Journal of Advances and Scholarly Researches in Allied Education
Vol. 13, Issue No. 1, April-2017, ISSN 2230-7540

An Analysis of Hand Written Strings of Devanagari Characters Using Artificial 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 – The identification of handwritten numbers or characters has applications in the area of pattern recognition, text preparation and interpretation. However, the identification of handwritten numbers or characters is a difficult activity for the machine, as the writing styles differ from individual to person. The identification of handwritten numbers or characters is typically a standard concern for Pattern Recognition and Artificial Intelligence. The suggested scheme shows that it is used for the identification of handwritten English numerals identified by the said feature set. This approach to the identification of handwritten English numerals by the MPL classifier can also be expanded to include handwritten English alphabet characters. The images are processed by human operators in many of these real world applications. Automation, however, will increase efficiency and reduce costs. To do this, an automated system's efficiency has to compare favorably with human performance.

Keywords – Hand Written Strings, Optical Character Recognition, Devanagari Characters, Artificial Neural Networks etc.

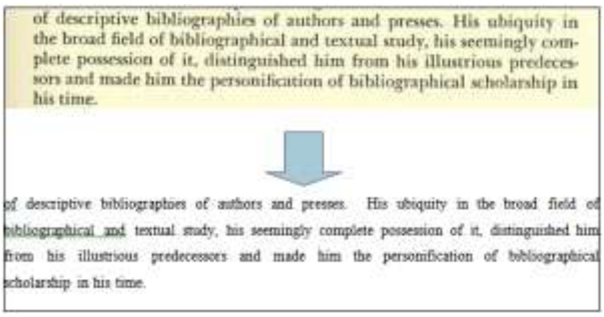
- X -

1. INTRODUCTION

The recognition of handwritten numeral character has been considerable interest to researchers working on OCR. Offline numeral identification has a range of practical applications. The interpretation of postal zip codes in addresses written or typed on envelopes is one of the most cited practical applications for offline numeral identification. The advantages of introducing these schemes in the postal services are immense. These systems will enable the automated sorting and routing of millions of mails passing through the postal system every day, reduce the human workload and speed up the entire operation. Nevertheless, despite extensive testing, current techniques do not produce results that meet the required precision. A robust and efficient numeral recognition device requires to be really accurate[1].

OPTICAL CHARACTER RECOGNITION

Optical Character Recognition (OCR) is the method of classifying optical objects, typically found in a digital image referring to an alphanumeric or other character. It involves many phases, including segmentation, isolation of features and grouping. That of these is a complex field, which is discussed briefly in this essay. Consider the OCR, which is part of the scanned text document, borrowed from the internet, shown below along with the corresponding (humanly recognized) characters from that language.


of descriptive bibliographies of authors and presses. His ubiquity in the broad field of bibliographical and textual study, his seemingly complete possession of it, distinguished him from his illustrious predecessors and made him the personification of bibliographical scholarship in his time.

of descriptive bibliographies of authors and presses. His ubiquity in the broad field of bibliographical and textual study, his seemingly complete possession of it, distinguished him from his illustrious predecessors and made him the personification of bibliographical scholarship in his time.

Figure 1: Scanned image of text and its corresponding recognized representation

Optical character recognition is sufficient if the text can be interpreted both to humans and to the computer; alternate inputs cannot be specified beforehand. Compared to other automated identification techniques, OCR is unique in that the mechanism that generates the information is not needed to be managed[2].

The problem of identifying optically rendered characters is concerned with optical character recognition. After the text has been written or printed, optical recognition is performed off-line, Unlike online recognition where the character is recognized as drawn by the machine. OCR's output depends directly on the consistency of the typed or handwritten input documents.

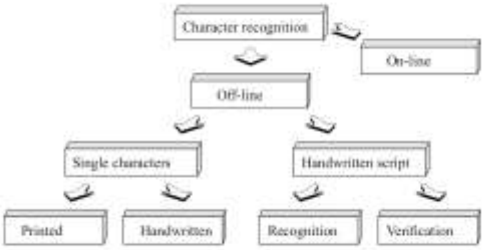


Figure 2: Different phases of character recognition

The less data, the higher the output of the OCR system. Nevertheless, OCR computers are still far from reading as well as humans when it comes to fully restricting free handwriting[4]. The machine learns easily, however, and technical advances bring the device closer to its day-to-day goal.

2. HISTORICAL DEVELOPMENT OF OCR

Systematically, the identification of a character is a separation of the prototype recognition. Nonetheless, the identification of words which provides the motivation to make mature fields of science for pattern recognition and image analysis.

2.1 The First thought

Since ancient times it has been a dream to imitate the person purpose by apparatus, creation the appliance able to execute errands similar to reading. In the year 1870 C. The retina scanner was invented by R. Carey of the Boston Massachusetts as a photocell mosaic system for transferring pictures. 20 years later, a major breakthrough came in the form of an invention by the Polish P. Nipkow who invented the serial scanner for both modern television and reading machines. Via experiments with OCR to support the blind, numerous attempts were made during the first decades of the 19th century to create devices. The modern version of OCR was created in the mid-1940s by the digital computer. Since then the focus for development was the possible applications of OCR in the business world [1].

2.2 The Beginning

By the year 1950 the world was witnessing technological revolution. Electronic data processing became a significant region. The admission of data was performed via punched cards and the through amount of data had to be treated cost-effectively. The screen reading technology was also established at the same period, and OCR machines hit the commercial market by the mid-1950s.

In 1954, Reader's Digest mounted the first digital OCR reading machine. Using this machine, the typewritten sales documents were converted onto punched cards for program input. Commercial OCR systems may be graded for the next four years on the level of versatility, robustness & efficiency[2-3].

2.3 First generation OCR

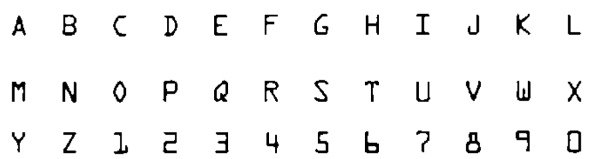
The OCR systems which emerged between 1960 and 1965 can be called OCR's first generation[1]. The restricted letter shapes read primarily characterized this generation of OCR machines. The symbols were originally designed exclusively for reading the computer and did not even look very realistic. Multifont computers were built later, which was interpret up to 10 different fonts. The numeral of fonts was constrained by pattern recognition approach used and the equivalent example, where each fonts character contrasts the glyph character with an interface sample collection.

2.4 Second generation OCR

The second generation analysis machinery came into being in the mid-1960s and early 1970s. Such devices were capable to identify both traditional machine-printed characters as well as hand-printed characters. Nevertheless, the numbers and a few letters and symbols were restricted to the character set when the hand-printed characters were listed.

IBM 1287 was the first and well-known machine of this kind to be demonstrated at the New York World Fair in 1965. Toshiba also produced the first automated letter taxonomy system for postal code digit during this time, and Hitachi developed a high-performance, low-cost OCR machine.

Significant work has been done in the field of standardization of OCRs during this time. A detailed analysis of OCR specifications was carried out in 1966, Define the U.S. OCR standard set called the OCR-A character set. This was a very fashionable font built to ease, understandable visual recognition. A European font named OCR-B was also produced with more realistic fonts than the US version. There have been few efforts to combine the two fonts into one uniform, but computers have been shown to interpret them.



A B C D E F G H I J K L

M N O P Q R S T U V W X

Y Z 1 2 3 4 5 6 7 8 9 0

Figure 3: OCR-A (top) and OCR-B (bottom)

2.5 Third generation OCR

During mid of 1970s, the third generation of OCR devices was faced with the problem of low quality records and wide printed and manuscript sets. The exponential advances in hardware technology were a major obstacle for low cost and high performance.

Although the industry started to manufacture increasingly sophisticated OCR equipment, simpler OCR systems were still being used. Until personal computers and laser printers in document processing became regulated, typing was a different area to the OCR. Since print isolation was common and a few fonts were very effective basic OCR devices. For the final editing, ordinary typewriters were used to produce rough drafts that were fed into the machine via an OCR tool. This way word processors, which at this time were a costly tool, could serve more people and could reduce the cost of the computer.

2.6 OCR in today (Fourth Generation)

Although OCR machines were already commercially available in the 1950s, up to 1986 only a few thousand systems had been sold worldwide. The main reason for this was the system cost. When hardware became cheaper, however, So revenues began to increase tremendously for OCR systems as product products were offered. Currently, several thousand devices are marketed each week, and the prices of an OCR OMNIFAT have decreased by ten every other year during the last six years.

This is the fourth era and could distinguished with the OCR of composite records intermixing towith content, illustrations, tables and numerical images, unrestrained transcribed characters, shading archives, low-quality loud reports, and so forth. Among the commercial items, postal location perusers, and perusing helps for the visually impaired are accessible in the market. These days, there is a lot of inspiration to give computerized record investigation frameworks. OCR adds to this advancement by giving strategies to change over enormous volumes of information consequently. Countless regulations and certificates report identification levels of up to 99,99%; which gives the impression of computerization issues being overcome. Despite the fact that OCR is generally utilized directly, its exactness today is still a long way from that of a seven-year old youngster, not to mention a tolerably gifted typist. [4]

The profound distortion of some real implementations demonstrates that difficulties with exhibits continue to exist in synthetic and deteriorated archival content (i.e. boisterous features, tilting, document blending, etc.). Various strategies to enlarge the precision of OCR have been suggested. Actually, different investigate labs, the test to create hearty techniques which expel however much could reasonably be expected the typographical and commotion confinements while keeping up rates like those gave by restricted text style commercial machines [5].

Along these lines, ebb and flow dynamic research regions in OCR incorporate handwriting recognition.

3. RECOGNITION OF HANDWRITTEN DEVNAGARI CHARACTERS USING FEED FORWARD NEURAL NETWORKS

For the written identification of the input-output relationship the 3-level neural network is important. The data neuron layers, i.e. Functional vectors F1, are independently liable for input. The neuron number in the output layer is calculated according to the required power set and each output is represented by a different neuron. Our research consists of 16 inputs, 20 first hidden layer neurons, 14 second hidden layer neurons, and 10 output layer neurons. This leads in the neural network for 16-14-10 back diffusion. As control tool, Sigmoid is used. The learning factor for back-propagation is 0.2 and the functional constant used is 0.9. The performance is 0.00126123 in 16400 times during the training process.

The current work on the detection of handwritten Hindi characters requires the creation of a neural network that can practice in recognizing the handwritten Hindi characters[9].

The steps in the system design are as follows:

do line and further character segmentation

- Extracting the features from local approach technique
- intend the neural network based upon the requisite and accessibility
- reproduce the software for network
- Train the network utilizing input statistics files until fault falls beneath the acceptance stage
- authenticate the potential of neural network in detection of test data

1	अ	91.3	9240
2	ब	82.7	7800
3	स	81.2	8980
4	द	[illegible]	7880
5	क	71.0	8210
6	ख	[illegible]	[illegible]
7	ग	82.4	8820
8	घ	80.1	18200
9	ङ	77.0	17140
10	च	87.1	7480
11	ढ	71.2	18250
12	ज	[illegible]	8810
13	झ	72.0	19890
14	उ	[illegible]	17000
15	ऊ	[illegible]	[illegible]
16	म	78.1	17710
17	ल	74.2	7980
18	ल	[illegible]	8770
S. No.	Handwritten Hindi Character	Recognition success rate	Epochs

ANATOMY OF DEVNAGARI CHARACTERS

The letter anatomy can be described as a scheme that depicts a letter's structural form, representing essential characteristics of a letter on the typeface. S was the first effort to describe Devanagari's script graphically. V.[4] Bhagwat. Based on graphical similarity as shown in Figure , he grouped messages.

Figure 4: The grouping of letters by Bhagwat based on graphic similarities[4]

For certain graphic elements in letters shown in Figure 3, it has also established guidelines for letters and termology.

Figure 5: Bhagwat's guidelines [4]

The topmost section is the Rafar Line, which passes along the Matra Line. Matra Line is the highest point of the upper matras. Head Line is shown after the Matra Column. Head Line is the highest point of Shiro Rekha. Head Line is preceded by Upper Mean Line and Lower Mean Line. Upper Mean Line Shows where the actual letter starts for instance the upper piece of the counter of 'ब' or 'व'. Lower Mean Line signifies where the characteristic element of the letter reaches a conclusion for instance the lower some portion of the counter of 'ब' or 'व'. This is trailed by the Base Line which means the finish of the character and where the lower matra starts. The lowermost line is the Rukar Line which means the finish of the lowest portion of the Rukar.

Bapurao Naik likewise endeavored the graphical gathering of letters. Naik sorted out letters graphically in five gatherings based on the situation of the kana or the vertical bar. The significant part of this gathering is that ए and ऐ are absent from the gathering. Naik's gathering of letters is appeared in figure 4.

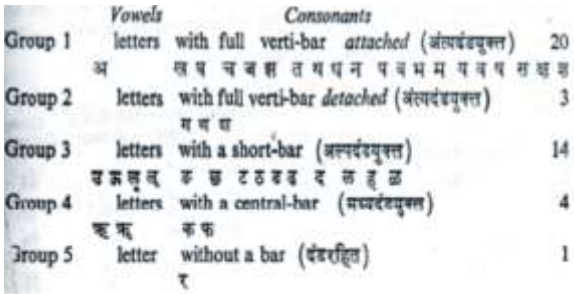


Figure 6 : Naik's grouping of letters on the basis of the position of the vertical bar or the kana [4]

Few men, such as M. W. Gokhale, Mahendra Patel also suggested various methods for grouping letters and creating a Devanagari script vocabulary. A detailed Devanagari script anatomy analysis can be contained in [5].

CONCLUSION

In this thesis a novel approach for handwritten numeral or characters recognition using neural network classifier and bounding box approach which was implemented successfully. The various writing styles of characters introduce large variations. In MLP Neuarl Network classifier the “thicken” operation is performed on the captured images from our database and to extract normalized features using the proposed bounding box approach. The major advantage of MLP classifier approach is robustness, easy to implement and relatively high recognition rate. Moreover, this method is insensitive to barbs and gaps that arise at the preprocessing stage. We can emphatically advocate the use of MLP Neural Network for document processing from this study. The automatic recognition of digits on scanned images has wide commercial importance. This approach of recognition of handwritten English numerals by MPL classifier can also be extended to include handwritten characters of English alphabet.

REFERENCES

1. J. Mantas (1986). “An Overview of Character Recognition Methodologies”,

Pattern Recognition, Vol. 19, No 6, pp. 425-430.

2. Eikvil Line (1993). "OCR-Optical character recognition system", Report No. 876, Document Image Analysis Publications, Norwegian computing center.

3. http://www.nr.no/documents/samba/ research_areas/BAMG/Publications/OCR.ps. gz.

4. U. Pal, B.B. Chaudhuri (2004). "Indian script character recognition: a survey" Pattern Recognition 37, pp. 1887 – 1899.

5. George Nagy, Thomas A. Nartker, Stephen V. Rice (2000). "Optical Character Recognition: An illustrated guide to the frontier", Proceedings of Document Recognition and Retrieval VII, SPIE Vol. 3967, Sanjose, pp. 58-69, January 2000.

6. Abdel Belaid (1997). "OCR Print - An overview" chapter 2 in "Survey of the state of the art in human language technology", pp. 71–74. http://www.loria.fr/~abelaid/publi_ps/ocr_art. ps

7. Sue Wu, Adnan Amin (2003). "Automatic Thresholding of Gray-level Using Multi-stage Approach" Proceedings of the Seventh International Conference on Document Analysis and Recognition.

8. N. Otsu (1979). "A threshold selection method from gray-level histograms". IEEE Transactions. Systems., Man., Cyber. vol. 9: pp. 62–66.

9. Graham Leedham, Saket Varma, Anish Patankar and Venu Govindaraju (2002). "Separating text and background in degraded document Images – a comparison of global thresholding techniques For multi-stage thresholding" Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition.

10. W. Niblack (1986). "An Introduction to Digital Image Processing", pp. 115-116, Prentice Hall.

11. J. Sauvola, M. Pietika Kinen (2000). "Adaptive document image Binarization" Pattern Recognition Vo.33, pp. 225-236.

12. Rangachar Kasturi, Lawrence O'Gorman and Venu Govindaraju (2002). "Document image analysis: A primer", Saadhanaa Vol. 27, Part 1, pp. 3–22.

Corresponding Author

Anita Venugopal*

Research Scholar, Motherhood University, Roorkee

aradhana.parmar14@gmail.com