Efficient Character Recognition Approach for Handwritten Documents

M.Arun[*1] , G. Anitha[*2] , S. Mathivani[*3]

[*1, *2, *3]*Dept of ECE, Mepco Schlenk Engineering College, Sivakasi, India*

Abstract : **An off-line handwritten character recognition system for segmented English characters using Support Vector Machine as classifier is described in this paper. The performance of an optical character recognition system mainly depends on the extracted features. Feature extraction plays a major role in achieving high recognition accuracy.Feature extraction helps to obtain the significant features that can be fed into the classifier since not all the features of an image are required for classification. Features that are included for the character recognition are Histogram of Oriented Gradient descriptor (HOG), Gabor features, Discrete Cosine transform features.This proposed system will be suitable for converting handwritten documents into structural text form and recognizing them.**

Keywords : **English Character, Feature Extraction, HOG. Offline Character Recognition, SVM Classifier.**

I. INTRODUCTION

Handwritten Character recognition is the ability of the computer to interpret intelligible handwritten input from several sources.In recent years,with increasing amounts of data being generated by businesses and researchers, there is a need for fast, reliable and accurate algorithms for data analysis. The intelligent data analysis has been developed by the contributions of artificial intelligence, improvements in databases technology and computing performances. Character recognition has now become a challenging task.It contributes a lot to the advancement of automation process and it can improve the interface between man and machine in different applications. Several works for research have been focusing on varying techniques and methods that would reduce the processing time while providing higher recognition accuracy. The input given to the character recognition system is divided into offline and online.

In offline character recognition system, scanned images of handwritten characters are processed. However, in online character recognition system, characters are processed in real time while writing takes place.The text contained in the images will be of different fonts, different formats , different languages, quality may be poor, or even images may be blurred and hence the proper text extraction becomes very difficult. If all these difficulties are overcome, the text extraction can be beneficial for numerous applications. Therefore to design a system which is so versatile as possible is a bit difficult. Handling the unknown text layout in the image, character fonts and sizes and variability in imaging conditions with uneven lighting, reflection, shadowing and aliasing makes character recognition a challenging task. All these challenges have been considered before developing a good character recognition system.

The initial step of a general recognition system is preprocessing. Preprocessing is undergone in order to obtain better feature sets. It includes noise removal, resizing, binarizing, autocropping etc of the image before features have been extracted. Feature extraction follows preprocessing. Several methods contribute to feature extraction such as projection histograms,fourier descriptors,spline curve approximation,contour profiles, zoning,histogram of oriented gradients,template matching,deformable templates,structural an statistical features etc.

II. ENGLISH CHARACTER SET

India is a linguistically rich country with different popular languages. It includes English as one of the languages. English language has 26 letters which makes easy to process and recognize. The data set for digits is collected from the MNIST database in which the digits counts to 5500.For English characters, (both capital and small letters) the samples are collected from 20 different writers and in total there are 1040 characters.

Character dataset is collected from different places under different age groups. Samples used for character recognition is written by the people in an unconstrained manner, ie, without forcing them to use specific pen, ink color, line thickness, writing style etc.

Some dataset images are shown in the figure below,

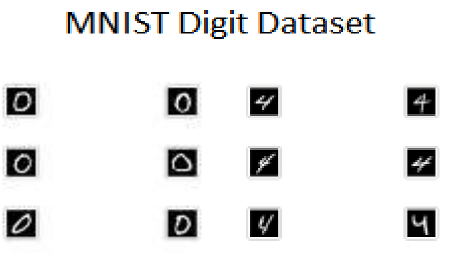


Figure 1. MNIST Digit Dataset

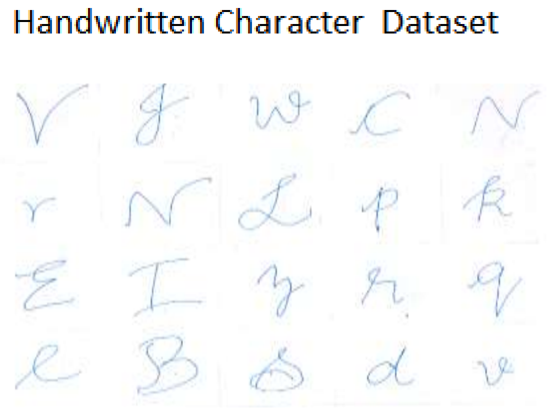


Figure 2.Handwritten Character Dataset

III. RELATED WORK

There are several researches have been done regarding the handwriting recognition in various fields.

Gururaj Mukarambiand et al proposed”Zone based character recognition engine for Kannada and English scripts “. In this paper, Optical character recognition engine is proposed based on zone features.The zone is one of the old concepts in case of docment image analysis research. Here the kannada consonants and English alphabets sample images are classified based on the SVM classifier.The average recognition accuracy for kannada consonants is 73.33% . The recognition accurcy is low for Kannada consonants because most of the characters are similar in shape.

J.Pradeep,E.Srinivasan and S.Himavathi scheduled “Diagonal based feature extraction for Handwritten alphabets recognition system using Neural network“[7].In thier a new method called diagonal base feature extractionis introduced for extracing the features of the handwritten alphabets.The features are extracted from the pixels of each zone by moving along their diagonals.This procedre is repeated for all the zones leading to the extraction of features for every character. The accuracy for this method totally depends on the number of features extracted from the character.However the feature extraction process is complex an time consuming.

Ranjan Jana and et al suggested “Optical Character Recognition from Text Image “[8] where the recognition approach is based on template matching. The text image is divided into several regions by isolating each line and then individual characters with spaces. After character extraction, the texture and topological features like corner points, features of different regions, ratio of character area and convex area of all characters of text image are calculated. Previously , features of each uppercase and lowercase letter, digit, and symbols are stored as a template.

Jino P Jand Kannan Balakrishnan designed“Offline Handwritten Recognition of Malayalam District Name“[9] - A Holistic Approach proposed by Features consider for the recognition are Histogram of Oriented Gradient descriptor, Number of Black Pixels in the upper half and lower half, length of image. The Holistic recognition provides good results to overcome smaller class problems. But in the larger class problems more number of samples are required with carefully designed features.

Sandhya Arora and et al recommended“Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition“[10].The paper takes Devanagiri script as input. Statistical classifiers are rooted in the Bayes decisionrule and can be divided into parametric ones and nonparametric ones. Three major features i.enumber of paths, direction of paths and region of the node were extracted from the middle zone.The concept is Shadow feature of character,chain code, histogram of character contour obtained accuracy is 92.38%.

Nibaran Das and et al considered “Handwritten *Bangla* Basic and Compound character recognition using MLP and SVM classifier“[11]. In this paper, 50 symbols of basic *Bangla* alphabets are considered as samples.The paper includes shadow features,long run features and Quad –tree based fature. On experimentation, the technique produces an average recognition rate of 80.5% using SVM after three fold cross validation of data for Bangla character.

IV. FRAMEWORK OF THE PROPOSED METHOD

The proposed method includes two phases. One is training phase and another one is testing phase. The OCR system primarily involves four steps: Pre-processing, Features extraction, Features training and classification.

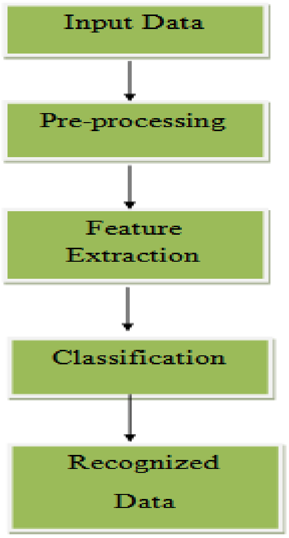


Figure 3. Describes the block diagram of the proposed method.

Initially in the training phase , the scanned input image is subjected to preprocessing. The preprocessed image has undergone to several feature extraction techniques.During training phase,classifier learns about the characteristics of classes based on the extracted features .During testing phase, classifier classifies the characters based on the training data.Finally the testing data is recognised appropriately.

A. IMAGE ACQUISTION :

The scanned image forms the input for the recognition system.The output from scanner,digital camera or from any other device in the form of image is given as input to the recognition system.

B. PREPROCESSING :

The pre-processing consists of series of operations performed on the scanned input image..In character recognition systems most of the applications use gray or binary images since processing colour images increases the complexity.Some of the difficulties in colour images is that it may also contain watermarks or non-uniform background making it difficult toidentify.Preprocessing Techniques in Character Recognition extract the document text from the image.

Scanned images may or may not contain noise. In case of noise it can be removed by preprocessing step using median filter.

B.1.MEDIAN FILTER

The main concept of median filter is, it scans each and every pixel of input image. It replaces pixel value with average of its neighbouring pixel. Median is easy to define if entries of window has an odd number. In this case just sort all values, middle number is a median value. But in even number case there is more than one possible median. The general equation for the median filter is given by equation (1).

F(x,y)= $Median_{(s,t)} \in s_{xy}\{g(s,t)\}(1)$

The maximum efforts are done by calculating the median of each window because filter must process every entry in signal. The first step is that all entries must be sorted, then select middle entry so selection sort is efficient for this. If signals use whole number representation histogram median can be efficient. It is simple to update the histogram by traversing from window to window, and finding the median of a histogram. Median filter is nonlinear filter which is one kind of smoothing technique, such as linear Gaussian filtering. All the filtering techniques are effective in noise removal of smooth patches or smooth areas of a signal, but it affect its edges a lot. In text extraction and recognition, it should be noted that the noise must be removed from image and it is important to preserve the edges. Edges play a vital role in the visual appearance of images. Because of this, median filtering is used for preprocessing.

C. FEATURE EXTRACTION :

Feature extraction technique is applied for all individual extracted characters.

C.1.Histogram of Oriented Gradients :

(i) Histogram of Oriented Gradients :

Histogram of oriented gradients (HOG) is a global feature descriptor used to detect objects in image processing. The HOG descriptor technique counts occurrences of gradient orientation in localized portions of an image. The HOG feature is used to extract the shape and appearance of the object by mapping the magnitude and direction of the image.

To calculate a HOG descriptor, we need to first calculate the horizontal and vertical gradient .After all, we want to calculate the histogram of gradients. This is easily achieved by using **Sobel** operator in Open CV with kernel size 1.The magnitude and direction of gradient is given by equation (2) and (3).

$$g = \sqrt{{g_x}^2 + {g_y}^2}(2)$$

$$\emptyset = \arctan\left(\frac{g_x}{g_y}\right)(3)$$

Implementation of the HOG descriptor algorithm:

- Cells are formed by dividing the image into small connected regions, and for each cell compute a histogram of gradient directions or edge orientations for the pixels within the cell.
- Discretize each cell into angular bins according to the gradient orientation.
- Each cell's pixel contributes weighted gradient to its corresponding angular bin.
- Adjacent cell groups are considered as spatial regions called blocks. The basis for grouping and normalization of histograms is the grouping of cells into blocks.
- Normalized group of histograms represents the block histogram. The set of these block histograms represents the descriptor.

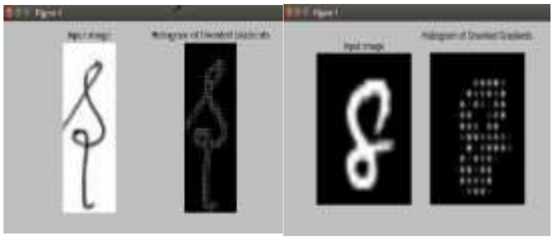


Figure 1
Input image
Histogram of Oriented Gradients
Figure 1
Input image
Histogram of Oriented Gradients

Figure 4. Input Image and its HOG Features visualizations

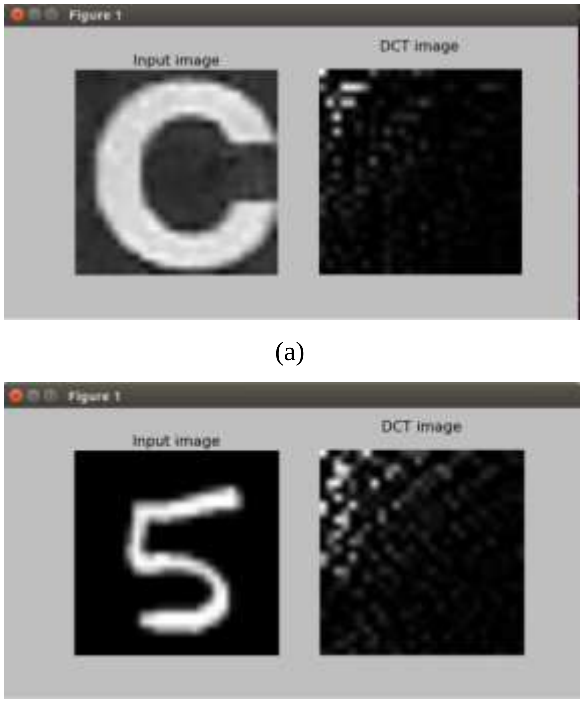
C.2.Discrete Cosine Transform:

A **discrete cosine transform** (**DCT**) expresses a finite sequence of data points in terms of a sum of **cosine** functions oscillating at different frequencies.DCT is similar to the discrete Fourier transform. It transforms a signal from spatial domain to the frequency domain. It is computationally easier to implement and more efficient to regard the DCT as a set of basis functions for which a known input array size (8x8 window) is given that get applied to the entire image. The most common variant of discrete cosine transform is type- II DCT, which is often called simply “the DCT ”.The 2d DCT is given by the equation (4).

F(u,v)=

$$1/N^2 \sum_{m=0}^{N-1} \cdot \sum_{n=0}^{N-1} f[m,n] \cos\left[\frac{(2m+1)u\pi}{2M}\right] \cos\left[\frac{(2n+1)v\pi}{2N}\right] \quad (4)$$

where,u, v = discrete frequency variables (0,1,2….N-1),f(m,n) = N*N image pixels (0,1,…N-1),F(u,v) = DCT result. The DCT implies an even extension of the original function. The use of cosine rather than sine functions is critical for compression, since it turns out that fewer cosine functions are needed to approximate a typical signal.



(a)

(b)

Figure 5.(a)Input image (b) DCT features visualizations

C.3.Gabor Features :

Gabor filters are defined by harmonic functions modulated by a Gaussian distribution.The Gabor filter has been used in many computer vision appliations including image compression,edge detection, texture analysis, object recognition an facial recognition. It is interesting to notice that in OCR area Gabor features is not that much popular as they have in face and Iris pattern recognition areas. This situation is difficult for the new comers to understand, especially considering the following facts:

1) Gabor features are well motivated and mathematically well-defined,
2) They are easy to understand, fine-tuned and easy to implement,
3) They have also been found less sensitive to noises, small range of translation, rotation, and scaling.

Introduction to Gabor Filter

Gabor filters have been used extensively in image processing, texture analysis for their excellent properties. Frequency and orientation representation of Gabor filters are similar to those of the human visual system, and they have been found to be particularly appropriate for texture representation and discrimination.

A Gabor Filter is a linear filter whose impulse response is defined by a harmonic function multiplied by a Gaussian function as given in the equation (5).

$$h(x, y) = g(x, y)\, s(x, y) \quad (5)$$

Where s(x, y) is a complex sinusoid, known as carrier and g(x, y) is a Gaussian shaped function, known as envelope.

D. CLASSIFICATION

Support Vector Machine is used as a classifier in this proposed method.SVMs are large margin classifiers.

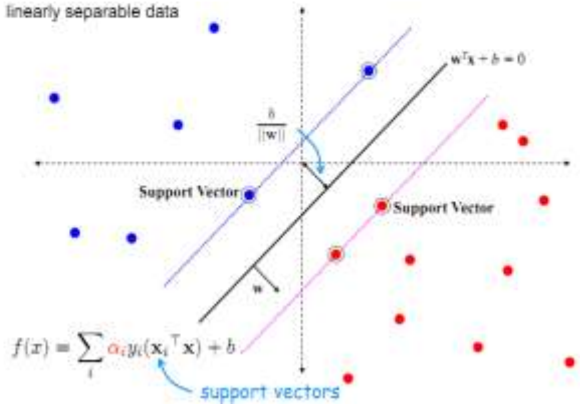


Figure 6. SVM Classifier [15]

SVM classifier is trained by a given set of training data and a model is prepared to classify test data based upon this model. A set of data is taken as input by the SVM classifier, it predicts and classify them in any one of the distinct classes. For multiclass classification problems, multiclass problems can be decomposed into multiple binary class problems and suitable combined multiple binary SVM classifiers can be defined. Several different types of kernel in SVM classifier are used. Commonly used kernels are:

Linear kernel, Polynomial kernel, Gaussian Radial Basis Function (RBF) and Sigmoid (hyperbolic tangent).

The performance of SVM depends on kernel used, kernel parameters and soft margin or penalty parameter C.

The most widely used choice is Linear kernel, which has dual parameter gamma and C. Since the database is huge one, linear kernel is optimal than RBF. And added to that RBF kernel results in time complexity, this proposed model is proceeded with the linear kernel. Default values of C (2.67) and Gamma (5.83) is used as mentioned in base paper.

V. RESULTS

The corresponding results for the propose method is tabulated as shown below **.**

Table 1 : For MNIST digit dataset

FEATURE EXTRACTED	ACCURACY (%)
Histogram of Gradients	88
Discrete Cosine Transform	86.28
Gabor Transform	89

Table 2 : For Handwritten character dataset

FEATURE EXTRACTED	ACCURACY (%)
Histogram of Gradients	92.36
Discrete Cosine Transform	90.45
Gabor Transform	92.05

VI. CONCLUSION

The strength of the selected feature and the effectiveness ofthe classifier are the two important key factors determining theperformance of a handwritten Character Recognition System.The result of the proposed system is better than publishedwork in the same area. It uses lesser number of features, which are highly uncorrelated.The computation time for recognition by using the extracted features will significantly reduced. Thus we can conclude that we have obtained the maximum average recognition rate as 87.76% for digits and 92% for characters approximately by using Histogram of gradients, Gabor Transform and Discrete Cosine Transform. The purpose of using Gabor Filters as mode of feature extractor is to promote its utility as major feature extraction technique in field of character recognition of the global language-English.

VII. FUTURE WORK

Very less literature is available on utilization of Gabor Filters for character Recognition. The work can be extended to increase the results by using or adding some more relevant features along with Gabor and DCT features.We can use some features specific to the mostly confusing characters, to increase the recognition rate. In future, design the systemusing different feature classifier pair for better recognition accuracy.

REFERENCES

[1]. Halima Begum and MuhammedMazharul Islam, “Recognition of Handwritten Bangla Characters using Gabor Filter and Artificial Neural Network”,Halima Begum et al , International Journal of Computer Technology &Applications, Vol8(5),618-621.

[2]. Nusaibath C, AmeeraMol P M,“Off-line Handwritten Malayalam Character Recognition Using Gabor Filters”,International Journal of Computer Trends and Technology (IJCTT) –volume 4 Issue 8–August 2013.

[3]. Sukhpreet Singh, Ashutosh Aggarwal, RenuDhir,”Use of Gabor Filters for Recognition of Handwritten Gurmukhi Character”,Volume 2,Issue 5, May 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.

[4]. Jie Yao, Patrick Krolak, and Charlie Steele,”The Generalized Gabor Transform”,IEEE TRANSACTIONS ON IMAGEPROCESSING, VOL. 4, NO. 7, JULY 1995.

[5]. Poovizhi P,”A Study on Preprocessing Techniques for the Character Recognition”,International Journal of Open Information Technologies ISSN: 2307-8162 vol. 2, no. 12, 2014.

[6]. Vladan Vuckovic, Boban Arizanovic, “Efficient character based segmentation approach for Machine typed documents”, Expert Systems With Applications, Volume 80, 2017, pp.210-231.

[7]. J.Pradeep,E.Srinivasan and S.Himavathi , ”Diagonal based feature extraction for Handwritten alphabets recognition system using Neural network”,IJCSIT,2011.3103.

[8]. Ranjan Jana, Amrita Roy Chowdhury, Mazharul Islam, ” Optical Character Recognition from Text Image ”,Intrenational Journal of Computer Applications Technology and Research Volume 3-Issue 4,239-243.

[9]. Jino P J , Kannan Balakrishnan , ”Offline Handwritten Recognition of Malayalam District Name - A Holistic Approach”IJET/2017v9i2/1170902250.

[10]. Sandhya Arora, ” Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition ”,IJCSI Volume 7,Issue 3,May 2010.

[11]. Nibaran Das , Brinda ban Das, Ram Sankar, Subhadip Basu, Mahantapas Kundu, Mita Nasipuri, ”Handwritten *Bangla* Basic and Compound character recognition using MLP and SVM classifier” ,Journal of Computing, Volume2, Issue 2, Feb 2010.

[12]. Gururaj Mukarambi, B.V.Dhandra, “A zone based character recognition for Kannada and English characters”, Procedia Engineering, Volume 37, 2012, pp. 3292-3299.

[13]. Ritesh Sarkhel, Nibaran Das, “A multi objective approach towards isolated handwritten and digit recognition”, Pattern Recognition, Volume 58,October 2016, Pages 172-189.

[14]. Ashfaqur Rahman ,Brijeshverma, “Effect of ensemble classifier composition on offline cursive character recognition”, Information Processing & Management, Volume 49, Issue 4, July 2013, Pages 852-864.

[15]. Anju K Sadasivana, T.Senthilkumar, “AutomaticCharacter Recognition in Complex Images”Procedia Engineering, Volume 30, 2012, Pages218-225.

[16]. Seiichi Uchida, Hiroaki Sakoe,” A Survey of Elastic Matching Techniques for Handwritten Character Recognition”, IEICE – Transactionson Information and Systems, Vols. E88 - D, no. 8, pp. 1781 - 1790, 2005.

[17]. D. Gabor, Theory of Communications, Journalof the Institute of Electrical Engineers, vol. 93,pp. 429-457, 1946.

[18]. S. Singh, A. Aggarwal, R. Dhir. Use of Gabor Filters for recognition of Handwritten Gurmukhi character, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 5,pp. 234 - 240, 2012.

[19]. Jomy John, Pramod K. V., Kannan Balakrishnan," Unconstrained Handwritten Malayalam Character Recognition using Wavelet Transform and Support vector Machine Classifier", International Conference oncommunication Technology and System Design, ELSEVIER 2011.

[20]. J. G. Daugman, "Complete discrete 2-D Gabortransforms by neural networks for image analysis and compression", IEEE Transaction on ASSP, Vo1.36,No.7, pp.1169-1179, 1988.

[21]. U.Pal, N.Sharma, R.Jayadevan "Handwriting Recognition in Indian Regional Scripts: ASurvey of Offline Techniques", ACM Transactions on Asian Language Information Processing, Vol. 11, No. 1, Article 1,Publication date: March 2012.

[22]. Vamvakas, G.; Gatos, B.; Petridis, S.Stamatopoulos, N., "An Efficient Feature Extraction and Dimensionality Reduction Scheme for Isolated Greek Handwritten Character Recognition", Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, vol.2, no., pp.1073-1077, 23-26 Sept. 2007.

[23]. U. Pal, T. Wakabayashi and F. Kimura,"Handwritten Bangla compound character Recognition using Gradient Feature", ICIT , 2007.