JISEAT Journals for International Shodh in Engineering and Technology

Journals for International Shodh in Engineering and Technology
Website: http://jiseat.com (Volume 02, Issue 3, July 2017)

A Theoretical Review on Optical Character Recognition

Sandli Bansal
M. Tech. Scholar
Shrinathji Institute of Technology and Engineering,
Nathdwara (India)
sandlibansal@gmail.com

Ms. Komal Paliwal
Assistant Professor
Shrinathji Institute of Technology and Engineering,
Nathdwara (India)
erkomalpaliwal.cs@gmail.com

Abstract – **Optical character recognition (OCR) is a technique of digitizing texts from images of symbols or characters that belong to a certain alphabet. Thanks to this, the recognized data can be identified and stored from the images and interact with these characters. Despite the efforts and intensive work done in the field of optical recognition of writing, no OCR system is considered 100% reliable. The objective of this paper is to present a theoretical review on optical character recognition.**

Keywords –**ANN, DTP, DWT, OCR, HCR, SVM.**

I. Introduction

Historically, man has been fascinated by the machines. The first controllers having fun and were performing repetitive tasks. Nowadays they are endowed with sensory organs allowing them to see, hear and even taste. Among the most sophisticated robots, there are those which are able to decipher the human writing. This field of writing still remains to be explored given its complexity and diversity. There are two areas of research: recognition of printed characters and handwriting recognition.

In the case of handwriting recognition, there are two fields of applications following the entry mode. If the input mode is dynamic, we talk about recognition in real time called online or on-line. In the case of the static entry is called deferred-time recognition called offline or offline [1]. The offline recognition applies once the writing is on paper that is scanned and saved to an image format. This image contains pixels either binary (black and white pixels) or integer (grayscale picture). This offline recognition includes two themes: the recognition of isolated handwritten characters and word recognition. Current applications are aimed more automatic reading of handwritten documents of the type checks or postal mail. Case studies focus on the recognition of handwritten characters from the segmentation of figures such as the digital amount of a check or postal code for an address or segmentation lowercase letters and / or capital in the case of such literal amount of a check or city name on an address. These isolated characters exhibit strong variations primarily caused by the position of the letter in the word.

Building a handwriting recognition system comprises several distinct stages. The steps that are described by Duda et al. [2] are shown in Figure 1.1. The system acquires an input form or a signal from a sensor (camera, scanner, tape recorder, etc.). This input can be a scanned image or a voice signal to be stored in a file. Subsequently several processes are made on these images and files. The purpose of these pre-processing is to eliminate the phenomena which cause degradation of system performance, reduce quantization noise (binarization) and preserve the connectivity of the connected components in the image [3]. The result of this phase will extract or highlight local or global features. This step will generate for each image, a vector features that serves as input to the head of the classification module [4].

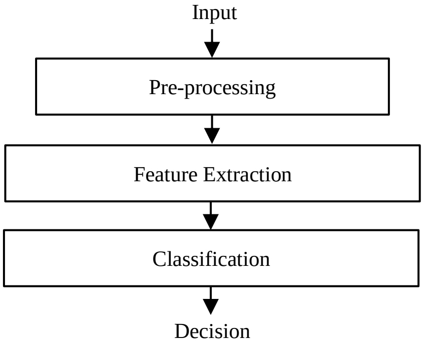

Figure 1: The steps of character recognition

.

II. Literature Review

Gauri Katiyar et al. presented a handwritten character recognition framework on the basis of assessment of neural network by conjoining various feature extraction methods (Mean and Gradient

Operations, Box Approach and Diagonal Distance Approach). The CEDAR (Centre of Excellence for Document Analysis And Recognition) dataset was taken for proposed framework [1]

Gaur et al. presented a three phase approach for recognition of characters. Pre-processing is performed in the initial phase which contains image binarization and characters separations. The second phase is feature extraction which is accomplished by k-means clustering and the feature vector is generated. Third and final phase is classification which is performed by support vector machine (SVM) [5].

These days character recognition has picked up a considerable measure of consideration in the arena of pattern recognition because of its usage in different aspects. Handwritten Character Recognition (HCR) and Optical Character Recognition (OCR) has particular space to relate. OCR framework is most reasonable for the uses like multi decision investigations, published mailing address determination and so on. While utilization of HCR is more extensive contrasted with OCR. HCR is helpful in a wide range of form processing frameworks and a great deal more. In future, character recognition framework may assist as a key component to make a paperless domain by digitizing and handling current paper forms. Sameeksha Barve gave a neural network based approach to deal with recognition of optical or visual characters. OCR (Optical Character Recognition) System or to enhance the accuracy of a current one. The utilization of artificial neural network rearranges advancement of an optical character recognition application, while accomplishing most elevated accuracy of recognition and great performance [6].

Sameeksha Barve exhibited an Optical character recognition framework in view of Artificial Neural Networks (ANNs). Training of neural network is accomplished by back propagation algorithm [7].

Handwritten character recognition is dependably an outskirts region of research in the field of pattern recognition and there is an expansive interest for Optical Character Recognition on hand written archives. Shabana et al. give an exhaustive survey of existing works in handwritten character recognition in light of soft computing strategy during the previous decade [8].

Patel et al. suggested a technique for the recognition of handwritten characters. The multiresolution approach is accomplished using DWT and Euclidean distance formula. This approach was found to be faster in the form of 26 patterns of characters. DWT features of handwritten characters are extracted with multiresolution approach, thereafter a mean vector is derived for every pattern class. Euclidean distance is measured between input vector and the mean vectors. The membership function for input vector is accomplished by the minimum distance. This technique gives great recognition precision of 90% [9].

R. K. Mandal et al. proposed another strategy to enhance the execution of the beforehand connected strategies. The input image matrix is compacted into a lower dimension matrix with a specific end goal to lessen non-significant features of the image matrix. The compressed matrix is segmented column-wise. Every segment of a specific image matrix is mapped to indistinguishable patterns for perceiving a specific character. The larger part of a known pattern chooses the presence of a specific character [10].

Pramod J Simha et al. exhibited the abilities of Artificial Neural Network usage in recognition of extended sets of optical language images. This paper portrays a propelled arrangement of classification utilizing probabilistic neural systems. Training of the classifier begins with the utilization of mutilation displayed characters from text styles. Factual measures are taken up against an arrangement of features figured from the twisted character [11].

Mitrakshi B. Patil et al. presented an offline recognition technique for characters using ANN systems. This approach is divided into two phases; the primary phase is the separation of characters into the line, word and characters while the secondary phase utilized the feed-forward neural network algorithm for recognition of chatacters [12].

Tirtharaj Dash et al. built up an Offline Hand Written English Character Recognition in light of Artificial Neural Network (ANN). The ANN executed in this work has single yield neuron which indicates whether the test character has a place with a specific cluster or not. The execution is done totally in "C" language. Ten arrangements of English letter sets (little 26, capital-26) were utilized to prepare the ANN and 5 sets of English letter sets were utilized to test the system. The characters were gathered from several persons over a span of nearby 25 days. The approach was tested with the sets of having 5 small letters and 5 capital letters [13].

Amarjot Singh et al. depicted an overview of uses of OCR in various fields and further exhibits the experimentation for three essential applications, for example, Captcha, Institutional Repository and Optical Music Character Recognition. This paper makes utilization of an improved image segmentation approach in light of histogram equalization utilizing genetic algorithms for optical character recognition [14].

J. Pradeep et al. depicted an offline handwritten alphabetical character recognition framework utilizing multilayer feed forward neural network system. A strategy, called, diagonal based feature extraction is presented for separating the features of the manually written letter sets. Fifty information sets, each containing 26 letters in order composed by different individuals, are utilized for preparing the neural system and 570 diverse handwritten in order characters are utilized for testing [15].

III. Theoretical Framework

A. *Introduction to OCR*

Optical Character Recognition (OCR) is the subject of the future of human-machine communication. It is used in several areas where the text is the basis of work, mainly in office automation, for purposes of indexing and automatic archiving of documents, in desktop publishing (DTP) to facilitate the composition from a selection of several documents, in the post for the automatic sorting of the mail, in a bank to facilitate the reading of the check amounts.

The recognition of writing is in the domain of pattern recognition that deals with character forms. The goal is to assign to a form an identifier of the previously determined reference prototypes. Optical Character Recognition (OCR) research, although less advanced than other languages, is becoming more intensive than before.

Any written information can be used in a computerized processing chain for different purposes: the writing and editing of reports, the distribution of documents in a messaging system lead to exploiting information available only on paper. Optical Character Recognition (OCR) is a fast computing operation that enables the transformation of a text written on paper into a text in the form of a computer file in symbolic representation of ASCII code (American Standard Code for Information Interchange).

B. *Different Aspects of OCR*

There is no universal OCR system that can recognize any character in any font. It all depends on the nature of data administered and obviously on the anticipated use [16]. There are several modes of classification of OCR systems among which we can mention:

- Systems qualified as "on-line" or "offline" depending on the acquisition mode.
- Global or analytical approaches depending on whether the analysis operates on the whole word, or by segmentation in characters.
- Statistical, structural or stochastic approaches to characteristic features extracted from the forms considered.

1) Online and Offline Recognition

These are two different OCR modes, each with its own acquisition tools and corresponding recognition algorithms.

a) Online Recognition

This recognition mode operates in real time (during writing). Symbols are recognized as they are written by hand. This mode is generally reserved for handwriting. It’s an approach signal where the recognition is performed on one-dimensional data. The writing is represented as a set of points whose coordinates are function of time [17], [18].

Online recognition has a major advantage is the possibility of correction and modification of writing interactively given the continuous response of the system [19]. The acquisition of writing is usually provided by a graphics tablet with an electronic pen.

b) Offline Recognition

Starts after the acquisition. It is suitable for printed documents and manuscripts already written. This mode can be considered as the most general case of the recognition of writing. It is getting closer to the mode of visual recognition. The interpretation of the information is independent of the generation source [19].

Offline recognition can be classified into several types:

Recognition of Text or Analysis of Documents: In the first phase it is a question of recognizing a text of structure limited to some sentences or words. The search is accomplished by a general recognition of words in the sentences followed by sepration of each word into characters [16]. In second phase (document analysis), it is well-structured data whose reading obliges understanding of the layout and typography of the document. Here the simplified pre-processing approach is not used while a specific method is used which contains [20]:

- Localization of regions.
- Separation of graphic and photographic regions.
- Semantic labeling of textual areas from models.
- Determination of order of reading and the structure of the document.

Recognition of Print or Manuscript: The approaches differ depending on whether print or manuscript recognition is involved. The printed characters are in the general case horizontally aligned and separated vertically, which simplifies the reading phase [16]. The shape of the characters is defined by a calligraphic style (font) which constitutes a model for identification. In the case of the manuscript, the characters are often ligated and their graphism is unevenly proportioned from within and intercribing variability. This usually requires the use of specific allocation methods and frequently appropriate understanding to guide interpretation [21].

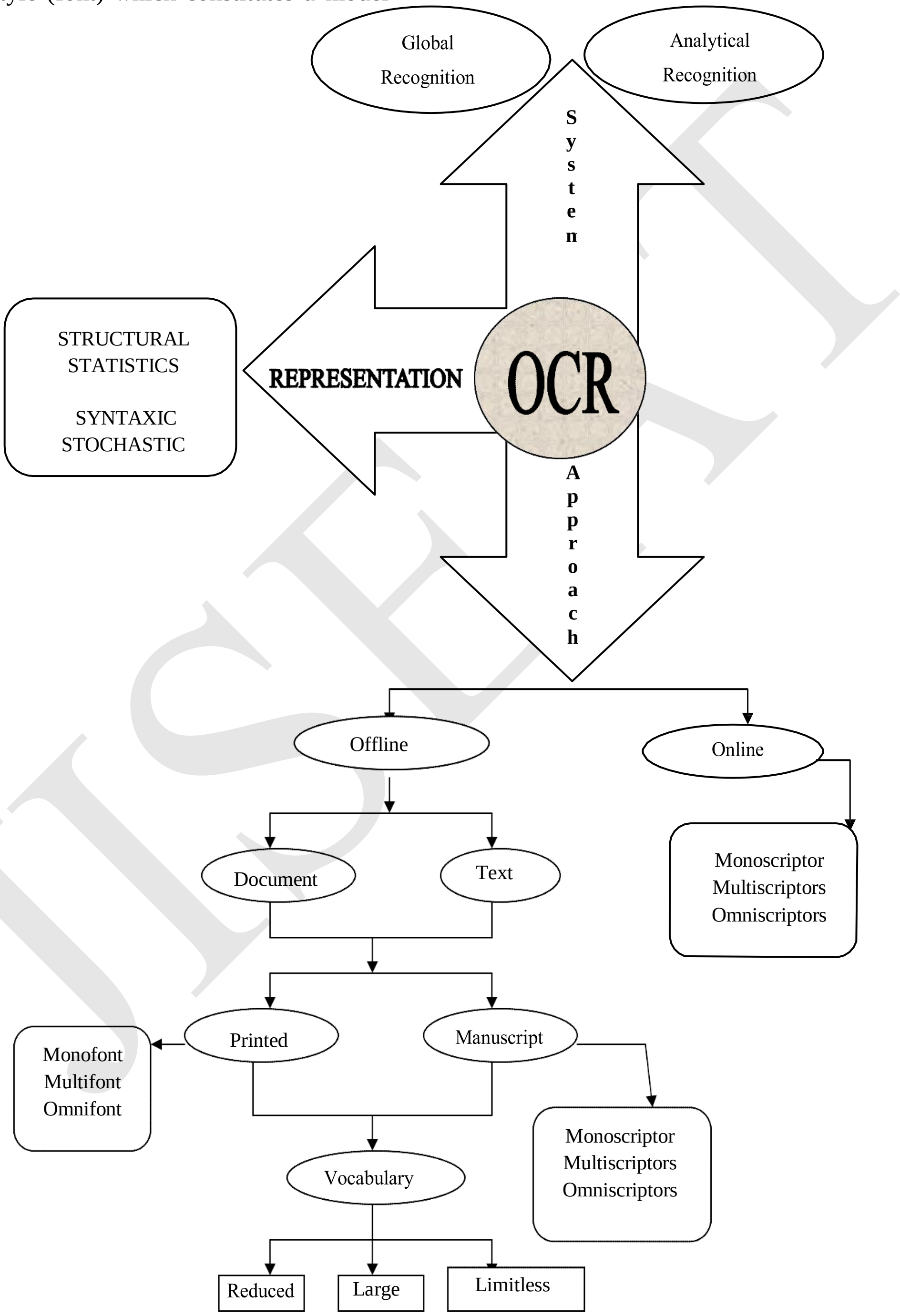

Figure 3.1: Different systems, representations and approaches of recognition [16]

In the case of print, the recognition can be monofonte, multifonte or omnifonte. A system is said to be monofonte if it can recognize only one font at a time that is to say that it knows graphics only from a single font. This is the simplest case of recognition of printed characters. A system is said to be multifont if it is able to recognize various types of fonts among a set of previously learned fonts [16]. And an omnifonte system is able to recognize any font, usually without prior learning. However this is almost impossible because there are thousands of fonts, some of which are unreadable by humans (except of course for the one who designed it) and with any font creation software anyone can design fonts at will.

2) Global or Analytical Recognition

The global approach considers the word as a single entity and describes it independently of the characters that make it up. This approach has the advantage of keeping the character in its surrounding context, which allows a more effective modeling of the variations of the writing and the degradation that it can undergo. However this method is penalizing by the memory size, the computation time and the complexity of the treatment which believe linearly with the size of the lexicon considered, from where a limitation of the vocabulary [18].

The Analytic Approach: unlike the global approach, the word is separated into characters or significant morphological fragments smaller than the character called graphemes. Word recognition consists in recognizing the segmented entities and then tending towards a recognition of the word, which constitutes a delicate task that can generate different types of errors, [17]. A recognition process according to this approach is based on an alternation between two phases: the segmentation phase and the segment identification phase. Two solutions are then possible: explicit (external) segmentation or implicit (internal) segmentation [22]. Moreover, the analytical methods as opposed to the global methods, have the advantage of being able to generalize to the recognition of a vocabulary without limit a priori, because the number of characters is naturally finished. Moreover, the extraction of primitives is easier on a character than on a string of characters [23].

C. Problems Related to OCR

The task of the OCR is not easy, various problems complicate the process of recognition, among which we can mention [18], [16]:

The Quality of the Document: A document that is faxed or photocopied multiple times is harder to process than the original copy. The writing can become thinner or on the contrary thicker, degraded with parts of the text that are missing or tasks that appear, openings or clogged loops.

Printing: A composite document is of better quality than a typewritten document which, in turn, is clearer than text from a dot matrix printer. An inkjet printer can introduce ink stains and spreading characters, a laser printer can generate lines or backgrounds.

The Discrimination of the Form: According to the style of the font used, its body and its fat, the character changes of graphic design. The number of shapes is all the more important as the number of writing styles is high. In addition, several characters have a strong resemblance such as:

- For English: U and V, O and 0, S and 5, Z and 2.

Information Support: such as paper, also plays on the performance of recognition by its quality: its grammage, granulation and color.

Acquisition: Real-time scanning often introduces distortions in the image. In the off-line case the quality of the scanned text is a compromise between the variations of the position (inclination, translation and shrinkage) the cleanliness of the glass of the digitizing device and its resolution.

Variations in Dimensions: A pitch of 10, 12 or 16 ... (10, 12 or 16 cpi (character per inch)). A pitch of 10 implies larger characters both in width and height than those of a pitch of 12.

In addition to these problems, an OCR system should be able to distinguish between a text and a figure, to recognize the ligature characters and to be independent of the variations of the space as well inter-words as of the line spacing.

The problems posed by the optical recognition of handwriting are more complex than those related to printed writing. The reading errors in the case of the manuscript are due to the infinite variations of the writing of random nature which depend on particular factors of the writer and the conditions of writing.

IV. Conclusion

Optical character recognition is one of the most established thoughts in the automated pattern recognition. In recent time, character recognition turns into the field of useful utilization. With character acknowledgment, the procedure begins with the perusing of a scanned picture of a progression of characters, decides their meaning, and lastly makes an interpretation of the picture to a computer written text content. Mostly, this procedure is done generally in the post-offices to automatically read the addresses and names and on wrappers and by the banks to peruse acknowledge the amount and number of checks. Likewise, different organizations and persons can utilize this technique to rapidly make an interpretation of paper reports to computer written documents.

REFERENCE

[1] Gauri Katiyar, Shabana Mehfuz, "MLPNN Based Handwritten Character Recognition Using Combined Feature Extraction", IEEE, International Conference on Computing, Communication and Automation (ICCCA2015), pp. 1155-1159, July 2015.

[2] Duda, R.O., Hart, P.E. and Stork, D.G., 2012. Pattern classification. John Wiley & Sons.

[3] Jayadevan, R., Kolhe, S.R., Patil, P.M. and Pal, U., 2011. Offline recognition of Devanagari script: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41(6), pp.782-796.

[4] Kolman, E. and Margaliot, M., 2008. A new approach to knowledge-based design of recurrent neural networks. IEEE Transactions on Neural Networks, 19(8), pp.1389-1401.

[5] Gaur, A. and Yadav, S., 2015, January. Handwritten Hindi character recognition using k-means clustering and SVM. In Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), 2015 4th International Symposium on (pp. 65-70). IEEE.

[6] Sameeksha Barve, "Artificial Neural Network Based On Optical Character Recognition", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 1, Issue 4, June 2012.

[7] Sameeksha Barve, "Optical Character Recognition Using Artificial Neural Network", International Journal of Advanced Technology & Engineering Research (IJATER), ISSN NO: 2250-3536 Volume 2, Issue 2, May 2012.

[8] Shabana Mehfuz, Gauri katiyar, "Intelligent Systems for Off-Line Handwritten Character Recognition: A Review", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2, Issue 4, April 2012.

[9] Dileep Kumar Patel, Tanmoy Som, Sushil Kumar Yadav, Manoj Kumar Singh, "Handwritten Character Recognition Using Multiresolution Technique and Euclidean Distance Metric", Journal of Signal and Information Processing, PP. 208-214., 2012.

[10] Rakesh Kumar Mandal, N. R. Manna, "Hand Written English Character Recognition using Column-wise Segmentation of Image Matrix (CSIM)", WSEAS Transactions On Computers, E-ISSN: 2224-2872, Issue 5, Volume 11, May 2012.

[11] Pramod J Simha & Suraj K. V., "Unicode Optical Character Recognition and Translation Using Artificial Neural Network", International Conference on Software Technology and Computer Engineering (STACE-2012), ISBN : 978-93-81693-68-1, 22nd July 2012.

[12] Mitrakshi B. Patil, Vaibhav Narawade, "Recognition of Handwritten Devnagari Characters through Segmentation and Artificial neural networks", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 1 Issue 6, August, 2012.

[13] Tirtharaj Dash, Tanistha Nayak, "English Character Recognition using Artificial Neural Network", Proceedings of National Conference on AIRES, Andhra University, 2012.

[14] Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin, "A Survey of OCR Applications", International Journal of Machine Learning and Computing, Vol. 2, No. 3, June 2012.

[15] J. Pradeep, E. Srinivasan, S. Himavathi, "Diagonal Based Feature Extraction For Handwritten Alphabets Recognition System Using Neural Network", International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

[16] Levin, E. and Pieraccini, R., 1993. Planar hidden Markov modeling: From speech to optical character recognition. In Advances in Neural Information Processing Systems (pp. 731-738).

[17] Lecolinet, E. and Baret, O., 1994. Cursive word recognition: Methods and strategies. In Fundamentals in Handwriting Recognition (pp. 235-263). Springer, Berlin, Heidelberg.

[18] Al-Badr, B. and Mahmoud, S.A., 1995. Survey and bibliography of Arabic optical text recognition. Signal processing, 41(1), pp.49-77.

[19] Lallican, P.M., Viard-Gaudin, C. and Knerr, S., 2000, September. From off-line to on-line handwriting recognition. In Proceedings of the seventh international workshop on frontiers in handwriting recognition (pp. 303-312).

[20] Trenkle, J., Schlosser, S. and Gillies, A., 1997. An off-line Arabic recognition system for machine-printed documents. Ann Arbor, 1001, pp.48113-4001.

[21] Sawant, S. and Baji, S., 2016. Handwritten character and word recognition using their geometrical features through neural networks. International Journal of Application or Innovation in Engineering & Management (IJAIEM), 5.

[22] Casey, R.G. and Lecolinet, E., 1995, August. Strategies in character segmentation: A survey. In Document Analysis and Recognition, 1995. Proceedings of the Third International Conference on (Vol. 2, pp. 1028-1033). IEEE.

[23] Al-Badr, B. and Haralick, R.M., 1994, March. Symbol recognition without prior segmentation. In Document Recognition (Vol. 2181, pp. 303-314).