International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2019): 7.583

Character Recognition using KNN Algorithm

Sai Jahnavi Bachu

B. Tech in K L University
E-mail: *bachusaijahnavi[at]gmail.com*

Abstract: ***Optical Character Recognition System offers the human machine interaction which is commonly used for several important applications. A lot of study has been conducted to accomplish the work on character recognition in different languages. This paper represents a technique for recognizing the characters from an image with noise using Optical Character Recognition (OCR). The important steps of this system are pre-processing of the text including converting the text image to black/white and remove the noise from the text image, segmentation of the text image to each character, Feature Extraction using KNN and classification. The System is implemented using Anaconda software application program. Noise is removed from all the text images. The quality ofthe input document is very important to achieve high accuracy rate.***

Keywords: Classification, Extraction, Noise Removal, Optical Character Recognition, Segmentation

1. Introduction

The Optical Character Recognition (OCR) field has gained more attention in the recent years because of its importance and many applications. Character recognition is the process of detecting and recognizing characters from the given input image (handwritten, printed, or typewritten) and converts it into American Standard Code for Information Interchange (ASCII) or other equivalent machine editable form. Character recognition is one of the most interesting areas of pattern recognition and artificial intelligence. Optical character recognition (OCR) is the branch of technology which deals with the automatic reading of text.

The goal is to classify the optical patterns (often contained in a digital image) corresponding to alpha-numeric or another characters. And to imitate the human ability to read - at a much faster rate - by associating symbolic identities with the images of characters. As emphasis shifts from recognizing individual characters to recognizing the whole words and pages, more general terms being used which includes optical text recognition and document image processing. The process of OCR involves several steps including segmentation, feature extraction, and classification. Recognizing the text in scene images is the most challenging due to it's many possible variations in backgrounds, textures, fonts, and lighting conditions which are present in such images.

Three major types of character recognition in computer science are in place:

1) Optical Character Recognition(OCR):
   Techniques based solely on image processing techniques which include extracting features from the image, comparing these features with predefined ones and finally recognizing characters.
2) Intelligent Character Recognition(ICR):
   Includes machine learning algorithms within the recognition process. Also targets handwritten script or cursive script one glyph or character at a time.
3) Intelligent word recognition(IWR):
   It targets handwritten printscript or cursive text, one word at a time. This is especially useful for languages where glyphs are not separated in the given cursive script.

Image processing is a method that performs some operations on an image, in order to get an enhanced image or to extract the useful information from it. It is a type of signal processing in which input is an image and output may be an image or characteristics/features associated with that particular image. Nowadays, image processing is one of the rapidly growing technologies. It forms core research area within engineering and computer science disciplines. Image processing basically includes the following three steps:

Importing the image through image acquisition tools;
Analyzing and manipulating the image;
Output in which result can be modified image or report that is based on image analysis;

There are two types of methods used for image processing namely, analogue and digital image processing. Analogue image processing is used for the hard copies like printouts and photographs. Image analysts use various fundamentals of interpretation while using these visual techniques. Digital image processing techniques help in manipulation of the image by using computers. The three general phases that all types of data have to undergo while using digital technique are pre-processing, enhancement and display, information extraction.

Recognition of cursive text is the present active area of research, with recognition rates even lesser 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 more 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 likely. The shapes of individual cursive characters themselves simply do not contain enough information accurately (greater than 98%) recognize all the hand written cursive script.

2. Literature Survey

Character recognition is not a new problem but can be traced back to systems before the inventions of computers. The

earliest OCR systems were not computers but mechanical devices that are able to recognize characters, but are very slow of speed and low at accuracy. In 1951, M. Sheppard invented a reading and robot GISMO that can be considered as the most earliest work on modern OCR. GISMO can read musical notations as well as words on a printed page one by one. However, it can only recognize 23 characters. The machine also has the capability to copy a typewritten page. J. Rainbow, in 1954, devised a machine which can read the uppercase typewritten English characters, one per minute. The early OCR systems were neglected due to errors and slow recognition rate. Hence, much research efforts were not put on the topic during 60's and 70's. The developments were done on government agencies and large corporations like banks, newspapers and airlines etc. Because of the complexities associated with recognition, it was felt that there should be standardized OCR fonts for making the task easy for recognition for OCR.

Hence, OCRA and OCRB were developed by ANSI and EMCA in 1970, which provided comparatively acceptable recognition rates. During the past thirty years, substantial research has been conducted on OCR. Which has lead to the emergence of document image analysis (DIA), multi-lingual, handwritten and omni-font OCRs. Despite these extensive research efforts, the machine's ability to read text is still far below the human. Hence, the current OCR research is being conducted on improving accuracy and speed of OCR for diverse style documents printed/ written in unconstrained environments. There has not been availability of any open source or commercial software available for complex languages like Urdu or Sindhi etc.

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method which is used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or the regression process.

In k-NN classification, the output is a class membership. An object which is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to a class of that particular single nearest neighbor.

A brief description on the history of OCR is as follows. In 1929 Gustav Tauschek obtained a patent on OCR in Germany, followed by Handel who obtained a US patent on OCR in USA in 1933. In 1935 Tauschek was also granted a US patent on his OCR method. Tauschek's machine was a mechanical device which used template and a photo detect. RCA engineers in 1949 worked on the first primitive computer-type OCR that helps blind people for the US Veterans Administration, but instead of converting the printed characters to machine language, their device converted it to machine language and then spoke the letters. It is too expensive and was not used after testing. In1978

Kurzweil Computer Products began selling a commercial version of the optical character recognition computer program. LexisNexis was one of the first customers, and bought the program to upload paper legal and news documents onto its nascent online database.

The methods are tedious and are highly involved in computational power. In Di gesu the idea of using both intensities and spatial information has been considered to take into account local information used in human perception. The number of new methodologies and strategies were proposed over the past years to find global as well as local solutions in nonlinear multimodal function optimization. In addition attempts were also been made to use Fuzzy Logic, Artificial Neural Network for optical character recognition. With the help of crowding multiple peaks can be maintained in multimodal optimization problem. Crowding method is extremely reliable in detecting the peaks on bimodal histogram. It makes use of an OCR system which uses the histogram equalization to extract images. The histogram used by the mentioned algorithm is bimodal in nature so it can be divided into two classes. Genetic algorithm is further used to select the threshold from the histogram for extracting the object from the background.

Claudiu was investigated using simple training data pre-processing which gave experts with less errors correlated than that of different nets trained on the same or bootstrapped data. Hence committees that simply average the expert outputs considerably improve recognition rates. Our committee-based classifiers of isolated handwritten characters were the first on par with human performance and can be used as the basic building blocks of any OCR system (all our results were achieved by software running on powerful yet cheap gaming cards).

Georgios has represented a methodology for offline handwritten character recognition. The proposed methodology relies on a new feature extraction technique based on recursive subdivisions of the character image so that the resulting sub-images at every iteration were balanced (approximately equal) numbers of foreground pixels, as far as this is possible. Feature extraction is allowed by a two- stage classification scheme based on the level of granularity of the feature extraction method. Classes with high values in the confusion matrix are merged at a certain level and for each group of merged classes, granularity features from the level that best distinguishes them were employed. Two handwritten character databases (CEDAR and CIL) as well as two handwritten digit databases (MNIST and CEDAR) which were used in order to demonstrate the effectiveness of the proposed technique.

Sankaran presented a novel recognition approach which results in a 15% decrease in word error rate on heavily degraded Indian language document images. OCRs have considerably good performance on good quality documents, but easily fail in presence of degradations. Also, classical OCR approaches perform poorly over complex scripts such as those for Indian languages. Sankaran addressed these issues by proposing to recognize character n-gram images, which are basically groupings of consecutive character/ component segments. Their approach was unique, since they use the character n-grams as a primitive for recognition rather than for post-processing. By exploring the additional

content present in the character n-gram images, we enable better disambiguation between confusing characters in the recognition phase. Labels obtained from recognizing the constituent n- grams are then fused to obtain a label for the word that emitted them. Their methods are inherently robust to degradations such as cuts and merges which are common in digital libraries of scanned documents. We also present a reliable and scalable scheme for recognizing character n-gram images.

Zhang discussed the misty, foggy, and hazy weather conditions lead to image color distortion and reduce the resolution and the contrast of the observed object in the outdoor scene acquisition. In order to detect and remove haze, this article proposes a novel effective algorithm for visibility enhancement from a single gray or color image. It can be considered that the haze mainly concentrates in one component of the multilayer image, the haze-free image is constructed through haze layer estimation based on the image filtering approach using both low-rank technique and the overlap averaging scheme. By using parallel analysis with Monte Carlo simulation from the coarse atmospheric veil by the median filter, the refined smooth haze layer is acquired with both less texture and retaining depth changes. With the use of dark channel prior, the normalized transmission coefficient is calculated to restore fogless image. Experimental results show that the proposed algorithm is a simpler and efficient method for clarity improvement and contrast enhancement from a single foggy image. Moreover, this can be comparable with the state-of-the-art methods, and even has better results than them.

3. Applications of OCR are:

OCR can be used for digitizing printed texts so that it can be Electronically edited, searched, stored more compactly, displayed on-line, and used in Machine Processes such as Machine Translation, Text- to Speech, Key Data and Text Mining.

Postal services use the OCR to read addresses from the letter envelopes.

It is widely used for information entry from the passports, invoices, bank statements, voter ID forms etc.

4. Theoretical Analysis

Existing Theory:
In this contemporary world there is a growing demand for users to convert the printed document in to an electronic document for maintaining the security of the data. Hence the basic OCR system was invented which converts the data available on papers in to computer process able documents, So that the documents can be editable and also be reused. The previous system of OCR on a grid infrastructure is just OCR without the grid functionality. This is an existing system that deals with the homogeneous character recognition or character recognition of single languages.

Limitations of Existing Theory:
The drawback in the existing OCR systems is they only have the capability to convert and recognize only the documents of English or a specific language only. That is, the older 4 OCR system is uni-lingual.

Proposed Theory:
In this paper, automatic character recognition system is implemented on text images which have noise. The system involves different stages including: Image acquisition, Preprocessing, Segmentation, Feature Extraction and Classification. The recognition of text and characters begin with obtaining a digitalized image of the text using a suitable scanning system. In second stage the Pre-processing of the image goes on Binarization and on to noise removal. In the third stage the segmentation of the text to individual characters goes on. Segmentation of the individual text to characters is an essential and difficult stage in text recognition system. The fourth stage is the Feature Extraction stage. The final stage is the Classification Process. This will compare feature vectors to the different models and find the closest match.

Advantages of Proposed Theory:
In this proposed theory the noise of the images is completely removed. We get the image with perfect accuracy without any disturbances. It recognizes both alphabets and numerals in both the hand written and digital Images.

5. Character Recognition Process

5.1 Image Pre-Processing

The digital image may contain a certain amount of noise depending on the resolution of the scanner. The recognition rates would be poor since the character may be smeared or broken. This can usually be eliminated by using a pre-processing technique to smooth the digital image.

5.2 Binarization

Binarization is a process of converting pixel image into a binary image. Binary image is also called as bi-level or two-level has only two possible values for each pixel which represent the color black or white.

5.3 Remove Noise

Noise can give a significant impact on the quality of digital images. Various techniques such as Mean Filter, Median Filter, Local Pixel Grouping, Adaptive Filter, Wiener Filter, etc. can be used to connect unconnected pixels, to remove isolated pixels, to smooth pixels boundary. Median filter is used to remove the noise that is added on character‘a’.

5.4 Segmentation

Segmentation is a process among the most crucial and is an essential step in an OCR. The most of optical character recognition techniques will segment the words into individual characters which can be recognized individually. We must find the regions of the document in which data are printed and separate them from figures and graphics. A poor segmentation process gives misrecognition or rejection.

5.5 Line Segmentation Process

The lines of a text block were discovered by checking or scanning the input image horizontally. Frequency of black pixels in each row is counted in order to build the row histogram. When black pixels frequency in a row is zero it indicates a boundary between two white pixels and consecutive lines.

5.6 Word Segmentation Process

When a line has been discovered, after that each line is scanned vertically for word segmentation. Number of black pixels in each column is determined to build column histogram. When no black pixel is found in vertical scan which is proved to be the space between two words. Therefore, we can separate the words.

5.7 Feature Extraction

The main purpose of feature extraction process is to capture the important characteristics of the symbols, which are usually accepted that it is among the most difficult problems of pattern recognition. Two kinds of features are found, statistical features and structural features. The majority of researchers accept that statistical features can be quickly found using those easy methods and might carry out high recognition results especially in closed testing data. Structural features are more conformed to the intuitive thinking of human mind, its more robust for the deformation of symbols. However, they usually depend on human summarized rules for the recognition algorithm.

5.8 Zoning

Zoning process is a well-known technique used in character recognition. In this technique, the rectangle shaped character images are divided into a number of overlapping or non-overlapping regions (zones) of predefined sizes. These predefined sizes are usually of the order 2×2, 3×3, 4×4 etc. Then features are computed for each zone. The average pixel density was found by dividing the number of foreground pixels by the total number of pixels in each zone.

5.9 The Classification Process

The character classification process would be used to assign certain part of text to one or more predefined classes or categories. This part of text could be a document, news article, search query, email, tweet, support tickets, customer feedback, user product review etc. Applications of classification include categorizing newspaper articles and news wire contents into topics, organizing web pages into hierarchical categories, filtering spam email, sentiment analysis, predicting user intent from search queries, routing support tickets, and analyzing customer feedback. There are two steps in creating a classifier: training and testing

5.10 K-Nearest Neighbor (K-NN)

K-Nearest Neighbor algorithm is a non-parametric machine learning algorithm which is used for classification. The view point behind this algorithm is very straightforward. To classify a new character, the system discovers the k nearest neighbors among the training datasets, and uses the categories of the k nearest neighbors to weight the category candidates. Several researchers had discovered that the k-NN algorithm comes up with very good performance for character recognition in this research projects on different data sets. A distance function is needed to compare point similarity. Euclidean Distance can be used between the test point and all the reference points in order to find K nearest neighbors, and then arrange the distances in ascending order and take the reference points corresponding to the k smallest Euclidean Distances. A test sample is then attributed the same class label as the label of the majority of its K nearest neighbors. Euclidean Distance can be calculated using the equation below:

The overall performance of this algorithm very much depends on two conditions or factors, that is, a suitable similarity function and an appropriate value for the parameter k.

6. Implementation

6.1 Image Pre-Processing

Image was selected from specific file and then converted to gray scale image. Median Filter is used to remove noise from the selected image by hitting the Remove Noise button. The last step of pre-processing is to convert the image to black and white just by clicking on the Convert to Binary button. In order to maintain uniform size of all the character images, they are resized into a standard dimension.

6.2 Segmentation and Feature Extraction

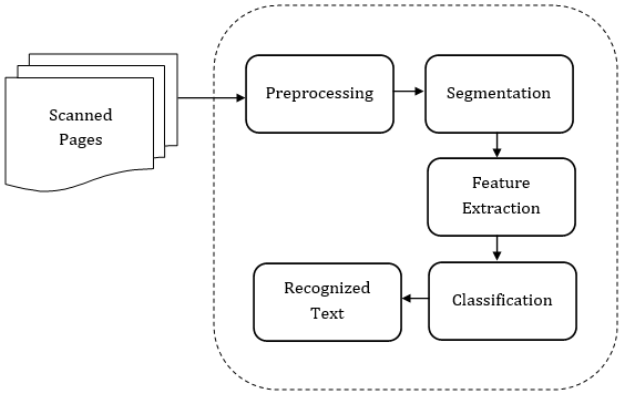
Segmentation and feature extraction are applied to the image after pressing the OCR button. After that the recognition process will continue to recognize characters inside the image and display the result. Segmentation and feature extraction process is powerful on image having no noise and acceptable on image with noise. In the proposed system, GUI selected image will be displayed on the left side and the result will be displayed on the right side.

The simplest method of segmentation is called thresholding method. This method is based on a clip-level (or threshold value) to turn a gray-scale image into a binary image. The key for this method is to select the threshold value (or values when multiple-levels are selected). Several popular methods are used in industry including the maximum entropy method, balanced histogram thresholding, Otsu's method (maximum variance), and k-means clustering. Recent methods have developed for thresholding computed tomography (CT) images. The key idea is that, unlike Otsu's method, the thresholds were derived from radiographs instead of an image.

New method suggests the usage of multi-dimensional fuzzy rule-based non-linear thresholds. In this, decision over each pixel's membership to a segment is based on multi-dimensional rules derived from fuzzy logic and evolutionary algorithms based on image lighting environment and application

6.3 Recognition Process

The proposed recognition system performance was evaluated on large number of images using two different datasets containing the same parameters. An independent English dataset was also created for numbers and characters. Testing is performed in two stages (on documents having no noise and on the documents with noise). The K-NN classifier is used on datasets for classification process. Three options can be occurred (text is localized and recognized which means that its matched, text is localized correctly but not recognized meaning that it is not matched, or text is not localized at all meaning that the text is not found).



6.1.1 Image representing the process for character recognition

7. Conclusion

In this paper, an effective character recognition system using OCR has been proposed. The proposed system is based on the image pre- processing, characters segmentation, feature extraction, and classification process. English numbers and characters dataset is created. Two types of images are used (images having no noise and images with noise). k-Nearest Neighbor algorithm is used for classification process. The last step, Euclidean Distance is used as a distance function which is needed to compare the similarity. The proposed character recognition system performance was evaluated and high rate of recognition was achieved.

8. Future Scope

The proposed algorithms are used for segmentation of handwritten and digital English characters which can be extended further for recognition of the other Indian scripts. The proposed algorithm of the segmentation can be modified further to improve the accuracy of segmentation. Many New features can be added to improve the accuracy of recognition. These algorithms are allowed to try on large databases of handwritten and digital English character. There is need to develop the standard database for recognition of the characters. The proposed work can be extended to work on degraded text or on broken characters. Recognition of half character and compound character can be done to improve recognition rate of a word.

References

[1] Das, R.L., B.K. Prasad & G. Sanyal. HMM based Offline Handwritten Writer Independent English Character Recognition using Global and Local Feature Extraction. International Journal of Computer Applications 46: 45–50(2012).
[2] Pradeepa, J., E. Srinivasana & S. Himavathib. Neural Network based Recognition System Integrating Feature Extraction and Classification for English Handwritten. International Journal of Engineering 25: 99–106(2012).
[3] Ganapathy, V. & K.L. Liew. Handwritten Character Recognition using Multiscale Neural Network Training Technique. World Academy of Science, Engineering and Technology 1:32–37(2012).
[4] Tokas, R. & A. Bhadu. A comparative analysis of feature extraction techniques for handwritten character recognition. International Journal of Advanced Technology & Engineering Research 2: 215– 219(2012)
[5] Sampath, A, C. Tripti & V. Govindaru. Freeman Code based Online Handwritten Character Recognition for Malayalam using Backpropagation Neural Networks. International Journal on Advanced Computing 3:51–58(2012)
[6] Pradeep, J., E. Shrinivasan & S. Himavathi. Diagonal based Feature Extraction for Handwritten Alphabets Recognition System using Neural Network. International Journal of Computer Science & Information Technology (IJCSIT) 3:27–38(2011)