Extraction and Recognition of Handwritten Hindi and Gujarati Character Using Artificial Neural-network Approach

Prof. Abhishek Mehta
PhD Research Scholar
Calorx Teachers University, Ahmadabad.
Assistant Professor at PICA, Parul University
Post Limda, Waghodia, Gujarat, 391760, India

Dr. Ashish Chaturvedi
Department of Computer Science
Calorx Teachers University, Ahmadabad.

Abstract—Hindi is that the most usually auditory communication in India, with in more than three hundred million speakers. As there's no division between the characters of writings written in Hindi as there's in English, the Optical Character Recognition (OCR) frameworks created for the Hindi language convey a poor recognition rate. During this paper we have a tendency to propose AN OCR for written Hindi content in Devanagari script content, utilizing Artificial Neural Network (ANN), that improves its productivity. one in every of the numerous functions behind the poor recognition rate is mistake in character division. The closeness of contacting characters within the examined records more entangles the division procedure, creating an interesting issue once designing a compelling character division methodology. Pre-processing, character division, embrace extraction; lastly, grouping and recognition area unit the important advances that area unit pursued by a general OCR. The pre-processing tasks thought of inside the paper conversion of gray scaled footage to binary footage, image rectification, and segmentation of the document’s matter contents into paragraphs, lines, words, thus at the extent of basic symbols, the basic symbols, obtained as a result of the essential unit from the segmentation methodology, recognized by the neural classifier. Neural Network is one in every of the foremost widely used and common techniques for character recognition downside. This paper discusses the classification and recognition of written Hindi Vowels and Consonants mistreatment Artificial Neural Networks. The vowels and consonants in Hindi characters are often divided in to sub teams supported bound vital characteristics for every cluster, a separate network is meant and trained to acknowledge the characters that belong to it cluster.

Keywords- Pattern Recognition, Character Recognition, Artificial Neural Network, Feature Extraction, Thinning, OCR, Pre-Processing, Segmentation, Feature Vector, Classification, Noise Removal.

Pattern Recognition is outlined because the field involved with machine recognition of significant regularities in shouting and complicated environments [1]. There square measure varied applications of pattern recognition like character recognition, online signature verification, and face recognition so on. Character Recognition is that the electronic conversion of scanned pictures of written or written text into computer readable text. Character recognition system is that the base for several differing types of applications in numerous fields, several of that we have a tendency to use in our daily lives. Hindi character recognition is that the difficult downside in Pattern Recognition and Neural Networks is one in every of the foremost normally used techniques for character recognition and classification because of their learning and generalization skills. This paper describes and discusses the classification and recognition of written Hindi characters victimisation Artificial Neural Networks. The introduction is roofed into 3 sub-sections. the primary defines the OCR and its basic applications, the second is regarding OCR generally, and therefore the third is regarding Nagari script, the mother script of the Hindi language.

What is Handwriting Recognition?

The importance of the piece of paper cannot be ignored in enhancing the people's memory and in facilitating communication between people. It is used for both personal (letters, notes, addresses on envelopes etc.) and business communications (bank cheques, tax forms, admission forms etc.) between person to person and for communications written to ourselves (reminders, lists, diaries etc). Handwriting is the most common and natural means of communication for humans. The concept of handwriting is very old and attributed by many civilizations and cultural ages. However, the solitary purpose is to facilitate communication and expand human memory.

"Handwriting Recognition is a process which allows computers to recognize written or printed characters such as numbers or letters and to change them into a form that the computer can use for editing and searching."

What is Optical Character Recognition?

CR (optical character recognition) is that the recognition of written or written communication characters by a laptop. This involves icon scanning of the text character-by-character, analysis of the scanned-in image, and so translation of the character image into character codes, like code, usually employed in processing. In OCR process, the scanned-in image or image is analysed for light-weight and dark
areas so as to spot every alphabetic letter or numeric digit. Once a personality is recognized, it's regenerate into ANN code. Special circuit boards and laptop chips designed expressly for OCR square measure accustomed speed up the popularity method. OCR is being employed by libraries to digitalize and preserve their holdings. OCR is additionally accustomed method checks and master card slips and type the mail. Billions of magazines and letters square measure sorted a day by OCR machines, significantly dashing up mail delivery.

II. REVIEW OF EARLIER APPROACHES

A good text recognizer has many commercial and practical applications such as processing cheques in banks, documentation of library materials, extracting data from paper documents, searching data in scanned book, automation of any organization like post office, which involve lot of manual task of interpreting text. The problem of text recognition has been attempted by many different approaches; some of them are Template matching, Feature extraction, Geometric approach and neural networks. Template matching approach is one of the most simplistic approaches. This is based on matching the stored data against the character to be recognized. Template matching involves determining similarities between the given template and stored database and output the image that produces the higher similarity measure. This technique works effectively with recognition of standard fonts, but gives poor performance with handwritten characters, noisy characters and deformed images.

The objective of feature extraction is to capture the essential characteristics of the symbols and this is one of the most difficult problems of pattern recognition. In this approach, statistical distribution of points is analyzed and orthogonal properties are extracted. For each symbol a feature vector is calculated and stored in database, and recognition is performed by finding distance of feature vector of input image with those stored in the database and giving the symbol with minimum deviation. This is very sensitive to noise and edge thickness, but performs well on handwritten character set. In geometric approach an attempt is made to extract features that are quite explicit and can be very easily interpreted. These features depend upon the physical properties, such as number of joints, relative position; number of end points, aspect ratio etc. Classes formed on the basis of these geometric features are quite distinct, with not much of overlapping. The main draw back with this approach is that this approach depends heavily on the character set. Neural network techniques are more popular to perform Character Recognition. It has been reported that Neural Networks could produce high recognition accuracy. Neural Networks with various architectures and training algorithms have been applied successfully for Character recognition. In this, neural network is first trained by the multiple sample images of each alphabet. Then, in the recognition processes, the neural network recognizes the given input symbol. Neural networks are capable of providing good recognition even at the presence of noise but the drawback is they require a lot of training time. Character recognition remains a highly challenging task. Hindi character recognition is one of the most difficult tasks of optical character recognition. This section gives a brief overview of related research work. The research work pertaining to character recognition of Indian languages is very limited.

Dr. P.S. Deshpande et.al, proposed a novel methodology on character encoding and ordinary articulations for shape recognition in their paper [2]. The strategy is autonomous of the particular part of individual shapes, for example, thickness of line, size of character and shapes. In this, highlights are extricated as customary articulation. They accomplished a precision of 90%.

Pooja Agarwal, Hanumandlu and Brijesh, in their paper Coarse Classification of Handwritten Hindi characters [5], depicted a framework for the arrangement of complete written by hand Hindi character set into subgroups dependent on some similitude measure. They proposed a calculation for finding and expulsion of header line and distinguishing proof of present position of vertical bar in written by hand Hindi character. Exploratory outcomes show that t beneficiary calculation is successful and accomplished an arrangement rate of 97.25%.

U. Pal, N. Sharma, in this paper we present a system towards the recognition of off-line handwritten characters of Devnagari, the most popular script in India. The features used for recognition purpose are mainly based on directional information obtained from the arc tangent of the gradient. To get the feature, at first, a 2×2 mean filtering is applied 4 times on the grayscale image and non-linear size normalization is done on the image. The normalized image is then segmented to 49 x 49 blocks and a Roberts filter is applied to obtain gradient image. Next, the arc tangent of the gradient (direction of gradient) is initially quantized into 32 directions and the strength of the gradient is accumulated with each of the quantized direction. Finally, the blocks and the directions are down sampled using Gaussian filter to get 392 dimensional feature vectors. A modified quadratic classifier is applied on these features for recognition. We used 36172 handwritten data for testing our system and obtained 94.24% accuracy using 5-fold cross-validation scheme.

Arora, S. Bhattacharjee, D. Nasipuri, in this paper a scheme for offline Handwritten Devnagari Character Recognition is proposed, which uses different feature extraction and recognition algorithms. The proposed system assumes no constraints in writing style, size or variations. First the character is pre-processed and features namely; Chain code histogram, four side views, shadow based are extracted and fed to Multilayer Perceptions as a preliminary recognition step. Finally the results of all MLP’s are combined using weighted majority scheme. The proposed system is tested on 1500 handwritten devnagari character database collected from different people. It is observed that the proposed system achieves 98.16% recognition rates as top 5 results and 89.58% as top 1 result.

Garg, Naresh Kumar Kaur, Lakhwinder, in this paper, author have discussed the new method for Line Segmentation of Handwritten Hindi text. The method is based on header line detection, base line detection and contour following technique. No pre-processing like skew correction, thinning or noise removal has been done on the data. The purpose of this paper is three fold. Firstly, we explained by experiments that this method is suitable for fluctuating lines or variable skew lines of text. Also, we confirm that this method is invariant of non uniform skew between words in a line (non uniform text line skew). Secondly, the contour following after header line detection
correctly separates some of the overlapped lines of text. Thirdly, this paper provides a brief review of text line segmentation techniques for handwritten text which can be very useful for the beginners who want to work on text line segmentation.

Sarvaramini, Farzin Nasrollahzadeh, Alireza, Convolutional Neural Networks (CNNs) have been confirmed as a powerful technique for classification of visual inputs like handwritten digits and faces recognition. Hindi handwritten character recognition (HHCR) is one of the challenging issues in machine vision. This study aims to investigate the performance of Convolutional neural networks (CNNs) on HHCR problems. To investigate the performance of different CNNs, a dataset of Hindi handwritten characters has been used as ground truth data. Different optimizers have been implemented on different parameters to determine the test accuracy of the proposed architecture.

Deepu Kumar, Divya Gupta, Off-line handwritten Devanagari script recognition is getting a brighter side of the research day by day. In India, millions of people use handwritten Devanagari script for documentation in northern and central parts of India. The optical character recognition for off-line Devanagari script has been improving day by day. Some innovative steps have been taken into consideration. A bunch of work has been also accounted on handwritten character recognition attempt for several Indian scripts, like Gurmukhi, Gujarati, Oriya, Telugu, Kannada, Tamil, Malayalam, etc. This Off-line handwritten Devanagari script recognition does not have enough reported works. As of late different techniques have been represented by the researchers in the direction of off-line handwritten Devanagari script recognition, many recognition systems for detached handwritten Devanagari characters present in the literature work. The objective of this review paper most desirable feature extraction techniques, as well as classification techniques used for the identification are reviewed in various segments of the paper. An effort is made to address the most crucial consequences reported so far and it is also tried to foreground the better directions of the research to date. This review paper is intended to serve as a guide for the readers, working in the field of off-line handwritten Devanagari character recognition.

Mahesh Jangid, handwritten character recognition is currently getting the attention of researchers because of possible applications in assisting technology for blind and visually impaired users, human–robot interaction, automatic data entry for business documents, etc. In this work, we propose a technique to recognize handwritten Devanagari characters using deep convolution neural networks (DCNN) which are one of the recent techniques adopted from the deep learning community. We experimented the ISIDCHAR database provided by (Information Sharing Index) ISI, Kolkata and V2DMDCCHAR database with six different architectures of DCNN to evaluate the performance and also investigate the use of six recently developed adaptive gradient methods. A layer-wise technique of DCNN has been employed that helped to achieve the highest recognition accuracy and also get a faster convergence rate. The results of layer-wise-trained DCNN are favourable in comparison with those achieved by a shallow technique of handcrafted features and standard DCNN.

III. RECOGNITION PROCESS

Character recognition is one all told the very important tasks in pattern recognition. The standard of the character recognition draw back depends on the listing to be recognized. Character recognition technique is dependent upon vary of things like varied font sizes, noise, broken lines or characters etc. and these factors influence the results of recognition system [11]. Artificial Neural Network is one all told the techniques wide used for character recognition draw back and thought of as a strong classifier on account of their high computation rate accomplished by massive parallelism [12, 14]. There unit four totally different phases in character recognition processes specifically Character acquisition, pre-processing stages, grouping of characters and Character Recognition.

<table>
<thead>
<tr>
<th>Character Acquisition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Processing</td>
</tr>
<tr>
<td>Grouping Characters</td>
</tr>
<tr>
<td>Characters Recognition</td>
</tr>
</tbody>
</table>

Figure 1: Stages of character recognition process
A. Character Acquisition:

Character acquisition is that the 1st innovate any image process or pattern recognition task. During this paper the images of Hindi characters, in tiff, jpg, bmp, and gif format square measure obtained through a scanner. once getting the digital image, ensuing step is to use pre-processing so as to boost the image clarity and conjointly the accuracy of recognition rates.

B. Pre-Processing:

Pre-processing is a very important step of applying variety of procedures for smoothing, enhancing, filtering etc, for creating a digital image usable by ulterior rule so as to boost their readability for Optical Character Recognition software system. The assorted stages concerned within the pre-processing are:

![Preprocessing Stages](image_url)

C. Grouping of Characters:

1. Binarization:

Image binarization converts a picture of up to 256 grey levels to a black and white image. Frequently, binarization is employed as a pre-processor before OCR. In fact, most OCR packages on the market work solely on bi-level (black & white) pictures. The simplest way to use image binarization is to settle on a threshold worth, and classify all pixels with values higher than this threshold as white, and every one alternative pixels as black. The matter then is the way to choose the right threshold. In several cases, finding one threshold compatible to the whole image is extremely tough, and in several cases even not possible. Therefore, accommodative image binarization is required wherever an optimum threshold is chosen for every image space.

2. Noise Elimination

Noise that exists in pictures is one amongst the most important obstacles in pattern recognition tasks. the standard of image degrades with noise. Noise will occur at completely different stages like image capturing, transmission and compression, varied normal algorithms, filters and morphological operations out there for removing noise that exists in pictures. Mathematician filter is one amongst the popular and effective noise removal techniques. Noise elimination is additionally known as as smoothing. It may be accustomed scale back fine rough-textured noise and to boost the standard of the image. The techniques like morphological operations accustomed connect unconnected pixels, to get rid of isolated pixels, and conjointly in smoothening pixels boundary.

3. Grouping of characters:

In the wake of pre-processing of character, alternatives of character square measure separated. This progression is heart of the framework. This progression helps in arranging the characters upheld their choices. The vowels and consonants of Hindi postin square measure partitioned into sub groups bolstered beyond any doubt imperative qualities. The vertical bar highlight and its situation inside the character is utilized to group the vowels and consonants in to sub groups. The characters square measure grouped in to three sub groups, the essential sub group comprises of character with none vertical bar. Characters with vertical bar at right aspect of the character square measure in second sub group and furthermore the third bunch incorporates the characters including a vertical bar inside the centre of the character.

D. Character Recognition:
Character recognition system is that the base for several differing types of applications in numerous fields, several of that we tend to use in our way of life, value effective and fewer time overwhelming businesses, post offices, banks, security systems, range plate recognition system and even the sphere of AI use this technique because the base of their operations. Recently, neural network became highly regarded as a way to perform character recognition it’s been reported that neural network are capable of providing sensible recognition rate even at the presence noise wherever alternative strategies ordinarily fail. The inherent pattern recognition skills of superimposed neural networks lend itself utterly to the current sort of task, by autonomously learning the advanced mappings in high dimensional input file. During this work, the matter of character recognition is resolved mistreatment neural networks. Neural network maps the set of input values to line of output values. The multi layer feed forward connectionist model trained by back propagation (gradient – descent) is employed to acknowledge the given input character.

IV. Result Compression of Previous work

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature</th>
<th>Classifier</th>
<th>Data Set Size</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Govindraj et al. [32]</td>
<td>Gradient</td>
<td>Neural Networks</td>
<td>4,506</td>
<td>84</td>
</tr>
<tr>
<td>Kompalli et al. [33]</td>
<td>GSC</td>
<td>Neural Network</td>
<td>32,413</td>
<td>84.77</td>
</tr>
<tr>
<td>Bansal et al. [14]</td>
<td>Statistical Structural</td>
<td>Statistical Knowledge Sources</td>
<td>Unspecified</td>
<td>87</td>
</tr>
<tr>
<td>Huanfen g Ma et al. [34]</td>
<td>Statistical Structural</td>
<td>Hausdorff image comparison</td>
<td>2,727</td>
<td>88.24</td>
</tr>
<tr>
<td>Natraj et al. [40]</td>
<td>Derivatives</td>
<td>HMM</td>
<td>21,727</td>
<td>88.24</td>
</tr>
<tr>
<td>Bansal et al. [8]</td>
<td>Filters</td>
<td>Five filters</td>
<td>Unspecified</td>
<td>93</td>
</tr>
<tr>
<td>Dhurandhar et al. [38]</td>
<td>Contours</td>
<td>Interpolation</td>
<td>546</td>
<td>93.03</td>
</tr>
<tr>
<td>Kompalli et al. [37]</td>
<td>GSC</td>
<td>K-nearest neighbor</td>
<td>10,297</td>
<td>95</td>
</tr>
<tr>
<td>Kompalli et al. [35]</td>
<td>SFSA</td>
<td>Stochastic finite state automaton</td>
<td>10,606</td>
<td>96</td>
</tr>
<tr>
<td>Dhingra et al. [39]</td>
<td>Gabor</td>
<td>MCE</td>
<td>30,000</td>
<td>98.5</td>
</tr>
<tr>
<td>Divakar et al.</td>
<td>Mean distance, pixel value and vertical zero crossing</td>
<td>Neural Network</td>
<td>1000</td>
<td>98.5</td>
</tr>
<tr>
<td>Govindraj et al. [32]</td>
<td>Gradient</td>
<td>Neural Networks</td>
<td>4,506</td>
<td>53</td>
</tr>
<tr>
<td>Kompalli et al. [37]</td>
<td>GSC</td>
<td>K-nearest neighbor</td>
<td>1,882</td>
<td>58.51</td>
</tr>
<tr>
<td>Kompalli et al. [33]</td>
<td>GSC</td>
<td>Neural Network</td>
<td>14,353</td>
<td>61.8</td>
</tr>
<tr>
<td>Ma et al. [34]</td>
<td>Statistical Structural</td>
<td>Hausdorff image comparison</td>
<td>2,727</td>
<td>66.78</td>
</tr>
<tr>
<td>Kompalli et al. [35]</td>
<td>SFSA</td>
<td>Stochastic finite state automaton</td>
<td>10,606</td>
<td>87</td>
</tr>
<tr>
<td>Divakar et al.</td>
<td>Mean distance, pixel value and vertical zero crossing</td>
<td>Neural Network</td>
<td>650</td>
<td>90</td>
</tr>
</tbody>
</table>

Table-1: Result comparison of OCR Systems
V. CONCLUSION

Gujarati Handwritten digit acknowledgment in computerized pictures is a testing issue because of contrasts in size, style, introduction, and arrangement, just as low picture differentiation and complex foundation. Discover a totally hearty and summed up method for digit acknowledgment; it is hard to give proper contribution to the optical character acknowledgment (OCR) framework. Numerous calculations have been proposed for perceiving content information in a picture. Every strategy gives hearty outcomes for determined arrangement of pictures. We utilized after technique for digit extraction: picture pre-handling, Segmentation and limitation, highlight Extraction, and grouping.

VI. REFERENCES


