Recent Trends and Tools for Feature Extraction in OCR Technology

Om Prakash Sharma, M. K. Ghose, Krishna Bikram Shah, Benoy Kumar Thakur

***Abstract*—This paper presents a recent trends and tools used for feature extraction that helps in efficient classification of the handwritten alphabets. Numerous models of feature extraction have been defined by different researchers in their respective dissertation. It is found that the use of Euler Number in addition to zoning increases the speed and the accuracy of the classifier as it reduces the search space by dividing the character set into three groups.**

***Index Terms*— Handwritten Character Recognition, Feature Extraction, Zoning, Euler Number, Classification.**

I. INTRODUCTION

Character recognition plays an important role in this modern world where there are heterogeneous representation of text based information. It is the mechanical or electronic translation of handwritten, typewritten or printed text into machine editable formats. Character recognition also popularly referred as optical character recognition (OCR) is a field of research that has immense potential in future where we want to track and locate every piece of information being exchanged. The problem with the hand written text is due to uncertainties such as variation in calligraphy over period of time, similarity in text, variation in styles of writing. These have made hand written text continues to be challenging area of research work. Given a grey scale image of characters as input the task is to recognise the character and assign the corresponding ASCII values to the recognized characters. In general the character recognition is basically classified into two types: offline handwritten text recognition, online handwritten text recognition [1].

II. MAJOR PHASES OF OCR SYSTEM

A. Preprocessing

The goal of pre-processing is to simplify the pattern recognition problem without missing any vital information. It reduces the noises and inconsistent data [7]. It enhances the image and prepares it for the next steps.

B. Segmentation

Segmentation is an integral part of any text based recognition system. It assures efficiency of classification and recognition [7]. Accuracy of character recognition heavily depends upon segmentation phase. Incorrect segmentation leads to incorrect recognition. Segmentation phase include line segmentation, character and word segmentation [9]. It is important to obtain complete segmented character without any noise to ensure quality feature extraction.

C. Normalization

The results of segmentation process provides isolated characters which are ready to pass through feature extraction stage, thus the isolated characters are reduced to a specific size depending on the methods used. The segmentation process essentially renders the image in the form of m*n matrix. These matrices are then generally normalized by reducing the size and removing the redundant information from the image without loosing any important information [10].

D. Feature Extraction

Feature extraction is the process of extracting the relevant features from objects/alphabets to form a feature vectors. These feature vectors is then used by classifiers to recognize the input unit with target output unit. It becomes easier for the classifier to classify between different classes by looking at these features as it allows fairly easy to distinguish [1].

E. Classification

The results Classification is the last stage where we train the neural net using the feature vectors obtained during feature extraction method against the required targets. To optimize the whole recognition process, several combination methods of multilayer perceptron have been devised.

III. TOOLS

There are several tools available in the market for implementation and testing of character recognition. Among them the most popular tools are tesseract from google as an open source and the other one is matlab, a premier product of Mathworks Inc.

A. Tesseract

Tesseract is a free optical character recognition engine. It was originally developed as proprietary software at Hewlett-Packard between 1985 until 1995. After ten years without any development taking place, Hewlett Packard and UNLV released it as open source in 2005. Tesseract is currently developed by Google and released under the Apache License.

Tesseract is a raw OCR engine. It has no document layout analysis, no output formatting, and no graphical user interface. It only processes a TIFF image of a single column and creates text from it. Tesseract can process English, French, Italian, German, Spanish, Brazilian Portuguese and Dutch [18].

B. Matlab

On the other hand, matlab is a premier commercial product of Mathworks Inc, originally created as a front end tool for

Manuscript received on January, 2013.

Om Prakash Sharma, Department of Computer Science and Engineering, Sikkim Manipal Institute of Technology, Majitar, Sikkim, India.

Dr. M. K. Ghose, Department of Computer Science and Engineering, Sikkim Manipal Institute of Technology, Majitar, Sikkim, India.

Krishna Bikram Shah, Department of Computer Science and Engineering, Sikkim Manipal Institute of Technology, Majitar, Sikkim, India.

Benoy Kumar Thakur, Department of Computer Science and Engineering, Rockvale Management College, Kalimpong, West Bengal, India.

one of these, the LINPACK package-a group of routines for working with matrices and linear alzebra. Over a period of year now matlab has evolve as the tool of choice for high productivity research, development and analysis.
Matlab is defined as high performance language for technical computing. It integrates computation, visualization, and programming in an easy-to use environment where problems and solutuions are expressed in familiar mathematical notation. Matlab comes with various toolboxes, out of which the image processing toolbox and neural network toolbox plays a vital role in pattern recognition.

IV. CURRENT TRENDS IN OCR TECHNOLOGY

The accurate recognition of Latin-script, typewritten text is now considered largely a solved problem. Typical accuracy rates exceed 99%, although certain applications demanding even higher accuracy require human review for errors. Other areas—including recognition of hand printing, cursive handwriting, and printed text in other scripts (especially those with a very large number of characters)--are still the subject of active research [18].

Optical Character Recognition (OCR) is sometimes confused with on-line character recognition. OCR is an instance of off-line character recognition, where the system recognizes the fixed static shape of the character, while on-line character recognition instead recognizes the dynamic motion during handwriting.

Recognition of cursive text is an active area of research, with recognition rates even lower than that of hand-printed text. Higher rates of recognition of general cursive script will likely not be possible without the use of contextual or grammatical information. For example, recognizing entire words from a dictionary is easier than trying to parse individual characters from script. Reading the Amount line of a cheque (which is always a written-out number) is an example where using a smaller dictionary can increase recognition rates greatly. Knowledge of the grammar of the language being scanned can also help determine if a word is likely to be a verb or a noun, for example, allowing greater accuracy. The shapes of individual cursive characters themselves simply do not contain enough information to recognize all handwritten cursive script accurately (greater than 98%) [18].

It is necessary to understand that OCR technology is a basic technology also used in advanced scanning applications. For more complex recognition problems, intelligent character recognition systems are generally used such as HMM, SVM and ANN. The artificial neural networks can be more advantageous and can be made indifferent to both affine and non-linear transformations.

V. ARTIFICIAL NEURAL NETWORK

An Artificial Neural Network (ANN) is an information processing paradigm. It is inspired by the way the biological nervous system function, such as brain processing the function of interest. The key element is the novel structure of information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working parallel to solve specific problems as shown in fig 1. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true in case of ANNs as well.

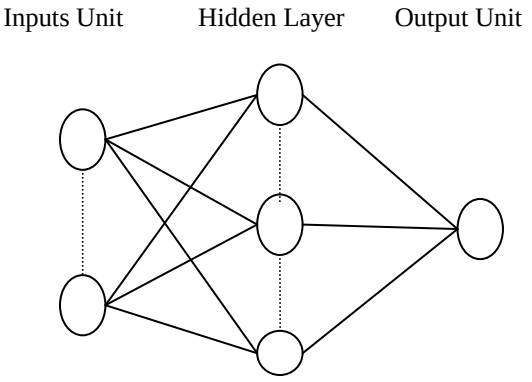

Fig 1: Neural Network

VI. COMPARATIVE ANALYSIS

After going through several dissertation and Internationals Journal on Hand Written Character Recognition based on Artificial Neural Network, it is found that, several methods has been devised at each steps of character recognition, of which the most of them has concentrated on the feature extraction process. Selection of feature is the single most important factor in achieving high recognition in performance in character recognition system [10]. The widely used feature extraction methods are: Template matching, Unitary Transforms, Projection Histograms, Zoning, Geometric Moments and Zernike Moments.

Dinesh et al [15] have used horizontal/vertical strokes, and end points as the potential features for recognition and reported a recognition accuracy of 90.50% for handwritten Kannada numerals. However, this method uses the thinning process which results in the loss of features [1].

U. Pal et al [16] have proposed zoning and directional chain code features and considered a feature vector of length 100 for handwritten numeral recognition and have reported a high level of recognition accuracy. However, the feature extraction process is complex and time consuming [1].

Anita Pal et al [12] have proposed the features extraction from Boundary tracing and their Fourier Descriptors. Also an analysis has been carried out to determine the number of hidden layer nodes to achieve high performance of back propagation network. A recognition accuracy of 94% has been reported for handwritten English characters with less training time.

E. Srinivasan et al [1] have proposed diagonal based feature extraction for handwritten alphabets recognition system using neural network and from the test results it has been identified that the diagonal method of feature extraction yields the highest recognition accuracy of 97.8 % for 54 features and 98.5% for 69 features.

Recently, an improved approach to feature extraction is cited in [17] which may be efficient as compared to former methods. In this approach first the Euler Number is computed. Based on the result of computation whether the result is positive or negative or zero, the character set is divided into three groups.

A. *Euler Number*

Euler Number is defined as number of connected components in the image minus the number of holes.
This will divide the characters into 3 groups [2]:
Euler number equal one and this contains: s, S, f, F, G, h, H, j, J, k, K, l, L, z, Z, x, X, c, C, v, V, n, N, m, M, w, W, E, r, t, T, y, Y, u, U, i, I.
Euler number equal zero and this contains: q, Q, R, o, O, p, P, a, A, d, D, g, b.
Euler number equal minus one and this contains: B.

B. *Zoning*

The Zoning is one of the most popular and simple to implement feature extraction method. The commercial OCR system developed by CALERA used Zoning mechanism on binary characters [10]. An (n*m) grid is superimposed on the character image and then for each zone average value is computed giving a feature vector of length (n*m), if required further we can compute the average on this zone again in row wise and column wise respectively.

After the categorization, Diagonal Based Feature Extraction method as implemented by J. Pradeep, E. Shrinivasan and H. Himavathi [1]. In this feature extraction process, the individual segmented characters are first resized into a size 90x60 pixel and then the Euler Number is computed for the character to identify in which class it belongs [17]. The result obtained is then followed by the diagonal based zoning operation is carried where an image is further divided into 54 equal zones, each of size 10x10 pixels.

Each zone has 10 horizontal lines and the foreground pixels present long each horizontal line is summed to get a single sub-feature, thus 10 sub-features are obtained from the each zone. These 10 sub-features values are averaged to form a single feature value and placed in the corresponding zone. This procedure is sequentially repeated for all the zones. There could be some zones which are empty of foreground pixels. The feature values corresponding to these zones are zero. Finally, 54 features are extracted for each character. In addition, 9 and 6 features are obtained by averaging the values placed in zones row-wise and column-wise, respectively. As result, every character is represented by 69, that is, 54 +15 features [1][11]. These extracted features are used to train a feed forward back propagation neural network for performing classification and recognition tasks.

VII. RESULT AND DISCUSSION

Analysis is been carried out by using Matlab 7.5. The scanned image and the image drawn using paint application is given as an input to the feed forward neural net architecture where it is first converted from .png to .tiff file and then it is resized to a standard format of 60*90 pixels image followed by the thresholding/binarization operation.

After the binarization the gradient descent back propagation method with momentum and adaptive learning rate and log-sigmoid transfer functions is used for neural network training [1] on the obtained feature vectors. A recognition system using two different features, one simple diagonal based feature and the other includes the Euler number. In both the cases size of feature vectors is same; the results obtained using these two different types of feature extraction are summarized in Table I [17].

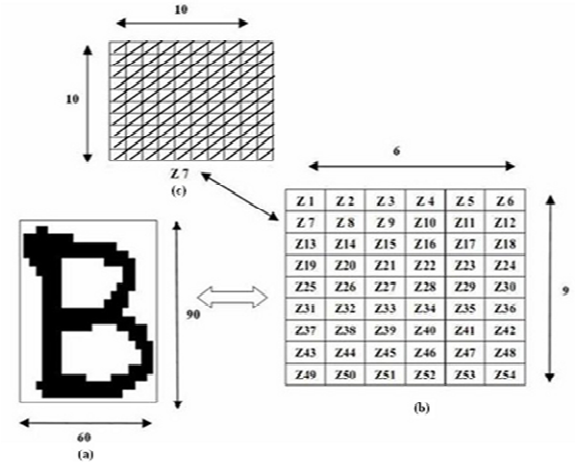

Fig 2: Diagonal Based Zoning Operation [1].

From the columns in Table I we can see that if some how the system misses the alphabets to recognize due to similarity in appearance as it has happened in column of diagonal based result where it has missed to recognize the alphabet E due to its similarity with alphabet B. But if we consider the column of Euler Based hybrid result, there we can see that use of Euler has helped to recognize the alphabet E because the result of Euler Computation categorizes E and B into two different distinct sets. For B the Euler is positive 1 and E is negative -1.

Table 1: Recognition result & time elapsed obtained based on diagonal based Hybrid Based feature extraction. [17]

Alpha bets	Diagonal Based Result	Elapsed Time (in sec.)	Hybrid Based Result	Elapsed Time (in sec.)
A	Recognise	0.4935	Recognise	0.4207
B	Recognise	0.7075	Recognise	0.1953
D	Recognise	0.6037	Recognise	0.5103
E	Missed	0.6055	Recognise	0.1941
P	Recognise	1.4506	Recognise	0.2986
Q	Recognise	0.1766	Recognise	0.1941
M	Recognise	0.6070	Recognise	1.0355
S	Recognise	1.9696	Recognise	0.4055
T	Recognise	0.1781	Recognise	0.5103
Y	Recognise	1.9688	Recognise	0.1941
Z	Recognise	0.5088	Recognise	0.1721

VIII. CONCLUSION

An improved feature extraction, namely, an efficient hybrid feature extraction model recently proposed can give high recognition accuracy while requiring less time for training and classification both. It not only yields higher levels of recognition accuracy but the overall time efficiency has increased as compared to the systems employing the diagonal based feature extraction [1] or any other conventional methods of feature extraction. From the result of Table I it can be concluded that the use of Euler Number computation

which is good in categorizing the alphabets into different groups, with zone based feature extraction method will be effective process for increasing the speed and accuracy.

REFERENCES

[1] J Pradeep, E Shrinivasan and S.Himavathi, "Diagonal Based Feature Extraction for Handwritten Alphabets Recognition System Using Neural Network", International Journal of Computer Science & Information Technology (IJCSIT), vol . 3, No 1, Feb 2011.

[2] M. Alata — M. Al-Shabi, " Text Detection And Character Recognition Using Fuzzy Image Processing", Journal of Electrical Engineering, vol. 57, no. 5, 2006, 258–267

[3] R. Plamondon and S. N. Srihari, "On-line and off- line handwritten character recognition: A comprehensive survey,"IEEE. Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 63-84, 2000.

[4] N. Arica and F. Yarman-Vural, "An Overview of Character Recognition Focused on Off-line Handwriting", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2001, 31(2), pp. 216 - 233.

[5] U. Bhattacharya, and B. B. Chaudhuri, "Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals," IEEE Transaction on Pattern analysis and machine intelligence, vol.31, No.3, pp.444-457, 2009.

[6] U. Pal, T. Wakabayashi and F. Kimura, "Handwritten numeral recognition of six popular scripts," Ninth International conference on Document Analysis and Recognition ICDAR 07, Vol.2, pp.749-753, 2007.

[7] Devinder Singh and Baljit Singh Khehra, "Digit Recognition System Using Back Propagation Neural Network", International Journal of Computer Science and Communication Vol. 2, No. 1, January-June 2011, pp. 197-205

[8] VENTZAS, DIMITRIOS 1, NTOGAS, NIKOLAOS , "A BINARIZATION ALGORITHM FOR HISTORICAL MANUSCRIPTS", 12th WSEAS International conference on Communications, Heraklion, Greece, July 23-25, 2008.

[9] Bindu Philip, R. D. Sudhaker Samuel and C. R. Venugopal, Member, IACSIT, "A Novel Segmentation Technique for Printed Malayalam Characters", International Journal of Computer and Electrical Engineering, Vol. 2, No. 4, August, 2010 1793-8163 Printed Malayalam Characters.

[10] Anil Kumar Jain and Torfinn Taxt, "Feature extraction Methods for Character Recognition- A Survey", Pattern Recognition, Vol.29, No.4, pp. 641-662, 1996.

[11] S.V. Rajashekararadhya, and P.VanajaRanjan, "Efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south-Indian scripts," Journal of Theoretical and Applied Information Technology, JATIT vol.4, no.12, pp.1171-1181, 2008.

[12] Anita Pal & Dayashankar Singh, " Handwritten English Character Recognition Using Neural Network", International Journal of Computer Science & Communication", Vol. 1, No.2, July-December 2010, pp. 141-144

[13] G. Vamvakas, B. Gatos, I. Pratikakis, N. Stamatopoulos, A. Roniotis, S.J. Perantonis, "Hybrid Off-Line OCR for Isolated Handwritten Greek Characters", The Fourth IASTED International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA'07), pp. 197-202, Innsbruck, Austria, February 2007.

[14] G. Vamvakas, B. Gatos, S. Petridis and N. Stamatopoulos, "An Efficient Feature Extraction and Dimensionality Reduction Scheme for Isolated Greek Handwritten Character Recognition", Proceedings of the 9th International Conference on Document Analysis and Recognition, Curitiba, Brazil, 2007, pp. 1073-1077.

[15] Dinesh Acharya U, N V Subba Reddy and Krishnamurthy, "Isolated handwritten Kannada numeral recognition using structural feature and K-means cluster," IISN-2007, pp-125 -129.

[16] M Arijit Bishnu, Bhargab B. Bhattacharya, Malay K. Kundu b C.A. Murthy, Tinku Acharya, "A pipeline architecture for computing the Euler number of a binary image", Journal of Systems Architecture 51 (2005) 470–487.

[17] Om Prakash Sharma, M. K. Ghose, Krishna Bikram Shah, "An Improved Zone Based Hybrid Feature Extraction Model using Euler Number", Innternationa Journal on Soft Computing and Engineering (IJSCE'12), ISSN 2231-2307, Volume -II, Issue- II, Article no-96, pp. 154-158.

[18] Bishnu Chaulagain, Brizika Bantawa Rai, Sharad Kumar Raya, "Final Report on Nepali Optical Character Recognition NepaliOCR", Submitted On July 29, 2009.

[19] R. J. Vidmar. (1992, August). On the use of atmospheric plasmas as electromagnetic reflectors. *IEEE Trans. Plasma Sci.* [Online]. *21(3).* pp. 876—880.
Available: http://www.halcyon.com/pub/journals/21ps03-vidmar