A Review on Character Recognition Using OCR Algorithm

MamtaKadyan
M.Tech Scholar,N.C.College of Engineering, Israna, Panipat, India.

Deepti Ahlawat
Assistant Professor ,N.C.College of Engineering, Israna, Panipat, India.

Abstract – Optical Character Recognition has a number of applications in day-to-day existing or OCR has been commonly applied to the plate identification, barcode recognition, legal billing document, finance and insurance documents many others. An optical character recognition system with a high identification rate is challenging to expand. One of the major features to OCR errors is mark characters. Many factor lead to the mark of nature such as bad scanning quality and a poor linearization technique. Typical path to nature segmentation go down into three major categories: Picture-based, identification-based, and holistic-based. By the whole of these approaches, the segmentation path can be linear or non-linear. Simply and versatile cameras build it possible to easily and quickly capture a broad variety of official paper. Even so, low resolution cameras present a difficult task to OCR because it is essentially not possible to do character segmentation independently from recognition. Handwriting approval has been one of the almost all interesting and inflexible analysis areas in discipline of photograph processing and pattern approval in the latest years. This paper relates the strategies for change textual content from a paper file into system readable shape. The computer honors accept the characters within the file through a transforming approach make reference to as Optical Character Recognition. This paper, many techniques like OCR using correlation approach and OCR the usage of neural networks has been mentioned.

Index Terms – Optical character recognition(OCR), segmentation, holistic approach, neural network.

1. INTRODUCTION

Highlight in 1950's [1], applied all through the spectrum of industries resulting into revolutionizing the document management system. Optical Character Recognition or OCR has enabled scanned documents to emerge as extra as just image documents, turning into fully searchable files with textual content material diagnosed by using computers. Optical Character Recognition extracts the relevant data and mechanically enters it into electronic database rather than the conventional way of manually retyping the textual content. Optical Character Recognition is a procedure by way of which we convert printed record or scanned web page to ASCII person that a pc can apprehend.[3] The report photograph itself can be both device printed or handwritten, or the aggregate of .OCR has three processing steps, Document scanning manner, Recognition manner and Verifying technique. In the report scanning step, a scanner is used to test the handwritten or published documents. The excellent of the scanned record depends up on the scanner. So, a scanner with high pace and shade fine is perfect. The recognizing system consists of several complicated algorithms and formerly loaded templates and dictionary which might be crosschecked with the characters within the record and the corresponding system editable ASCII characters. The verifying is finished either randomly or chronologically by way of human Intervention. Difference in font and sizes makes popularity undertaking hard if preprocessing, characteristic extraction and recognition aren't robust. There may be noise pixels which can be introduced because of scanning of the image. Besides, equal font and size can also have formidable face character as nicely as regular one. Thus, width of the stroke is likewise a factor that affects reputation. Therefore, a terrific character recognition method have to cast off the noise after studying binary picture facts, smooth the photograph for higher popularity, extra capabilities correctly, teach the gadget and classify styles. Segmentation of a document into traces and words and of words into person characters and logos represent a vital project within the optical analyzing of texts. Presently, maximum recognition mistakes are because of character segmentation errors. Very frequently, adjacent characters are touching, and can also exist in an overlapped. Therefore, it's miles a complex challenge to phase a given phrase correctly into its individual components. The system of handwriting recognition entails extraction of a few described traits called functions to classify an unknown handwritten person into one of the acknowledged classes. A regular handwriting recognition gadget includes numerous steps, particularly: preprocessing, segmentation, feature extraction, and classification, several types of decision strategies, along with statistical strategies, neural networks, structural matching (on trees, chains, and so forth). The stochastic processing (Markov chains, and so on.) have been used at the side of special types of functions [5]. The benefit of HMM method over ANN approach in optical man or woman reputation is that it could be effortlessly extendible to the reputation of handwritten characters. In this paper, we are able to talk how artificial

neural network, genetic set of rules and fuzzy common sense may be used in optical man or woman popularity for the use of character recognition. The closing part of this survey paper is prepared as follows:-In phase II, we can discuss the literature overview in the area of person reputation and in section III we describe the diverse strategies used for individual reputation the usage of OCR, the comparative look at of strategies mentioned in phase III given in phase IV and in segment V, we will conclude the paper and supply the destiny scope of this paper.

2. CHARACTER RECOGNITION TECHNIQUES

Optical Character Recognition may be applied to recognize text from any multimedia along with image, audio, video. Automatic multimedia popularity is primarily based on the laptop imaginative and prescient and pattern reputation software [16]. We can use photo processing, character positioning, character segmentation, neural community to clear up the problem of picture to text recognition.

2.1 HMM Approach

A hidden Markov version is a doubly stochastic process, with an underlying stochastic manner that isn't observable (for this reason the word hidden), however may be observed through some other stochastic process that produces the series of observations [11],[14],[16][17]. The hidden method consists of a hard and fast of states linked to every other by way of transitions with probabilities, even as the discovered system includes a set of outputs or observations, each of which may be emitted through every state in line with a few output chance density feature (PDF) [9][12]. Depending on the character of this PDF function numerous sorts of HMMs can be distinguished.

2.2 Neural community technique

Character Recognition within the registration code popularity has crucial position in optical popularity machine which is related at once with success or failure of the machine. We used Back Propagation Neural Network to optically understand the photograph. The primary idea of BP set of rules is the gaining knowledge of manner is split into stages:

PHASE I: Forward Propagation

PHASE II: Back Propagation

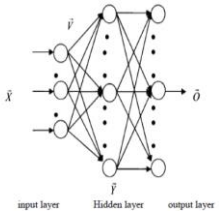

Figure 2.1: Character reputation the use of neural network three.

2.3 Character Normalization

It is vital to normalize the individual, letters and numbers to standard length. We can normalize the characters of different size into one fixed length to make our mission smooth for optical character popularity Figure.

Figure 2.2: Normalization

2.4 Correlation technique for unmarried man or woman popularity Preprocessing

The photograph is taken and is transformed to grey scale photo. The grey scale photograph is then converted to binary image. This procedure is known as Digitization of photo. Practically any scanner is not ideal; the scanned image may additionally have a few noises. This noise may be because of a few pointless information gifts in the picture. So, all of the items having pixel values less than 30 are eliminated. The de-noised image hence received is stored for further processing. Now, all the templates of the alphabets which are pre-designed are loaded into the machine.

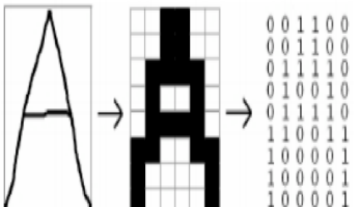

Figure 2.3: Digitized Image

2.4.1Segmentation

In segmentation, the location of the item i.e., the man or woman within the photo is found out and the dimensions of the photo is cropped to that of the template length. Recognition: The image from the segmented level is correlated with all of the templates which are preloaded into the gadget. Once the correlation is completed, the template with the maximum correlated price is said as the person gift inside the photograph.

2..4.2Recognition

The picture from the segmented stage is correlated with all of the templates which might be preloaded into the gadget. Once the correlation is completed, the template with the most correlated cost is asserted because the person present inside the photo.

3. REVIEW OF LITERATURE

A quick description of the history of OCR is as follows. In 1929 Gustav Tauschek obtained a patent on OCR in Germany, observed with the aid of Handel who received a US patent on OCR in USA in 1933. In 1935 Tauschek turned into additionally granted a US patent on his technique. Tauschek's device becomes a mechanical device that used templates and a image detector.

G. Siebra Lopes, *et.al* [1] There is a growing need for recognition of digits manuscripts for use in various situations, such as recognition of handwritten postal address digits for automated redirection of letters in the mail, acknowledgment of nominal values in bank checks. Recognition of handwritten digits faces great difficulty in dealing with intra-class variation due to different writing styles, different degrees of inclination of the characters. Optical character recognition systems, also known as OCR, identifying and recognizing printed characters through images, an already widespread functionality in scanners, mobile devices, among others. This paper presents the use of the classifier Optimum-Path Forest (OPF) applied in handwriting recognition digits. A new feature extraction method is proposed using signature of the characters, and the OPF algorithm is used in the classification.

G. Peng, *et.al* [2] The text line segmentation process is a key step in an optical character recognition (OCR) system. Several common approaches, such as projection-based methods and stochastic methods, have been put forward to fulfill this task. However, most of existing methods cannot be directly applied to process the palm leaf manuscripts of Dai which the images have poor quality and include smudges, creases, stroke deformation and character touching. To solve this problem, an improved Viterbi algorithm based on Hidden Markov Model (HMM) is proposed to find all possible segmentation paths firstly. And then, a path filtering method is used to detect the optimal paths for the segmented text blocks. The performance of the method is compared with relevant methods and the experimental results demonstrate the effectiveness of the proposed method.

A.Sanjrani, *et.al* [3] Sindhi language is script language like Arabic and Persian. Its origin is 2500 years old and spoken in various countries in Asia. In this paper, we propose an Optical Character Recognition (OCR) system which recognizes handwritten Sindhi numeral expressions (i.e. Sindhi handwritten numeral strings) without using common input devices such as keyboard and storage device memory. Our experiments focus on character recognition which later can be used for various applications such as tutoring, mathematical kids games, and automatic telephone number conversion from sign boards in India and Pakistan.

J. Ryu, *et.al* [4] Segmentation of handwritten document images into text-lines and words are an essential task for optical character recognition. However, since the features of handwritten document are irregular and diverse depending on the person, it is considered a challenging problem. In order to address the problem, we formulate the word segmentation problem as a binary quadratic assignment problem that considers pairwise correlations between the gaps as well as the likelihoods of individual gaps.

S. F. Rashid, *et.al* [5] Optical character recognition (OCR) of machine printed Latin script documents is ubiquitously claimed as a solved problem. However, error free OCR of degraded or noisy text is still challenging for modern OCR systems. Most recent approaches perform segmentation based character recognition. This is tricky because segmentation of degraded text is itself problematic. This paper describes segmentation free text line recognition approach using multi layer perceptron (MLP) and hidden markov models (HMMs).

B. VijayKumar, *et.al* [6] Neural network based radial basis function networks (RBFN) and subspace projection approach have been employed to recognize printed Kannada characters. RBFNs are trained with wavelet features using K-means and subspace method is applied on normalized image. Use of structural features for disambiguating confused characters improved the recognition accuracy by 3% in case of subspace and by 1.6% using RBFN. Compared to subspace, a maximum recognition rate of 99.1% is achieved with RBFN using Haar wavelets and structural features.

S. Farkya, *et.al* [7] Handwritten optical character recognition is one of the major research area due to its complexity in segmenting the character which increases in the case of Devnagri Script due to Modifiers and compound characters. Thus this paper shows an adaptive segmentation technique which shows less error. This paper shows an implementation of Handwritten Devanagri character recognition system. Keeping in mind the impairment of blind people, OCR system is extended to Text to Speech System.

S. H. Tanvir,*et.al* [8] This paper needed to train our classifier in case we are considering to use data mining techniques for such purposes. There are several established generic classification techniques that can be used together with feature extraction mechanisms but it is important to know which of them do better under which circumstances. This evaluates three approaches for OCR from handwritten manuscripts and also studies their results.

Ahmad, *et.al* [9] This paper proposed the use of stacked denoising auto encoder for automatic feature extraction directly from raw pixel values of ligature images. Such deep learning networks have not been applied for the recognition of Urdu text thus far. Subsequently, trained networks are validated and tested on degraded versions of UPTI data set. The experimental results demonstrate accuracies in range of 93% to

96% which are better than the existing Urdu OCR systems for such large dataset of ligatures.

T. Mantoro, *et.al* [10] this paper proposed a framework of Optical Character Recognition (OCR) on mobile device using server-based processing. Comparison methods proposed by this paper by conducting a series of tests using standalone and server-based OCR on mobile devices, and compare the results of the accuracy and time required for the entire OCR processing.

W. Q. Khan, *et.al* [11] This paper developed a technique using point feature matching on cropped Urdu newspaper clippings with font Jameel Noori Nastaleeq and converted them into editable textual Unicodes. The objective of the technique is that could be applied to any Urdu script font size, without worrying about the variation of characters/words caused by the disposal of ink in Urdu newspaper clippings.

R. E. Precup, *et.al* [12] The NIOAs minimize the objective functions to achieve optimal fuzzy control systems with reduced parametric sensitivity, and optimal PI-FCs for nonlinear servo systems are offered. The NIOAs are next applied to the optimal tuning of the parameters of Takagi-Sugeno fuzzy models for Anti-lock Braking Systems and for magnetic levitation systems. The NIOAs are inserted in optimal path planning algorithms for mobile. The multi-objective optimization is considered as the NIOAs use two to four objective functions to generate optimal trajectories for mobile robots in static environments while avoiding collisions with the obstacles and danger zones that might exist in the environment. The NIOAs solve the optimization problems by minimizing the objective functions, producing optimal collision-free trajectories in terms of minimizing the length of the paths and also assuring that the generated trajectories are at a safe distance from the danger zones.

E. Hassan, *et.al* [13] The paper proposed a novel multi-modal document image retrieval framework by exploiting the information of text and graphics regions. The framework applies multiple kernel learning based hashing formulation for generation of composite document indexes using different modalities. In the subsequent contribution propose novel multi-modal document indexing framework for retrieval of old and degraded text documents by combining OCR'ed text and image based representation using learning.

K. Ait-Mohand,*et.al* [14] This paper presented two algorithms that extend existing HMM parameter adaptation algorithms (MAP and MLLR) by adapting the HMM structure. This improvement relies on a smart combination of MAP and MLLR with a structure optimization procedure. Structure optimization is based on state splitting and state merging operations and proceeds so as to optimize either the likelihood or a heuristic criterion. These algorithms are successfully applied to the recognition of printed characters by adapting the HMM character models of a polygons printed text recognizer to new fonts.

D. Tao, *et.al* [15] for robust CCFR, this integrate a principal component convolution layer with the 2-D long short-term memory (2DLSTM) and develop principal component 2DLSTM (PC-2DLSTM) algorithm. PC-2DLSTM considers two aspects: 1) the principal component layer convolution operation helps remove the noise and get rational and complete font information and 2) simultaneously, 2DLSTM deals with the long-range contextual processing along scan directions that can contribute to capture the contrast between character trajectory and background. Experiments using the frequently used CCFR dataset suggest the effectiveness of PC-2DLSTM compared with other state-of-the-art font recognition methods.

C. C. Chang, *et.al* [16] An optical identification (ID) system has been built, and dynamic ID tags are compared, which include a home-made tunable cat's eye retro reflector array, a liquid crystal (LC) spatial light modulator (SLM) combined with a retro reflective tape, a LC SLM combined with a plane mirror, and a LC SLM combined with normal white paper.

4. CONCLUSION

A range of strategies that are used for optical character popularity have been discussed which makes use of correlation and neural networks. Much other development in Optical Character Recognition is being under improvement. The paper provides a short survey of the packages in numerous fields together with experimentation into few selected fields. The proposed technique is extremely green to extract all styles of bimodal snap shots such as blur and illumination. The paper will act as a good literature survey for researchers beginning to work within the discipline of optical individual reputation. The cause of its complexities are its characters shapes, its top bars and quit bars extra over it has some changed, vowel and compound characters and additionally one of the important motives for poor recognition in OCR system is the error in character recognition.

REFERENCES

[1] G. Siebra Lopes, D. Clifte da Silva, A. W. Oliveira Rodrigues and P. P. Reboucas Filho, "Recognition of handwritten digits using the signature features and Optimum-Path Forest Classifier," IN IEEE Latin America Transactions, vol. 14, no. 5, pp. 2455-2460, May 2016.

[2] G. Peng, P. Yu, H. Li and L. He, "Text line segmentation using Viterbi algorithm for the palm leaf manuscripts of Dai," 2016 International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, 2016, pp. 336-340.

[3] A. Sanjrani, J. Baber, M. Bakhtyar, W. Noor and M. Khalid, "Handwritten Optical Character Recognition system for Sindhi numerals," 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube), Quetta, 2016, pp. 262-267.

[4] J. Ryu, H. I. Koo and N. I. Cho, "Word Segmentation Method for Handwritten Documents based on Structured Learning," in IEEE Signal Processing Letters, vol. 22, no. 8, pp.1161-1165,Aug.2015. doi: 10.1109/LSP.2015.2389852.

[5] S. F. Rashid, F. Shafait and T. M. Breuel, "Scanning Neural Network for Text Line Recognition," 2012 10th IAPR International Workshop on Document Analysis Systems, Gold Cost,QLD,2012,pp.105-109. doi: 10.1109/DAS.2012.77.
[6] B. VijayKumar and A. G. Ramakrishnan, "Radial basis function and subspace approach for printed Kannada text recognition," 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, pp. V-321-4 vol.5.
[7] S. Farkya, G. Surampudi and A. Kothari, "Hindi speech synthesis by concatenation of recognized hand written devnagri script using support vector machines classifier," 2015 International Conference on Communications and Signal Processing (ICCSP), Melmaruvathur, 2015, pp. 0893-0898.
[8] S. H. Tanvir, T. A. Khan and A. B. Yamin, "Evaluation of optical character recognition algorithms and feature extraction techniques," 2016 Sixth International Conference on Innovative Computing Technology (INTECH), Dublin, Ireland, 2016, pp. 326-331.
[9] I. Ahmad, X. Wang, R. Li and S. Rasheed, "Offline Urdu Nastaleeq optical character recognition based on stacked denoising autoencoder," in China Communications, vol. 14, no. 1, pp. 146-157, Jan. 2017.
[10] T. Mantoro, A. M. Sobri and W. Usino, "Optical Character Recognition (OCR) Performance in Server-Based Mobile Environment," 2013 International Conference on Advanced Computer Science Applications and Technologies, Kuching, 2013, pp. 423-428.
[11] W. Q. Khan and R. Q. Khan, "Urdu optical character recognition technique using point feature matching; a generic approach," 2015 International Conference on Information and Communication Technologies (ICICT), Karachi, 2015, pp. 1-7.
[12] R. E. Precup, "Nature-inspired optimization algorithms applied to fuzzy control, fuzzy modeling, mobile robots and optical character recognition," 2014 IEEE 9th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, 2014, pp. 11-11.
[13] E. Hassan, S. Chaudhury and M. Gopal, "Multi-modal Information Integration for Document Retrieval," 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, 2013, pp. 1200-1204.
[14] K. Ait-Mohand, T. Paquet and N. Ragot, "Combining Structure and Parameter Adaptation of HMMs for Printed Text Recognition," *in* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 9, pp. 1716-1732, Sept. 2014.
[15] D. Tao, X. Lin, L. Jin and X. Li, "Principal Component 2-D Long Short-Term Memory for Font Recognition on Single Chinese Characters," in IEEE Transactions on Cybernetics, vol. 46, no. 3, pp. 756-765, March 2016.
[16] C. C. Chang, Y. C. Yang, M. C. Su and J. c. Tsai, "Tunable Micro Cat's Eye Array in an Optical Identification System and Comparison of Different ID Tags," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 21, no. 4, pp. 130-136, July-Aug. 2015.

Authors

I am Mamta Kadyan.I am Pursing M.Tech Form N.C.College of Engineering, Israna (Panipat).My Interested Area Is Image Processing.

Techniques	Author's name	advantages	disadvantages
Optimum-Path Forest(OPF) algorithm	G. Siebra Lopes, D. Clifte da Silva, A. W. Oliveira Rodrigues and P. P. Reboucas Filho	Less expensive	Low mobility
Hidden Markov Model	G. Peng, P. Yu, H. Li and L.	Better performance	poor quality and include smudges
Sindhi numeral expressions	A. Sanjrani, J. Baber, M. Bakhtyar, W. Noor and M. Khalid,	low complexity	Less in use
Binary quadratic assignment	J.Ryu, H. I. Koo and N. I. Cho	Less time to operate	word segmentation problem, irregular Features
Multi layer perceptron (MLP) and hidden markov models (HMMs)	S. F. Rashid, F. Shafait and T. M. Breuel	free text line recognition	noisy text
Radial basis function networks (RBFN)	B. VijayKumar and A. G. Ramakrishnan	Good accuracy	Expensive
Adaptive segmentation technique	S. Farkya, G. Surampudi and A. Kothari	less error	Complex
data mining technique	S. H. Tanvir, T. A. Khan and A. B. Yamin	Low cost	Less mobility
Stacked de-noising auto encoder	I. Ahmad, X. Wang, R. Li and S. Rasheed	96% accuracy	Less performance
Stacked denoising auto encoder	I. Ahmad, X. Wang, R. Li and S. Rasheed	Good accuracy	Difficult to maintain
server-based processing	T. Mantoro, A. M. Sobri and W. Usino,	Easy to portable	Time consuming
point feature matching	W. Q. Khan and R. Q. Khan,	No variation of characters/words	disposal of ink

NIOAs, optimal fuzzy control systems	R. E. Precup	minimizing the length	optimization problems
multi-modal document image retrieval framework	E. Hassan, S. Chaudhury and M. Gopal	Better performance	degraded text
HMM parameter adaptation algorithms (MAP and MLLR)	K. Ait-Mohand, T. Paquet and N. Ragot	Less time to operate	Less performance
2-D long short-term memory (2DLSTM) and develop principal component 2DLSTM (PC-2DLSTM) algorithm	D. Tao, X. Lin, L. Jin and X. Li	remove the noise, good results	Difficult to maintain
Optical identification (ID)	C. C. Chang, Y. C. Yang, M. C. Su and J. c. Tsai	Good precision value	More complex

Table 1.Comparison of Survey