Pattern Recognition using Adaptive Optical Character Recognition Technique

Faisal Nabi Mir
Computer Science Engineering
Desh Bhagat University, Mandi Gobindgarh, Punjab, India

Khushboo Bansal
Assistant Professor
Department of Computer Science Engineering
Desh Bhagat University, Mandi Gobindgarh, Punjab, India

Abstract- Character recognition is a field of research in pattern recognition, artificial intelligence and machine vision. Existing methods of pattern recognition were time consuming, high cost and low recognition accuracy. The methods proposed in this research work attempts to overcome the limitations of existing methods to improve recognition accuracy. The techniques for the recognition of handwritten English text by segmenting and classifying the characters have been proposed in this paper. The problems in handwritten English text written by different persons are identified after carefully analyzing the text. To solve these problems new techniques have been developed for segmentation, feature extraction and recognition.

Keywords –Pattern recognition, Character recognition, Feature extraction, Adaptive histogram equalization.

I. INTRODUCTION

Pattern recognition [1] is a scientific discipline with the goal to classify objects into a number of categories or classes. Pattern recognition has a strong applied aspect, with contributions towards many facets of daily life.
Pattern recognition [2] is a branch of machine learning that emphasizes the recognition of data patterns or data regularities in a given scenario. Pattern Recognition is a subject researching object description and classification method.

1.1 CHARACTER RECOGNITION

Character recognition [4] is one of the fields of research in pattern recognition. Recognition of hand-written characters can be done either On-line or Offline as shown in figure 1.

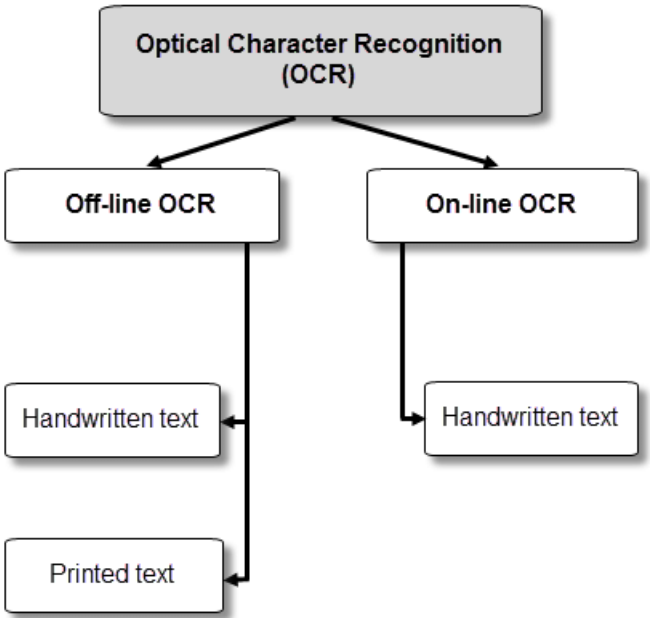

Figure 1. Types of character recognition

Character recognition is an easy task for human being, but for computer it is a difficult task. Computer recognizes it as document of pixels. Hence it is an important application of pattern recognition. It is used in banks, post-offices for processing of handwritten characters. Optical character recognition (OCR) is automatic reading of optically sensed document text materials to translate human-readable characters into machine-readable codes. On the other hand Handwritten Character Recognition is a difficult task when compared to OCR. Handwritten documents contain structurally several different handwritten styles for each symbol in the language.

Several applications [7] including mail sorting, bank processing, document reading and postal address recognition require off line handwriting recognition systems. As a result, the off-line handwriting recognition continues to be an active area of research towards exploring the newer techniques that would improve recognition accuracy. Selection of relevant feature extraction plays important role in performance of character recognition.

1.2 ARCHITECURE OF A GENERAL CHARACTER RECOGNITION SYSTEM

The major steps involved in recognition of characters include acquisition, pre-processing, segmentation, feature extraction and classification (as shown in figure 2).

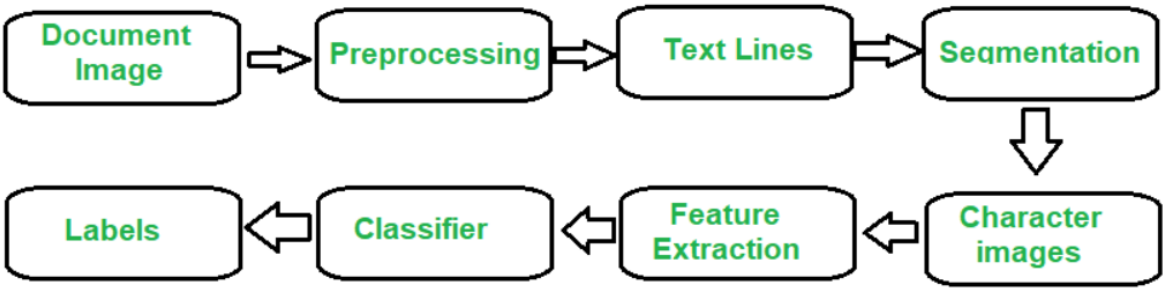

Figure 1.2 Architecture of a general character recognition system

1.2.1 Image Acquisition

Image Acquisition is the way toward getting a digitized picture from a certifiable source. Each progression in the procurement procedure may bring irregular changes into the estimations of pixels in the picture which is alluded to as commotion. A digital image is captured and stored into the computer for further processing.

The recognition system acquires a scanned image as an input image. Handwritten data samples for the experiment have been collected from 300 different individuals. Writers were provided with the plain A4 sheet and made to write uppercase and lowercase English characters. The images of the created database were scanned using HP-scan jet 5400c. The digitized images have been stored in PNG format. This created database is used in this research work.

1.2.2 Pre-processing

The purpose of pre-processing is to improve the quality of the image being processed [8]. There are reasons for the need of image pre-processing:

- improvement of image quality to meet the requirements of physician
- noise reduction
- contrast enhancement
- correction of missing or wrong pixel values
- elimination of acquisition-specific artifacts

The sequences of pre-processing steps are as follows:

1.2.2.1 Noise Removal

Noise is defined as any degradation in the image due to external disturbance [13]. Quality of handwritten documents depends on various factors including quality of paper, aging of documents, quality of pen, color of ink etc. Some examples of noise are salt and pepper noise, Gaussian noise. These noises can be removed to certain extent using filtering technique.

1.2.2.2 Image Enhancement

The main objective of image enhancement is to procedure the image in order that the output image will be better than data image. So this technique enhances and improves the quality of the image.

- Highlighting interesting details in images.
- Making images more visually appealing

1.2.2.3 Thresholding

The task of thresholding is to extract the foreground (ink) from the background (paper) [14]. Given a threshold, T between 0 and 255, replace all the pixels with gray level lower than or equal to T with black (0), the rest with white

(1). If the threshold is too low, it may reduce the number of objects and some objects may not be visible. If it is too high, we may include unwanted background information. The appropriate threshold value chosen can be applied globally or locally. Otsu's algorithm is the commonly used global thresholding algorithm.

1.2.2.4 Skeletonization

Skeltonization is an image preprocessing operation performed to make the image crisper by reducing the binary valued image regions to lines that approximate the skeletons of the region.

1.2.3 Segmentation

Image segmentation [15] is the one of the first step in image analysis and pattern recognition in the field of image processing and it is important and its necessary part of investigation framework, and which is one of the most troublesome assignments in picture preparing that, decides the nature of the last consequence of examination. It ought to be noticed that picture division technique is the procedure which it partitions a picture into various areas and where every district is homogeneous.

Segmentation step contains line segmentation, word segmentation and character segmentation. Methods for character segmentations are based on i) white space and pitch ii) projection analysis and iii) connected component labelling

Segmentation [16] is done to make the separation between the individual characters of an image. Segmentation of unconstrained handwritten word into different zones (upper middle and lower) and characters is more difficult than that of printed documents. This is mainly because of variability in inter-character distance, skew, slant, size and curved like handwriting. Sometimes components of two consecutive characters may be touched or overlapped and this situation complicates the segmentation task greatly. In Indian languages such touching or overlapping occurs frequently because of modified characters of upper-zone and lower-zone. Segmentation is an important stage, because the extent one can reach in separation of words, lines or characters directly affects the recognition rate of the script. Line, word and character segmentation will be done in segmentation unit.

1.2.4 Normalization

It is the process of converting the random sized image into standard sized image. This size normalization avoids inter class variation among characters. Bilinear, Bicubic interpolation techniques are a few methods for size normalization

1.2.5 Feature Extraction

Features are a set of numbers that capture the salient characteristics of the segmented image. There is different feature extraction methods proposed for character recognition. In feature extraction stage each character is represented as a feature vector, which becomes its identity. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements.

Due to the nature of handwriting with its high degree of variability and imprecision obtaining these features, is a difficult task. Feature extraction methods are based on 3 types of features:

- Statistical
- Structural
- Global transformations and moments

1.2.5.1 Statistical Features

Representation of a character image by statistical distribution of points takes care of style variations to some extent. The major statistical features used for character representation are:

- Zoning
- Projections and profiles
- Crossings and distances

1.2.5.2 Structural Features

Characters can be represented by structural features with high tolerance to distortions and style variations. This type of representation may also encode some knowledge about the structure of the object or may provide some knowledge as to what sort of components make up that object.

Structural features are based on topological and geometrical properties of the character, such as aspect ratio, cross points, loops, branch points, strokes and their directions, inflection between two points, horizontal curves at top or bottom, etc.

1.2.5.3 Global Transformations - Moments

In this, the Fourier Transform (FT) of the contour of the image is calculated. Since the first n coefficients of the FT can be used in order to reconstruct the contour, then these n coefficients are considered to be an n-dimesional feature vector that represents the character.

Central, Zenrike moments that make the process of recognizing an object scale, translation, and rotation invariant. The original image can be completely reconstructed from the moment coefficients.

1.2.6 Classification

The feature vector obtained from previous phase is assigned a class label and recognized using supervised and unsupervised method. The data set is divided into training set and test set for each character. Character classifier can be Bayes classifier, Nearest neighbour classifier, Radial basis function, Support vector machine, Linear discriminant functions and Neural networks. In this thesis, we use the neural networks for classification.

Neural network (NN) offers a promising solution as classifiers in the handwritten character recognition systems. But the classification capability depends on the architecture and learning rule. The network learns the characteristics of the pattern from the training data and then classifies a new test pattern to the appropriate class.

The simplest approach to handwritten English character recognition is to use the image as an input to the classifier. All the image pixels are used as features. Each pre-processed images is resized to maintain the aspect ratio. Two neural architectures, namely, (i) Feed forward NN and (ii) Radial basis function NN are investigated.

II.RELATED WORK

Anwar et al. [3] discussed that a fundamental problem in image deblurring is to recover reliably distinct spatial frequencies that have been suppressed by the blur kernel. To tackle this issue, existing image deblurring techniques often rely on generic image priors such as the sparsity of salient features including image gradients and edges. However, these priors only help recover part of the frequency spectrum, such as the frequencies near the high-end. To this end, they pose the following specific questions: (i) Does any image class information offer an advantage over existing generic priors for image quality restoration? (ii) If a class-specific prior exists, how should it be encoded into a deblurring framework to recover attenuated image frequencies? Throughout this work, they devise a class-specific prior based on the band-pass filter responses and incorporate it into a deblurring strategy. More specifically, they show that the subspace of band-pass filtered images and their intensity distributions serve as useful priors for recovering image frequencies that are difficult to recover by generic image priors. They demonstrate that our image deblurring framework, when equipped with the above priors, significantly outperforms many state-of-the-art methods using generic image priors or class-specific exemplars.

Prasad et al. [4] discussed that Character recognition is one of the fields of research in pattern recognition. Recognition of hand-written characters can be done either On-line or Offline. Not much substantial work has been published in the past on the development of hand-written character recognition (HWCR) systems for Telugu text. However none of them give 100% accuracy in recognition of Telugu characters. Their effort is intended to improve the accuracy in Telugu character recognition. Zonal based feature extraction is used in the present proposed work. They presented two methods for this purpose. First method is based on Genetic Algorithm and uses adaptive zoning topology with extracted geometric features. In second method, zoning is done in static way and uses distance, density based features. In both the contexts, we use K-Nearest Neighbor (KNN) algorithm for classification purpose. Using first method we obtained accuracies of 100 percent and 82.4 percent for 11 and 50 symbols respectively. Using second method we obtained accuracies of 100 percent and 88.8 percent for 11 and 50 symbols respectively.

Fedorovici et al. [5] presented the architecture of an Optical Character Recognition (OCR) technology application based on two approaches, a multilayer neural network and a Support Vector Machine (SVM) classifier using Zernike moments for feature extraction. The performance comparison of the two approaches is based on the similar layout of most of the characters that must be recognized. The comparison shows that the improvement of the processing performance can be obtained by creating classes of blobs that use geometric similarities, and doing OCR only on the representative blob from each class.

Chen et al. [6] purposed a handwritten character recognition algorithm based on artificial immune In order to improve the rate of character recognition and decrease the time of recognition training, referencing to immune biological principle. The antigen and memory cell in the artificial immune system are described. The equations of clone selection principle and of evolving memory cell are established. Finally, the process of character recognition is given. The algorithm steals the merit of self-adaptive learning, and immune memory in the biology immune system, which can also be applied to abnormity detection and pattern recognition.

Sahu et al. [7] presented detailed review in the field of Off-line Handwritten Character Recognition in this paper. Various methods are analyzed that have been proposed to realize the core of character recognition in an optical character recognition system. The recognition of handwriting can, however, still is considered an open research problem due to its substantial variation in appearance. Even though, sufficient studies have performed from history to this era, paper describes the techniques for converting textual content from a paper document into machine readable form. Offline handwritten character recognition is a process where the computer understands automatically the image of handwritten script. This material serves as a guide and update for readers working in the Character Recognition

area. Selection of a relevant feature extraction method is probably the single most important factor in achieving high recognition performance with much better accuracy in character recognition systems.

Hamad et al. [9] discussed that in many different fields, there is a high demand for storing information to a computer storage disk from the data available in printed or handwritten documents or images to later re-utilize this information by means of computers. Optical character recognition is an active research area that attempts to develop a computer system with the ability to extract and process text from images automatically. Some major challenges need to be recognized and handled in order to achieve a successful automation. The font characteristics of the characters in paper documents and quality of images are only some of the recent challenges. In this paper they investigate OCR in four different ways. First they give a detailed over view of the challenges that might emerge in OCR stages. Second, they review the general phases of an OCR system such as pre-processing, segmentation, normalization, feature extraction, classification and post-processing. Then, we highlight developments and main applications and uses of OCR and finally, a brief OCR history are discussed.

Chandarana et al. [10] explains comparative analysis between Random Transform and Hough Transform, which are applied for error detection and correction. This paper explains implementation of OCR in Matlab, compared with current working method of OCR. This system achieved recognition rate near about 92%.

Prasad and Sanyal [11] discussed the recognition of off-line English character. This explains a new model Hidden Markov Model (HMM) for character recognition. The Novel feature Extraction method is used for implementing HMM. By collecting 13000 samples from 100 writers they have tested performance of OCR technique and got accuracy of near about 94%.

Tiwari et al. [12] implemented the OCR technique in Matlab. This paper explained how Matlab is more convenient and effective for OCR technique. The performance of OCR has been tested with samples in this approach.

III. RESULTS

An implementation is the realization of a technical specification or algorithm as a program, software component, or other computer system through computer programming and deployment. The proposed method is simulated using MATLAB. To see the qualitatively as well as quantitatively performance of the proposed algorithm, some experiments are conducted on several images. The effectiveness of the approach has been even exploitation completely different images.
The figures from 3 to 8 show the different images which consists of original images and output images.

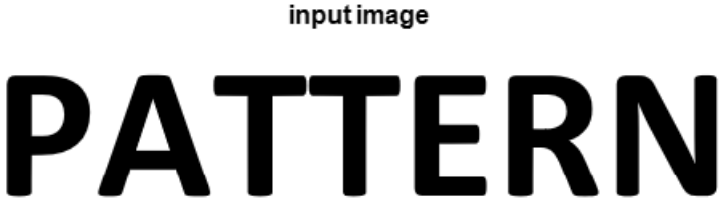

Figure 3. Input image

Using MATLAB ‘imread’ function of, a noiy scanned RGB image is loaded to the input system. Figure 3 shows the input noisy image.

Figure 4. Gray scale image

Now input image is converted into gray scale image for various image processing operations. After performing Grayscale conversion by ‘rgb2gray’, figure 4 is thus obtained; ‘rgb2gray’ eliminates the hue and saturation information of the RGB image while retaining the luminance.

Figure 5. Adaptive histogram equalized image

Histogram equalization is a global type method that gives the histogram of the whole pixels (0-255) range and this technique can be applied in image understanding problems to normalize variations in illumination. Figure 5 shows the image after applying the adaptive histogram equalization.

Figure 6. Binary image

The grayscale image is further processed into a binary image which replaces all pixels in the input image with luminance greater than level with the value 1 (white) and replaces all other pixels with the value 0 (black). Figure 6 shows the outcome after binarization (using 'im2bw').

Figure 7. Character extraction

The image is then sent through connectivity test in order to check for the maximum connected components and the properties of each component, which is in the form of a box. After locating the box, the individual characters are then cropped into different sub images that is the raw data for the following feature extraction routine. The size of the sub-images is not fixed since they are exposed to noises which will affect the cropping process to vary from one to another. This will causing the input of the network become non- standard and hence, prohibit the data from feeding through the network. To avoid this, the sub-images have to be resized and then by finding the average value in each 10 by 10 blocks, the inputs for the network can be determined. Figure 7 shows the character extraction. By this, extraction of the character is possible and could be passed to another stage for future classification and training purpose of the neural network.

Recognized characters are shown in the command window of the Matlab as shown in figure 8.

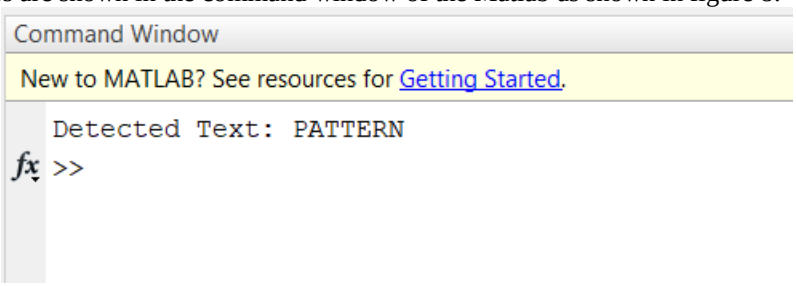

Figure 8. Recognized characters in the command window

IV.CONCLUSION

Pattern recognition is used for most of the pre-processing and analysis, including the identification of the characters and classification of characters. For many document-input tasks, pattern recognition is the most cost-effective and speedy method available. In this paper, classification of the characters is done using neural networks.

V. REFERENCE

[1] Poonam Mahana and Gurbhej Singh (2015), "Comparative Analysis of Machine Learning Algorithms for Audio Signals Classification", IJCSNS International Journal of Computer Science and Network Security, Vol.15, No.6.

[2] https://www.techopedia.com/definition/8802/pattern-recognition-computer-science

[3] Saeed Anwar, Cong Phuoc Huynh, Fatih Porikli (2018), "Image Deblurring with a Class-Specific Prior", IEEE, Transactions on Pattern Analysis and Machine Intelligence.

[4] Sanugula Durga Prasad and Yashwanth Kanduri (2016), "Telugu Handwritten Character Recognition using Adaptive and Static Zoning Methods", Proceedings of the 2016 IEEE Students' Technology Symposium.

[5] Lucian-Ovidiu Fedorovici (2011), "A Comparison between Two Character Recognition Approaches", Series: Automatic Control and Robotics Vol. 10, No 2, pp. 125-140.

[6] Yuefeng Chen, Chunlin Liang, Lingxi Peng, Xiuyu Zhong (2010), "A Handwritten Character Recognition Algorithm based on Artificial Immune", International Conference on Computer Application and System Modeling.

[7] Vijay Laxmi Sahu1, Babita Kubde (2013), "Offline Handwritten Character Recognition Techniques using Neural Network: A Review", International Journal of Science and Research (IJSR), India, Vol. 2, Issue 1.

[8] Chaganti, Venkata Ravikiran (2005), "Edge Detection of Noisy Images using 2-D Discrete Wavelet Transform" Electronic Theses, Treatises and Dissertations.

[9] Karez Abdulwahhab Hamad and Mehmet Kaya (2016), "A Detailed Analysis of Optical Character Recognition Technology", IJAMEC, pp. 244–249.

[10] Jagruti Chandarana, Mayank Kapadia (2014), "Optical Character Recognition", International Journal of Emerging Technology and Advanced Engineering, Vol. 4, Issue 5.

[11] Binod Kumar Prasad, Goutam Sanyal (2012), "A model Approach to Off-line English Character Recognition", International Journal of Scientific and Research Publications, Vol. 2, Issue 6.

[12] Sandeep Tiwari, Shivangi Mishra, Priyank Bhatia, Praveen Km. Yadav (2013), "Optical Character Recognition using MATLAB", International Journal of Advanced Research in Electronics and Communication Engineering, Vol. 2, Issue 5.

[13] Jomy John, Pramod K. V, Kannan Balakrishnan (2011), "Handwritten Character Recognition of South Indian Scripts: A Review", National Conference on Indian Language Computing, Kochi, Feb 19-20.

[14] Snigdha Mohanty and Mahesh Prasad Sahoo, "Edge detection: A comparison".

[15] Li, C., Li, Y. and Wu, X., 2012, ''Novel Fuzzy C-Means Segmentation Algorithm for Image with the Spatial Neighborhoods", In Proceedings of IEEE Trans. Engineering, Conference, pp. 1-6.

[16] Vijay Laxmi Sahu and Babita Kubde (2013), "Offline Handwritten Character Recognition Techniques using Neural Network: A Review", International Journal of Science and Research, India, Vol. 2, Issue 1.