Journal of Advances and Scholarly Researches in Allied Education
Vol. XV, Issue No. 1, April-2018, ISSN 2230-7540

A Study of Indian Script Handwritten Characters Using Neural Networks

Anita Venugopal[1]* Dr. Anil Kumar Kapil[2]

[1] Research Scholar, Motherhood University, Roorkee

[2] Professor, Faculty of Mathematics and Computer Sciences, Motherhood University, Roorkee

Abstract – In this paper hand-printed numeral identification, the numerals are supposed to be rendered legibly allowing for minor differences in the form of a numeral. Handwritten numeral identification is always the subject of research in the field of picture process and pattern recognition. Compared to the question of scanned numeral recognition, the issue of handwritten numeral recognition is exacerbated by differences in the shapes and sizes of handwritten English numerals. This makes the recognition of handwritten numerals or characters an intensive area of research in pattern recognition. Considering all these, the problem of handwritten numeral recognition is addressed under the present work in respect to handwritten English numerals. This dissertation presents a new approach for handwritten numeral recognition method using the MLP (Multi-Layer Perceptron) neural network for classification purpose.

Keywords- Hand Written Strings, Devanagari Characters, Artificial Neural Networks etc.

----------X----------

INTRODUCTION

Numeral variations in scale, form, slant and writing style render work more challenging. Numeral character recognition remains the most difficult field, as both the economic and theoretical challenges have not been resolved by the large research and development initiative that has gone into it. Manuscript character identification is an important step in many text processing applications. Virtual paper delivery is becoming increasingly common for applications of office and library automation, banking and postal services, publishing houses and communications technology[1]. The magnitude of the issue is greatly increased by data noise and the endless variation in handwriting owing to the temperament of the user and the essence of learning. The identification of handwritten digits has been a popular subject of research for many years. The identification of handwritten numbers has multiple applications, such as workplace automation, mail sorting, bank check recognition, etc. Because handwritten numbers and characters are used in highly sensitive fields such as banking and accounting, the recognition system should be reliable, quick and easy to introduce. Prior technologies classified under optical character recognition (OCR) could only recognise typed or handwritten numbers in set sizes and fonts. Yet, as stated in the literature, the 100 percent identification figure is beyond the scope of these programs. The present study therefore aims to create an effective program targeting 100% identification in the face of varying sizes, shapes and fonts[2].

FEATURES OF INDIAN LANGUAGES AND SCRIPTS

We focus on certain attributes and scripts of Indian languages that may influence how OCRs are constructed for these languages.

- **Common Alphabet**

Yes, every language in Indian has the same alphabet as the Sanskrit script. The common script comprises 33 consonants and 15 vowels in frequent observe. For different languages or other traditional forms, extra 3-4 consonants and 2-3 vows are used. In addition, this is not an essential difference. The alphabet's main letters are common consonants and vowels. Tamil is the southern language with around 12 fewer consonants, the most prominent exception to this regular script. Nevertheless, the form in Tamil is not so special as some consonants can be based on the change in the master list to be dropped. The Indian Standard Interchange Code (ISCII) now standard explicitly accepts the commonality of the alphabet. Of Indian languages, Unicode has followed the same theme[4].

- **Basic Character Unit**

In addition, Indian Akshara languages have an extra complicated definition of an English-like word unit or akshara to form the essential linguistic component. The 0, 1 or 2 or 3 consonants and the vowel shall consist of akshara. Names are one or two aksharas.

Because the languages are fully phonetic, each akshara may be pronounced independently. Aksharas are labeled samyuktaksharas or combo-characters of more than one consonant. The last consonant in the samyuktakshara is the main consonant. This akshara structure is recognized and encoded by the ISCII standard. Thus, ISCII's essential unit of depiction will vary in length from one to six bytes. A series of text bytes stored in ISCII can be broken into its aksharas.

- **Different Graphemes**

The alphabet's commonality does not include the graphic types used for printing them. Each language uses different scripts for printing, consisting of different graphics. Therefore, printed content is unavailable to readers of one language in other languages. India has 10-12 big plays. The most widely used Devanagari script for Nepal, Hindi (the mainly verbal), Marathi, Konkani and Nepali is the speech of Nepal's neighbor. Various scripts combine different philosophies of unique visuals and their variations. Several have a shirorekha or headline that persists for a word. Others have graphics that are not touching[5]. The middle of the written akshara is usually one of the consonants ' maps. The vowel serves as a matra or vowel trigger. It can be seen in or below configurations at top, center or above. Supporting consonants of a samyuktashara often refer to the l as modifier graphemes in some cases or may be distinct. Even within a script, these laws are not consistent and definitely not across texts. In different scripts and fonts, a language handwriting and print reliant feature are transferred to a tongue driver to print ISCII string. Similar scripts are using dissimilar drivers. A obstruct of manuscript from solitary language to other must be transmitted by another driver.

- **Unstructured Font Design**

Over the past few years, various fonts has been developed to every Indian script. The fonts are then constructed from glyphs and follow the graphical structure of each script, which is different for different languages[6]. For all plays, the normal set of rules for this point cannot be used. Unfortunately, no guidelines have been observed to establish the glyphs for various print families of a similar language, which is quite possibly.

The computer fonts were specified for specific purposes to increase the ambiguity (for example, For its online website per electronic newspaper identifies its own font). For ISCII converters, you don't exceed the glyph set. Due to the very limited use of local machines, the ISCII guidelines were not widely adopted by the different players, such as each newspaper.

CHALLENGES WITH INDIAN SCRIPTS

OCR is the automated translation of typed document pictures (normally obtained from a scanner) to apparatus-editable transcript. OCR has been researching for concerning 60 years, except since only the past decade, Indian script OCR has been studied seriously. English OCR systems are willingly existing, however OCR's for specific Indian languages has not yet come to filled maturity[7]. There has been a number of attempts to create OCRs for Indian writes similar to Devanagari, Malayalam, Telugu, Tamil Bangla, Gurumukhiand Kannada. Several of that efforts have been made. These OCRs do not cover several problems related to document image analysis, which are script-independent. In this article we portray an automated OCR that uses ordinary features and specific scripting approaches to create a robust structure. The result of a collaborative project sponsored by the GOVT Department of Information Technology at several institution for the creation of Indian print-script OCR structures. Pakistan. Global. Pakistan. Global.

The pre-processing methods were freely selected by each center, such as symbol division, piece collection, engine recognition selection, and the synthesis of results in words and phrases. Conversely, continuity in the method as a whole an incorporated architecture was accepted.

The main dispute associated with the implementation of incorporated OCR proposal as:

- ► Determine the design designs for entire practical element to enable the creation, test and execution of the modules independently, thereby the interconnection between the modules.
- ► Maintain the code standards established during the initial project phases (C++, compliance with Fedora core 6, Doxygen, etc.).
- ► Rendering and processing of records online in various texts.
- ► Development of success strategies, testing and evaluation.

It was a two-stage system for Devanagari and Bangla OCR. We also graded the character of the study into one of the sections by separating like characters into categories in the first level. By the subsequent phase, we were confidential as characters of the cluster. The linking character box was split into 5x5 windows and the total value was calculated for four directions (updown, left-right, 45 ° left and 45 ° slant) when the limit was reached. So we had 5x 5x 4= 100 dimensional characteristics in both points. At first, the emphasis was on minimum distance and classification of K-NN. The results were subsequently noted as more consistent and reliable by the SVM classifier. The icons of the middle zone were a two-phases ring. Due to 1st chapter, numerous characters with strong contour similarities were grouped in one category and, according to the single sign SVM, are highly

confusing. This has led to the development of several classes. The prototypes for such groups were educated by SVM classifiers. Then for the individual group symbols for the second phase, another set of SVM classificators was created. Because in the top area and the lower part of the text line there are few symbols, a single classifier method for each zone has been created.

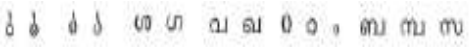
To integrate the effects of acceptance, language grammar information is essential. The principles used to merge ortho-syllables (Akshara) must have been defined using symbols and partially the symbols in the top, middle and bottom parts. It is always a challenge to keep exhaustive, and erroneous and dubious outputs from three areas frequently cause obscurity in term creation.

Nevertheless, at least the Banglady OCR systems were not too poor though we need some improvements for Devanagari, which in the second stage of the Tamil Program is one of the oldest and most recently recognised by the Government of India as a classical language. Throughout Tamil there are eighteen consonants and twelve vowels[8]. There are however 5 additional diagrams used to identify the Sanskrit consonants borrowed. There's a different character/Aytam/. About 200 grades are included in the latest Tamil OCR version, including all signs, special symbols and Indo-Arabic digits. It is often used to convert printed tamil books, based on the Tamil OCR's high performance, into Braille by Worth Trust, a non-governmental organization in Chennai. In the last one year, about 150 nursery, colleges and other books have been adapted to braille, which is used by about 100 individuals with a visual incapacity, and they include some 25,000 words. It has been observed that OCR operates without any noticeable degradation in output with an arbitrary script. In the analysis of around 1,500 characters, OCR achieved an average performance of approximately 93 percent, thus addressing the major challenges of married and divorced character division. This production is base on a raw understanding lacking incorporate a vocabulary form or a curse improvement dictionary. Further successful accuracy development can therefore be achieved in recognition. The writing system in Malayalam is primarily syllabic. A vowel termination syllable through a canonical (C)V construction is the dominant orthographic unit. The compulsory V is a small or extended vowel.

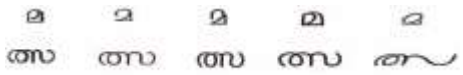
The usable C displays one or several consonants and, in certain case, the shape complies with the laws of phonology and primarily with articulation. In Malayalam, in addition to many conjugated and complex tones, there are 56 words, 15 vowels and 36 consonants[9]. There are two containing variations of the conjugated characters, but they are composed individually.

Text Change: Almost all of the traditional lipical characters got away with both the advent of modern word processors that can create some type of difficulty. Standards can be found across text processors and fonts. Most word processors are using a combination of old and new lipi characters[25].

Specific Characters: Such group of characters appear the same. The gap between these characters is so minimal that generally even people read the text[20] from their context.

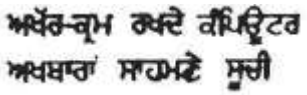


Glyph Variation: There are a number of characters that are identical. The gap is so tiny, that normally only people read the text by its meaning[9].



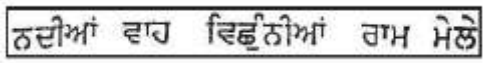
Inside a SVM classifier is used by Malayalam character recognition engine. In Directed Acyclic Graph (DAG) design, SVM classifiers are organized in parallel. All of the pairwise classifiers are actually linear. The same function definition is used by all pairwise classifiers. Simple statistical characteristics such as PCA and arbitrary ridge yield adequate outcomes. However, HOG (Histogram Of Gradient) features are being examined for broader support (for multi-font system). The initial results are good. The analysis of the related elements relies on the segmentation of character. These characters / symbols are recognized and a loop-up table is used to transfer the class marks to UNICODE. At this stage, further linguistic clues are used. Initial studies utilizing language subsymbols suggest that post-processing will significantly improve the recognition results. In the beginning, 32, 3 vaudron-bearers, 10 vowel modifiers (with no mukt symbol) and 3 auxiliary signs were included in the Gurmukhi syllabary. This script was eventually replaced with six further consonants. Six consonants of this kind are multi-component characteristics that can be divided into separate sections. Some characters actually change the consonants just below them when added. Half characters or roles subordinated to them are named protagonists.

Touching Characters [11]



This is Gurmukhi's printed script's most famous deterioration. Two adjacent characters contact each other in this type of degraded text. Another popular problem with old Gurmukhi text experiences is the survival of several skews on the identical paragraph,

which are critical in distinguishing the touching characters, is the proper segmenting of the characters, i.e. the situation to that moving couple ought to segmented several skews in the documents. That term or line could be distorted differently, so skew identification and correction could entail the creation of global and local algorithms.



The language is distinct from Devanagari-like scripts like Bangla and Gujarati, and also from the conventional Dravidian handwriting [10] of Tamil. The present Canada script, called varnamale, contains 48 characters. Consonants are categorized and consonants are ungrouped as consonants. In 14 vowels there are 34 consonants and 10 digit. Specific character are vowels and consonants. Vowel alter may emerge on the right top of or at the bottom of a bottom consonant. As shown in table 1, however, the two-dimensional arrangement of consonant clusters in Kannada[10].

Table 2.1: Some examples of CCV and CCCV

CCV Combinations	
CCCV Combinations	

The Kannada OCR[10] switch these plain, complicated symbols, the kannada numerals and the most common QWERTY symbols. There are therefore approximately 300 classes. In order to extract functionality of the structured frames of the segmented part, Karhunen Leuve Transformation is used. The SVM form is here. This is the SVM form. The functions, that comprise of several different connected components, are grouped into one category by a substantial combination of examination of connected components and vertical projectionThe segmentation precision of the item is very strong even when the characters are fused or divided. Therefore, it's true to achieve high quality primitives segmentation to develop a better Kannada (such other manusript thereby). Telugu is a phonetic tongue to sounds (often syllables), which are loosely spoken. In Telugu the script includes 16 vowels (12), 36 consonants (35) and 2 vowel transform with rounded features in complex forms with no vertical lines.

Complex configurations of extremely local configuration (e.g.)

Relationships division / Superset (e.g.)

Sets which are visually alike and puzzling (e.g.)

Modifiers of vowels add their forms to consonants and consonants. A significant number of similar form type characters (basic and conjunct) are found in Indian languages. From that point of view, we have suggested novel features based on the topography of the character to improve the performance of the current OCR in Indian script, particularly in Bengali and Hindi documents. The main features of the proposed method are as follows:

1. The key difficulty in developing an OCR for Indian script is to accommodate large-scale differences in the form of various scripts. Structures of characters can be decomposed into fragments that are straight lines, convex lines or closed boundaries (holes). In our research, we call the topography of the character strokes in four directions. In addition to the different convex forms created by the strokes of the characters, we also notice the appearance of the closed borders of the area.

2. The extracted features are represented by a shape-based graph, where each node comprises a topographic element, and all of them are positioned in the original character picture with respect to their centroids and relative positions.

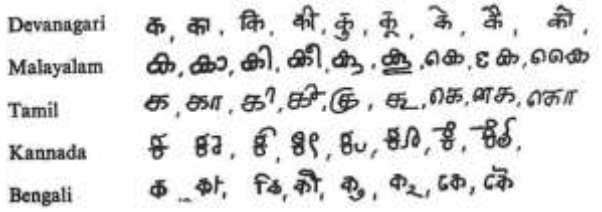
3. The approach suggested is specific to different languages.

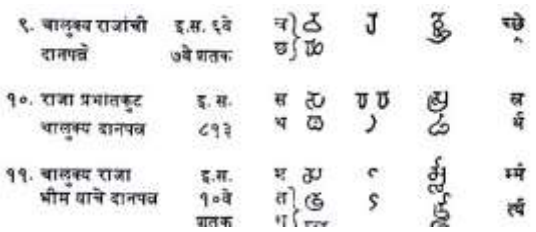
4. This technique extends to both typed and handwritten text documents.

5. This collection of topographic features allows to discern very similar type characters (shape similarity) in the right way.

BASIC ARRANGEMENT OF PHONEMES IS PHONETIC

The Brahmi script has developed from the phonemic experiments in the Dravidians, which are entirely phonetic. Many foreign language pronunciations are labelled with a dot at the bottom of the closest phoneme. The nearest phoneme displays half open vowels borrowed from European languages with a half-moon symbol.





Learning a different script may in any case seem to be forbidden, but the handwriting so that Hindi is write quite simple to study, enjoyable to sketch and very delightful. Hindi is composed primarily in a language known as Nagari or Devanagari. It is typically talked about utilizing a mixture of about 52 sounds, 10 vowels, 40 consonants, nasalism and sort of desire[12].

The script for Devanagari has 13 characters that are generally known as vowels and 40. Four standard conjunct consonants must also be known, a concept I would further clarify. We will consider how characters and sounds are interpreted, and how terms and their resonance are created. The major action you should execute is fundamentally an imaginative activity. You will learn how to connect sounds and pictures to figure out what you are doing as a symbol to help your friend with term and its tone. Such as apparently the character for क ka appears to some degree akin to a key, that is useful considering the way that I able to set the resonance and condition of the term together. While I can relate the character and the sound for the spinning condition of the character — that is, the cha beat. When you study Devanagari content and hindi sounds, you should remember two things. In the first case, but Hindi is printed in only a preordinate numeral of words, normally 33 or 52, and it is solidified in many respects before long, despite that an enormous portion of the major mixes is hardly adaptable once the central characters are identified. Subsequent, don't stress in case you can't well-spoken entire sounds effectively in any case (or even hear the differentiations every so often) as you get acquainted with your ear will 'tune in' to the noise and you will bit by bit make sense of how to verbalize them accurately.

THE HINDI SOUND SYSTEM

The sound framework (phonology) of present day conveyed in Hindi can be addressed in different contents with Devanagari. In English Devanagari is frequently referred to as syllabary rather than a letter in order, given that a consonant and a mix of vowels or vowel in loneliness is addressed regularly in every Devanagari character. Devanagari consonants are regularly regarded as having a key structure, containing a consonant which is stated in English words and young people with a natural' a'-sound like a vowel.

Each Devanagari character speaks to one absolute syllable regularly[14]. It is fairly effortless to study Devanagari since mostly phonetic, meaning that the language is usually represented by real sounds. You will learn two forms in that hindi understands resonance that was not common to English presenter until you try to work out how to express Hindi. There are most Hindi consonants, 2 by 2 where one consonant form is unpirated and the other is absorbed. Neither or very little, the unpirated consonants are voiced with the tone and the consonants are pulled with a very strong wind. Almost all consonants are absorbed in English. The Hindi dental consonants ' t' and 'd' are also known, with retroflex consonants ' t' and' d.' The dental consonants has created through tetchy the upper part of your mouth by pressing down and removing the edge of the language. T ' and d'sounds between indish consonants and retroflexes are again marked. When you train, the nuances continue to be heard and how the various sounds can be expressed.

THE TRADITIONAL ORDER OF DEVANAGARI

You can get acquainted to the customary Devanagari sort. It's because speakers of mother language want to say you that in this manner and dictionaries are in t order. You see the tabular since but it was page, top-down, right-hand side[15].

1	अ आ इ ई उ ऊ ऋ
2	ए ऐ ओ औ अं अः
3	क ख ग घ ङ
4	च छ ज झ ञ
5	ट ठ ड ढ ण
6	त थ द ध न
7	प फ ब भ म
8	य र ल व
9	श ष स
10	ह
11	क़ ख़ ग़ ज़ फ़ ड़

Remember the characters changed by points are not mentioned separately in the standard Devanagari order but treated as variants of their base characters.

CONCLUSION

The identification of handwritten numbers has many uses, such as office automation, postal processing, bank check recognition, automated pin code recognition, data collection from completed documents, etc. Although some commercially available software is available, primarily for the identification of typed characters in some languages, but the popularity is still to be applied to hand-written characters. This methodology is designed to facilitate better contact between man and machine. Because handwritten numbers and characters are used in highly sensitive areas such as finance and administration, the recognition program should be accurate, simple and easy to implement. Our future work is dedicated to improving the identification rate of indistinct unregulated numerals utilizing advanced extraction techniques.

REFERENCES

[1] O. D. Trier, Anil K. Jain, Torfinn Taxtt (1996). "Feature extraction methods for character recognition- A survey", Pattern Recognition, Vol. 29, No.4, pp. 641-662.

[2] Basu, S. et. al. (2005) "Handwritten 'Bangla' alphabet recognition using an MLP based classifier," Proceeding of 2nd National Conference on Computer Processing of Bangla, pp. 285-291.

[3] Pal, A. and Singh, D. (2010) "Handwritten English Character Recognition Using Neural Network," International Journal of Computer Science and Communication, Vol.1, No.2, pp. 141-144.

[4] Dinesh, A. U. et. al. (2007). "Isolated handwritten Kannada numeral recognition using structural feature and K-means cluster," IISN, pp. 125-129.

[5] Perwej, Y. and Chaturvedi, A. (2011). "Neural Networks for Handwritten English Alphabet Recognition," International Journal of Computer Applications, Vol. 20, No. 7, pp. 1-5.

[6] "Document image processing of Indian scripts," Special Issue of Sadhana, 2002.

[7] C. V. Jawahar, M. N. S. S. K. Pavan Kumar and S. S. Ravi Kiran (2003). "A bilingual OCR for hindi-telugu documents and its applications," in International Conference on Document Analysis and Recognition.

[8] C. V. Jawahar, M. N. S. S. K. Pavan Kumar, and S. S. Ravi Kiran (2002). "Recognition of Indian Language Characters using Support Vectors Machines," Technical Report TR-CVIT-22, International Institute of Information Technology, Hyderabad.

[9] V. Bansal (1999). "Integrating knowledge sources in Devanagari text recognition," doctoral thesis, IIT Kanpur, Department of Computer Science and Engineering.

[10] V. Bansal and R.M.K.Sinha, "A Devanagari OCR and a brief overview of OCR research for Indian scripts," http://www.cedar.buffalo.edu/ilt/research.htm . Home Page Denanagari OCR.

[11] B. B. Chaudhuri and U. Pal (1997). "An OCR system to read two Indian language scripts: Bangla and devnagari (hindi)," in Proc of ICDAR, pp. 1011–1015.

[12] A. Negi, Chakravarthy Bhagvathi, and B. Krishna (2001). "An OCR system for telugu," in Int. Conf. Document Analysis and Recognition (ICDAR).

[13] T. V. Ashwin and P. S. Sastry (2002). "A font and sizeindependent OCR system for printed kannada documents using support vector machines," Sadhana, vol. 27, pp. 35–58, February 2002.

Corresponding Author

Anita Venugopal*

Research Scholar, Motherhood University, Roorkee

aradhana.parmar14@gmail.com