognition in non-Indian scripts but limited work is available on Indian scripts. In this paper, we have presented an approach to recognize handwritten city names written in Gurmukhi script for postal automation. For recognizing words we have used a tree-diagonal feature extraction technique with SVM and *k*-NN classifiers. In this work, we have collected 18,000 samples of handwritten city names in Gurmukhi script. These samples have been collected from 60 different writers and each writer has written 30 city names. Using this approach, we have achieved a recognition accuracy of 90.8%.**

Keywords: **Recognition, Classification, Handwritten recognition system.**

I. Introduction

Word recognition is the capability to make out the words including their shape, form and their meaning. Recognition of handwritten words has been become a research area of pattern recognition from many years because of its various applications in today's world. Some of its applications are:

- Analysis of postal address which includes the recognition of address through state name, city name, zip code *etc*.
- Signature verification verifies the signature of writer.
- Writer recognition depicts the writing and then identifies the writer.
- Bank cheque involves the recognition of amount written on bank cheque.
- Form processing such as tax forms and census forms in which address blocks are recognized.

There are two foremost approaches for recognizing handwritten words. One approach is segmentation approach which means to fragment the word into individual characters and then after identifying each character, concatenate the identified characters to form the word. Another approach is holistic approach which treats the whole word as one and tries to discover that word by considering their different statistical and structural features. Postal automation is the major application for city name recognition system. There are many systems available for postal automation in non-Indian languages but very few works has been done on Indian scripts [1].

II. Review of literature

Gaur and Singh [2] have used a feature named as gradient to recognize the words written in Sanskrit language. They describe the use of sobel operator in gradient feature to detect the edges of documents written in Sanskrit. They also used a classifier named as neural network to recognize the words. The result concludes that if number of hidden nodes increases, then number of epochs also increases. Vaseghi and Hashemi [3] have designed a system for recognizing handwritten words in Farsi/Arabic script by using kohenen self-organizing vector quantization and a classifier named as right-left hidden markov model for reading city names in postal addresses. They used a sliding window to scan the word from left to right for feature extraction. Alkhateeb *et al.* [4] have proposed a word based off-line recognition system by using hidden markov model classifier and re-ranking feature. They used IFN/ENIT database containing 32,492 handwritten Arabic words. They used two features, one is intensity feature for training the HMM and structural feature for re-ranking to improve accuracy. Alkhateeb *et al.* [5] have compared two techniques for recognizing handwritten words in Arabic language and evaluated the efficiency of two classifiers namely as hidden markov models and bayesian networks. They evaluated the performance of both classifiers and found that HMM is more efficient than DBN in classifying various scripts. They also found that HMM is less complex and faster than other classifiers. Vajda *et al.* [1] have presented a system for postal automation by using two parameters. One parameter is pin code and the other is city name. They used a feature known as run length smoothing approach to divide the image into blocks and then detect various postal stamps and postal seals and they also used positional information for finding destination address block. They also used a feature based on water reservoir to find the script of the word and proposed a technique known as NSHP-HMM (Non Symmetric Half Plane Hidden Markov Model) to identify the city name. Koerich *et al.* [6] have presented an approach by combining high and low level features for recognizing the words. They presented this approach by extracting high level features from segmented words and these features are used with a classifier HMM. They also extract low level features from characters produced by the boundaries generated by HMM which are used with segmental neural network classifier. The scores produced by these classifiers are combined by using combination rules. The result produced by this combination reduces the word error rate in almost 71.0%. Ebrahimpur *et al.* [7] have proposed a new method for classifying handwritten words in Farsi language using gradient feature. They extract the features by using multilayer perceptron classifier which is combined with decision templates. This method generates an accuracy of 91.6%. Assabie and Bigun [8] describe two

approaches for Amharic handwritten word recognition using HMMs. One approach is to generate word models from concatenated features of individual characters and second approach is by concatenating HMMs of individual characters to form word model. Prum *et al.* [9] have introduced a bi-character model by joining characters for handwritten word recognition. They used HMM to represent a graph for every word and then decode that graph by using *viterbi* algorithm. This model is used to solve the problem arise in segmentation approaches. Wang *et al.* [10] have presented a framework to combine the results of multiple classifiers and generate an approach known as run time weighted opinion pool for identifying cursive handwritten words. They used the ROVER algorithm to combine different strings of characters generated by each classifier. They demonstrate that this approach gives better results by improving performance and reducing rate of error. Nuzaili *et al.* [11] have investigated two different feature extraction methods for Arabic handwritten word recognition. They presented angular span and distance span method to represent the distribution of pixels in word. Nemmour and chibani [12] have proposed a method to recognize words written in Arabic language by using combination of a technique known as Ridgelet transform and a classifier named as support vector machine. Ridgelet transform is used for finding linear regularity in words. They used the IFN/ENIT database and compared the performance of the proposed technique with radon and zoning features and found that the efficiency of the proposed technique is more as compared to other techniques. Kumar *et al.* [13] have presented a scheme for recognizing offline handwritten Gurmukhi characters. They have used SVM as classifier to recognize the characters and used diagonal, intersection and open end points feature extraction techniques for extracting features for a given character.

III. Gurmukhi script and data collection

Gurmukhi is the script used for Punjabi and Sindhi language. Gurmukhi is the Sikh language of prayer in which the Guru Granth Sahib is written. The word "Gurmukhi" is derived from the word "Guramukhi" which means "Guru's Mouth". Features of Gurmukhi script are: [14]

- Gurmukhi script has forty one alphabets which include thirty eight consonants and three vowel sign bearers.
- Gurmukhi script does not bother about upper and lower case letters.
- It is a script of syllables which includes consonants having an inherent vowel so it is also known as syllabic script.
- The alphabets in Gurmukhi script have a horizontal line over it.
- The alphabets in Gurmukhi script are joined by a line to form a word.
- Gurmukhi word is divided into three zones. Upper zone is the region above head line in which vowels are written, middle zone is the region below head line in which consonants and some part of vowels are written and the lower zone which is the region below middle zone in which some vowels and half characters are written.

In our data set we have considered 30 city names written in Gurmukhi script. This database is collected from 60 writers from different schools and colleges, and each writer has written each city name 10 times forming a total of 18,000 images. A sample of five handwritten Gurmukhi city names by three different writers is given in Figure1.

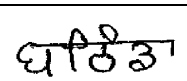
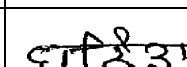
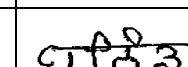
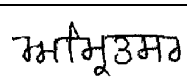
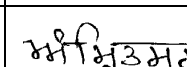
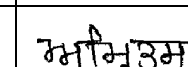
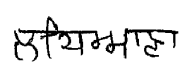
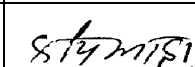
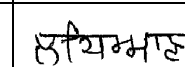
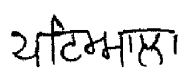
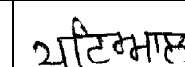
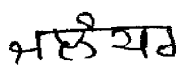
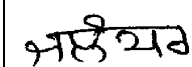
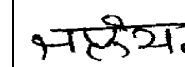
City Name	W_1	W_2	W_3
C_1			
C_2			
C_3			
C_4			
C_5			

Fig. 1. Samples of handwritten Gurmukhi city names

IV. Handwritten word recognition system

This system involves a number of steps which include digitization, preprocessing, feature extraction and classification.

A. *Digitization*

It is a process to convert handwritten city name image into digital format by using a scanner which captures the image and convert it into a digital image file. In digitization phase, the image is converted into binary by using a threshold value of 220.

B. *Preprocessing*

This phase of the system is used to convert the digital image file into bitmap image. To convert the image it follows two steps: one is normalization in which the image is normalized into a fixed size image of 300×100. And second step is converting that normalized image into bitmap image.

C. *Feature Extraction*

In this phase the bitmap word image is used and by the process of locating the pixels in the image the characteristics of the word image are extracted. There are various methods of feature extraction that are useful to extract useful information about the word which helps to recognize the word easily and accurately.

D. *Classification*

This Phase is the decision making phase of handwritten word recognition system. It takes feature vectors extracted from feature extraction phase and uses that feature vectors to predict the unknown class labels. In this paper we have used support vector machine and k-nearest neighbour as classifier. The principle of SVM is to map the inputs to high dimensional feature space which is non-linearly available to input space and to determine a separating hyper plane with maximum margins between two classes. SVM has different kernels which include linear kernel and RBF kernel [15]. Block diagram of the handwritten word recognition system is shown in figure 2.

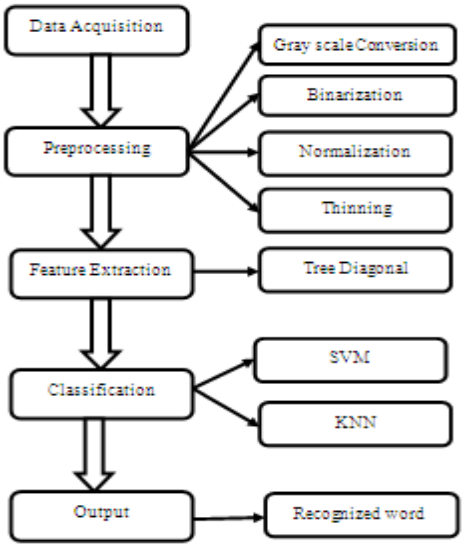


Fig. 2. Block diagram of handwritten word recognition

V. Proposed method

Except from using the segmentation approach for handwritten city name recognition we used holistic approach in which we have considered the whole word. In this we used a technique which comprises of known as tree-diagonal, in which a tree structure comprises of zoning and diagonal feature extraction technique, for handwritten city name recognition system. The steps used for extracting these features are given as below:

Algorithm:

Step I. Input the city name image and resize the image into 88×88 dimensions.

Step II: Divide the city name image into equal sized zones in hierarchal structure.

Step III. Each zone has 2×*n*-1 diagonals; foreground pixels present along each diagonal is summed up in order to get a single sub-feature.

Step IV: Corresponding to the zone whose diagonals do not have a foreground pixel then the feature value is taken as zero.

Step V: First consider the whole image as single word at level *k*=0, apply steps from II to IV.

Step VI: Divide the image into four parts at *k*=1 and then calculate the features in each part by applying the same steps making a total of four features.

Step VII: Then again divide each part into further four parts at *k*=2 making a total of 16 features.

Step VIII: Then again divide these 16 parts into further four parts at *k*=3 each making a total of 64 features.

Step IX: Then all the features extracted are summed up to form a total of 85 features which make a single feature vector.

Step X: Finally, normalize the feature vector in scale from 0 to 1.

VI. Results and Discussions

In this section, experimental results of handwritten city name recognition system are presented. The results are calculated by applying a technique of Tree diagonal. For generating the results, we have taken a sample of 18,000 images, out of which 16200 images are used for training and 1800 images are used for testing. Maximum recognition accuracy of 90.8% has been achieved with SVM classifier.

VII. Conclusion and future scope

This paper has presented a new feature extraction technique for handwritten word recognition system. This feature technique is based on sliding window combined with template matching. This technique generates an accuracy of 90.83% with SVM linear kernel and 72.83% with Polynomial kernel. This accuracy can be increased further by taking a larger sample. This technique can also be applied on other languages such as Hindi, English etc.

REFERENCES

[1] S. Vajda, K. Roy, U. Pal, B. B. Chaudhary and A. Belaid, "Automation of Indian postal documents written in Bangla and English," *International Journal of pattern recognition and artificial intelligence*, Vol. 8(1), pp.1599-1632, 2009.

[2] N. Gaur and D. Singh, "Sanskrit word recognition using gradient feature extraction," *VSRD International journal of computer science and information technology*, Vol. 2 (3), pp.167-174, 2012.

[3] B. Vaseghi and S. Hashemi, "Farsi handwritten word recognition using discrete HMM and self- organizing feature map," *International congress on informatics, environment, energy and applications*, Vol. 38, pp. 2012.

[4] J. H. Alkhateeb, J. Ren, J. Jiang and H. A-Multaseb, "Offline handwritten Arabic cursive text recognition using Hidden Markov Models and re-ranking," *Pattern Recognition Letters*, Vol. 32(8), pp.1081-1088, 2011.

[5] J. H. Alkhateeb, O. Pauplin, J. Ren and J. Jiang, "Performance of hidden Markov model and dynamic bayesian network classifiers on handwritten Arabic word recognition," *Knowledge-Based Systems*, Vol. 24(5), pp. 680-688, 2011.

[6] A. L. Koerich, A. S. B. Jr., L. E. S. D. Oliveira and R. Sabourin, " Fusing High- and Low-Level Features for Handwritten Word Recognition," in Proc.Tenth International Workshop on Frontiers in Handwriting Recognition, pp. 151-156, 2006.

[7] R. Ebrahimpur, R. D. Vahid and B. M. Nezhad, "Decision templates with gradient based features for Farsi handwritten word recognition," *International Journal of Hybrid Information Technology*, Vol. 4(3), pp. 1-12, 2011.

[8] Y. Assabie and J. Bigun, "Offline handwritten Amharic word recognition," *Pattern Recognition Letters*, Vol. 32(8), pp. 1089-1099, 2011.

[9] S. Prum, M. Visani and J. M. Ogier, "On-line handwriting word recognition using a bi-character model," in Proc. Twentieth International Conference on Pattern Recognition, pp. 2700-2703, 2010.

[10] W. Wang, A. Brakensiek and G. Rigoll, "Combination of multiple classifiers for handwritten word recognition," in Proc. eighth International Workshop on Frontiers in Handwriting Recognition, pp. 117-122, 2002.

[11] Q. A. Nuzaili, D. Mohamad, N. A. Ismail and M. S. Khalil, "Feature extraction in holistic approach for Arabic handwriting recognition system: a preliminary study," in Proc. IEEE 8th International Colloquium on Signal Processing and its Applications, pp. 335-340, 2012.

[12] H. Nemmour and Y. Chibani, "Handwritten Arabic word recognition based on ridgelet transform and support vector machines," in Proc. International conference on High Performance Computing and Simulation (HPCS), pp. 357-361, 2011.

[13] M. Kumar, M. K. Jindal and R.K. Sharma, "SVM Based Offline Handwritten Gurmukhi Character Recognition," in Proc. SCAKD, pp. 51-62, 2011.

[14] G. S. Lehal and C. Singh, "Feature Extraction and Classification for OCR of Gurmukhi Script," *VIVEK*, Vol.12 (2), pp. 2-12, 1999.

[15] Arora, D. Bhattacharjee, M. Nasipuri, L. Malik, M. Kundu and D.K. Basu, "Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition," *International Journal of Computer Science Issues*, Vol. 7(3), pp.18-26, 2010.