Review of Offline Handwriting Recognition Techniques in the fields of HCR and OCR

Hemant Khandelwal[1], Sakshi Gupta[2], Arihant Kumar Jain[3]

[1,2]B.Tech. Student, [3]Assistant Professor, Department of Computer Science & Engineering,

Jaipur Engineering College & Research Centre, Jaipur, India

Abstract— *Recently, numerous progressions have occurred in the area of Offline Handwriting Recognition, for example, Feature Extraction Techniques, Character Recognition Techniques and so on. HCR, Handwriting Character Recognition is the ability of a framework to interpret intelligible handwritten input from sources, for example, paper records, photos and might be sensed offline by Optical Scanning and Intelligent Word Recognition. Likewise OCR, Optical Character Recognition is the mechanical or electronic transformation of pictures of typed, manually written or printed content into machine-encoded content. It is a typical technique for digitizing printed writings and utilized as a part of machine procedures, for example, Cognitive Computing, Machine Translation and so forth. It is a field of research in Pattern Recognition, Artificial Intelligence and Computer Vision. The aim of this research is to actualize the different methods of offline handwriting recognition like, Support Vector Machine, Artificial Neural Network, Hidden Markov Model etc. in the fields of HCR and OCR. In this paper, different methods are analyzed and included to enhance programmed evaluation framework's ease of use and enhance the character recognition system for the fields of HCR and OCR.*

Keywords— *Artificial Neural Network, Hidden Markov Model, HCR, OCR, Support Vector Machine.*

I. INTRODUCTION

Character recognition is a process to identify individual printed symbol. These symbols may be alphabetic, numeric, punctuation etc. These symbols may either be machine printed or handwritten in a variety of different font and sizes. Here optical character recognition is for printed text and handwriting character recognition is for hand-written text. Optical character recognition technology is a way that enables us to convert printed paper documents, PDF Files, or images captured of printed data into digital format i.e. editable or searchable data [1]. On the other hand, off-line handwriting character recognition deals with scanned handwritten documents. Handwriting Character Recognition involves the automatic conversion of handwritten text in an image into letter codes which are usable within computer and text-processing applications. Precisely character recognition is a

process of detecting and recognizing character from input image and converts it into ASCII or any other equivalent editable form.
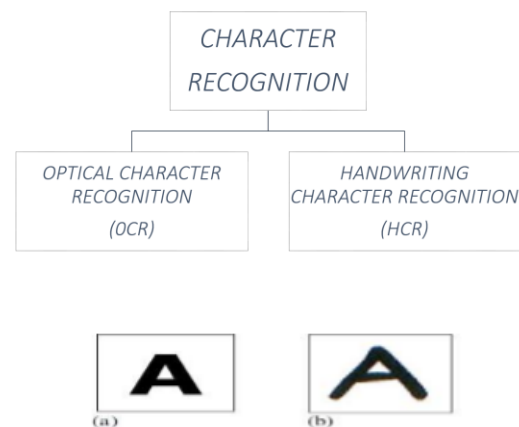

Fig 1: (a) Optical character (b) Handwritten character

A. Optical Character Recognition

Extracting data from scanned documents, camera images or image only PDFs, enabling you to access and edit the content is what OCR means. Advanced OCR systems can read text in large variety of fonts, but they still have difficulty with handwritten text. The potential of OCR systems is enormous because they enable users to harness the power of computers to access printed documents. OCR is already being used widely in the legal profession, where searching once required hours or days can now be accomplished in a few seconds. OCR is being used by libraries to digitize and preserve their holdings. OCR is also used to process checks and credit card slips and sort the mail. Billions of magazines and letters are sorted every day by OCR machines, considerably speeding up mail delivery [2].

B. Handwriting Character Recognition

Handwriting character recognition [3] is similar to optical character recognition, but is a slightly different and more difficult process since OCR is from printed text as opposed to handwritten characters. The storage of handwritten documents has to be bulky in size and many processing applications such as searching, editing, maintenance are either hard or impossible. Handwriting character recognition system translates such images of

documents into machine encoded text. It is really a challenging issue to develop and to identify a practical handwritten character recognition system which can maintain high recognition accuracy. Handwriting Character Recognition is one of the most important topic to study having its applications in cheque processing, form processing, mail address reading, and extracting data from handwritten old documents.

II. COMPONENTS OF OCR AND HCR SYSTEM

A. Optical Scanning

OCR [4] optical scanners consist of a transport mechanism plus a sensing device. Sensing device convert the multi-image into bi-level image. A bi-level image or a binary image is a series of black text written on a white background. Thus it induces uniformity to the input image. This process is known as thresholding and is performed on a scanner to save memory space and computational effort.

B. Pre-Processing

The image resulting from scanning process may have some noise elements like distorted line segments, bumps and gaps in lines, filled loops etc [5]. It is necessary to remove these noise elements prior to further steps. A pre-processor denoise the image and enhances its quality. For handwritten characters, some variants in an image are also removed, which do not affect the identity of the word. It is known as normalization. It includes image cleaning, skew correction, slant and slope removal and character size normalization [6].

C. Segmentation

In the segmentation stage, an image consisting of a sequence of characters is decomposed into sub-images of individual characters. The main goal is to divide an image into parts that have a strong correlation with objects or areas of the real world contained in the image. Segmentation is very important for recognition system. Image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. In Character Recognition techniques, the Segmentation is the most important process. Segmentation is done to make the separation between the individual characters of an image. The goal of segmentation is to simplify and change the representation of an image into something that is more meaningful and easier to analyse.

D. Feature Extraction

The objective of feature extraction [7] is to capture those properties that can identify a character uniquely and maximizes the recognition rate. These features are used to train the system. The features extracted are given as input to the classification stage and the output of this stage is a recognised character. The selection of the combination of feature-classifier contributes to the performance of the system [8].

E. Classification And Recognition

The classification phase is the decision making part of the recognition system. The performance of a classifier based on the quality of the features. This stage uses the features extracted in the previous stage to identify the character. When input image is presented to HCR system, its features are extracted and given as an input to the trained classifier like artificial neural network or support vector machine. Classifiers compare the input feature with stored pattern and find out the best matching class for input.

III. EXISTING METHODOLOGIES FOR OFF-LINE CHARACTER RECOGNITION SYSTEM

Offline character recognition operates on images generated by optical scanner. The approach taken to solve the character recognition problem is based on psychology of the characters as perceived by the humans. Thus the geometrical features of a character and its variants are considered for recognition.

Artificial Neural Network [9] is a system that can perceive and recognize data based on topological features. These features include shape, symmetry, closed and open areas, number of pixels etc. ANN is first trained on samples, and then is used to recognize characters having a similar feature set. ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.

A tool which can be used to train the system with the obtained feature vectors is HMM-Hidden Markov Model. A HMM model [10] keeps information for a character when the model is trained properly and the trained model can be used to recognize an unknown character.

SVM-Support Vector Machine [11] is one of the prominent techniques in pattern recognition and is a successful tool in linear and non-linear classification. SVM is based on statistical learning theory. Today, there are many successful applications in different domains of knowledge based on SVMs, such as text categorization, digital image analysis, character recognition and bioinformatics.

A. Character Recognition Using Artificial Neural Networks

ANN gets its inputs in the form of feature vectors. Every feature or property is separated and assigned a numerical value that can be used to identify a character called its vector. Vector database is usually utilized to train the network so that each character can be effectively get recognized on the basis of its topological properties. Computer vision is aimed at developing the ability to read and in some cases, when ambiguity lies in recognition of a character which needs to be processed as information such as

a message or sign boards. Lexicographic lookup is probably used under the main framework of ANN to remove all the ambiguities.
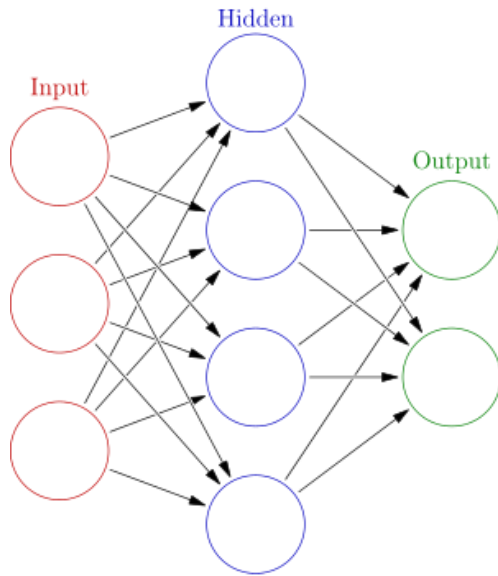

Fig 2: Artificial Neural Networks

Working Methodology

- Any image which needs pre-processing, according to ANN module should get converted into binary form (the one having pixel values 0 & 1 only).
- Binarization converts any image into a series of black text on a white background.
- Once the image get binarized, all the effects such as contrast, sharpness etc. can be handled.
- In the segmentation process, each character where the sum of black pixels is zero, the edges get calculated along the periphery of the character.
- Each character is normalized in terms of size and focus to resemble the templates (used for training).
- ANN, use its basic functioning to extract features and generate vectors to give highly precise results.

Artificial neural network is a powerful data modelling tool that is able to capture and represent complex input/output relationships. In the field of ANN, Back-propagation algorithm can be used for training and classification purposes. ANN is high noise tolerant recognition system.

Neural network technology stemmed to develop an artificial system that could perform intelligent task despite the computational complexity involved. ANN offers several advantages in the sense of emulating adaptive intelligent systems to some extent.

B. Character Recognition Using Hidden Markov Model

Mixed cursive is the most general and difficult type handwriting style, and in view of its automatic recognition HMM is used. HMM is a doubly stochastic process that is not observable under some observations but can be processed due to its stochastic approach. It consists of a set of states connected to each other by transition with probabilities while the observed process consists of a set of operations [12].

HMM calculates the hidden states chain which has their basis on the observation chain of Counter & Viterbi algorithm. Both the algorithms have the most likely result and work in a specific way.
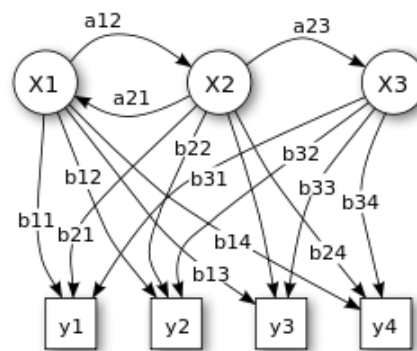

Fig 3: Hidden Markov Model

Training

- First the data is processed in a way so that it can be used for training.
- By extracting relevant data the size of content or data is reduced.
- It is essential to normalize the features & obtain the specific information in the form of vector of scalar values.
- Statistical estimation of feature vectors is essential for each class so the training process could have been done properly.

Testing

- In the testing module, pre-processing of the data is needed.
- The process of feature extraction has to be done so the results can be analysed.
- Finally the feature vectors of the various models get compared and matched up to the closest match and the process is done.

Application of counter algorithm on various written and printed character, generates feature vectors which gets first trained and then used to compare with the feature vector of the input during testing or recognition phase.

It also works with the sub algorithms but not detect the results when the character gets coloured.

So for this first the input should have to be converted in gray and then the output will be given by the process.

So, Hidden Markov Model uses the counter algorithm to analyse and recognise the characters (sometimes it uses the hidden chain mechanism), and finally match them with the closely related input processes.

C. Character Recognition Using Support Vector Machine

Support vector machine [13] is one of the top ranked algorithms in different domains of knowledge such as text categorization, digital image analysis, character recognition and bioinformatics. SVM is a classifier that performs classification tasks by constructing a hyper-plane or a set of hyper-planes in a high dimensional space. A Support Vector Machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyse data and recognize patterns, used for classification and regression analysis.

- After the pre-processing (i.e. conversion of the image into gray scale image, filtration, noise removal etc.), a clean image is available that goes for segmentation phase.
- Segmentation includes line segmentation, word segmentation and character segmentation.
- For ease of implementation and good recognition, a good set of features is extracted.
- SVM requires a training dataset and a testing dataset.SVM is a classifier that performs classification tasks by constructing a hyper-plane or a set of hyper-planes in a high dimensional space.

SVM approach has some advantages compared to other classification techniques. SVM technique shows greater ability to generate good classifiers.. Even when number of training samples is small, the result of SVM technique is robust, accurate and effective.

IV. CONCLUSIONS

Various techniques of offline handwriting recognition, support vector machine, artificial neural networks and Hidden Markov Model are implemented in HCR and OCR. These techniques are analysed and comprised to improve the system's usability and enhance the character recognition's framework. This material serves as a guide and update for readers working in the Character Recognition area.

REFERENCES

[1] Beigi, H.S.M., "An overview of handwriting recognition", Proc. The 1st Annual Conference on Technological Advancements in Developing countries, 1993. Columbia University. p. 30-46.

[2] Er. Neetu Bhatia, "Optical Character Recognition Techniques: A Review", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014

[3] Ranyang Li, Hang Wang, KaifanJi, "Feature extraction and identification of handwritten characters", 8th International Conference on Intelligent Networks and Intelligent Systems, 2015

[4] N. Venkata Rao, Dr.A.S.C.S.Sastary, A.S.N.Chakravarthy, KalyanChakravarthy P, "Optical Character Recognition Technique Algorithms", Journal of Theoretical and Applied Information Technology, Vol.83. No.2, 20th January 2016

[5] V. S. Dhaka, Manoj Kumar, Harshit Sharma, "Character Recognition of Offline Handwritten English Scripts: A Review", National Conference on Cloud Computing & Big Data

[6] Mr. Awasare Bhushan, Ms. Patil Kalyani, Ms. Jolapure Supriya, "Handwritten Script Recognition", IOSR Journal of Computer Engineering (IOSR-JCE)

[7] Sameeksha Barve, "Optical Character Recognition Using Artificial Neural Network", International Journal of Advanced Research in Computer Engineering & Technology, Volume 1, Issue 4, June 2012

[8] Te´ofilo E. de Campos, Bodla Rakesh Babu, Manik Varma, "Character Recognition In Natural Images"

[9] VivekShrivastava and Navdeep Sharma, "Artificial Neural Network Based Optical Character Recognition", Signal & Image Processing : An International Journal (SIPIJ) Vol.3, No.5, October 2012

[10] ElieKrevat, Elliot Cuzzillo, "Improving Off-line Handwritten Character Recognition with Hidden Markov Models"

[11] Eugen-DumitruTautu And Florin Leon, "Optical Character Recognition System Using Support Vector Machines", Gheorghe Asachi Technical University of Iaşi, Faculty of Automatic Control and Computer Engineering

[12] KarishmaTyagi, VedantRastogi, "Implementation of Character Recognition using Hidden Markov Model", International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 2, February – 2014

[13] Antonio Carlos Gay Thomé, "SVM Classifiers – Concepts and Applications to Character Recognition", Federal University of Rio de Janeiro, Brasil