Offline Handwritten English Script Recognition: A Survey

V. S. Dhaka,
School of Engineering and Technology, Jaipur National University Jaipur, India
vijaypal.dhaka@gmail.com
Manoj Kumar,
School of Engineering and Technology, Jaipur National University Jaipur, India
manoj186@gmail.com
Prashant Chaudhary,
School of Engineering and Technology, Jaipur National University Jaipur, India
pcbcs41@gmail.com

----------ABSTRACT----------

Intensive research has been done on optical character recognition (OCR) and a large number of articles have been published on this topic during the last few decades. Many commercial OCR systems are now available in the market. The character recognition problem itself can be considered as a mostly solved. So here we are giving review of some of the methods for detection of characters/word easily with less error in retrieved text. This material serves as a guide and update for readers working in the Offline English Characters/Word Recognition area. First, the historical evolution of OCR systems and English Script Properties is presented. Then, the available Offline English Characters/Word Recognition techniques with their superiorities are reviewed. Finally, the current status of Offline English CR is discussed, and directions for future research are suggested. Moreover, the paper also contains a comprehensive bibliography of many selected papers appeared in reputed journals and conference proceedings as an aid for the researchers working in the field of Offline English CR.

Keywords— Handwritten Character Recognition, Image processing, Feature extraction, Feed forward Neural Network, Convolutional NN, SVM, LeNet-5, HMM, Hybrid HMM/ANN, Projection Based Notch-Elimination, Artificial NN, Stroke length, Contour directional angle, Stochastic Context-Free Grammars (SCFG), Lexicon-driven and Segmentation, Statistical Model.

I. INTRODUCTION

OFFLINE handwriting recognition is the task of determining what letters or words are present in a digital image of handwritten text. It is of significant benefit to man-machine communication and can assist in the automatic processing of handwritten documents. It is a subtask of Optical Character Recognition (OCR), whose domain can be machine-print or handwriting but is more commonly machine-print. The recognition of English handwriting presents unique challenges and benefits and has been approached more recently than the recognition of text in other scripts. This paper describes the state of the art of this field. Handwritten recognition is usually classified into two groups which are online and offline. Online character recognition deals with information about writing dynamics as the text is being written while offline character recognition deals with static Information in which acquisition is done after all the text is written. Offline character recognition usually uses other medium of written text such as papers. One of the main issues in handwritten text recognition is that its accuracy, in human depends on knowledge about which language the text is written. Same text of the same corpus but written in different language can result in different accuracy [1]. It is "offline" if it is applied to previously written text, such as any images scanned in by a scanner. The online problem is usually easier than the offline problem since more information is available. This survey is restricted to offline systems. Standard database is also important for handwriting recognition research. The database is used in the development, evaluation, and comparison of different handwriting recognition algorithms [2]. Many databases have been developed in the character recognition community ranging from printed, handwritten, isolated or cursive, and also in various scripts. Some of them are publicly available such as IAM [2-3] and IRONOFF [4].

In this paper, we present a review of the offline handwritten English Recognition (OHER) work done on English language scripts. The review is organized into VI sections. Sections I cover introduction and section II or III cover properties on OCR or English scripts. In Section IV and V, we discuss different methodologies and performances in OCR development as well as research work done on English characters and word recognition. In Section VI, we discuss the scope of future work and conclude the paper.

II. PROPERTIES OF ENGLISH OCR

Machine replication of human functions, like reading, is an ancient dream. However, over the last five decades, machine reading has grown from a dream to reality. Optical character recognition has become one of the most successful applications of technology in the field of character recognition and artificial intelligence. Optical character recognition is the past when in 1929 Gustav

Tauschek got a patent on OCR in Germany followed by Handel who obtained a US Patent on OCR in USA in 1933. Since then number of character recognition systems have been developed and are in use for even commercial purposes also. But still there is a hope to build some more intelligent hand written character recognition system because hand writing differs from one person to other. His writing style, shape of alphabets and their sizes makes the difference and complexity to recognize the characters [5].

Depending on versatility, robustness and efficiency, the commercial OCR systems can be divided into four generations. The first generation systems can be characterized by the constrained letter shapes which the OCRs read. Such machines appeared in the beginning of the 1960s. The first widely commercialized OCR of this generation was the IBM 1418, which was designed to read a special IBM font, 407 [6]. The recognition method was logical template matching where the positional relationship was fully utilized.

The next generation is characterized by the recognition capabilities of a set of regular machine printed characters as well as hand-printed characters. At the early stages, the scope was restricted to numerals only. Such machines appeared in the middle of 1960s to early 1970s. In this generation, the first and famous OCR system was IBM 1287, which was exhibited at the 1965 New York world fair [6]. In terms of hardware configuration, the system was a hybrid one, combining analog and digital technology. The first automatic letter-sorting machine for postal code numbers of Toshiba was also developed during this period. The methods were based on the structural analysis approach.

The third generation can be characterized by the OCR of poor print quality characters, and hand –printed characters for a large category character set. Commercial OCR systems with such capabilities appeared roughly during the decade 1975 to 1985 [6–8].

The fourth generation can be characterized by the OCR of complex documents intermixing with text, graphics, table and mathematical symbols, unconstrained hand written characters, color document, low-quality noisy documents like photocopy and fax, etc. Some pieces of work on complex documents provided good results. Although many pieces of work on unconstrained hand written character are available in the literature, the recognition accuracy hardly exceeds 85%. Very few studies on color documents have been published and research on this problem is continuing. Also, research on noisy document is in progress [9, 10].

Writing may be classified as culture specific artifact. Even when using the same language, the motor-behavior of how the text is taught and learned at early school could be different for different people. The study investigates the direction of the CR research, analyzing the limitations of methodologies for the systems, which can be classified based upon two major criteria: 1) the data acquisition process (on-line or off-line) and 2) the text type (machine-printed or handwritten) [11]. No matter in which class the problem belongs, in general, there are five major stages in the CR problem in Offline Handwritten English Recognition [13].1) Preprocessing; 2) Segmentation; 3) Feature Extraction; 4) Classification; 5) Post processing.

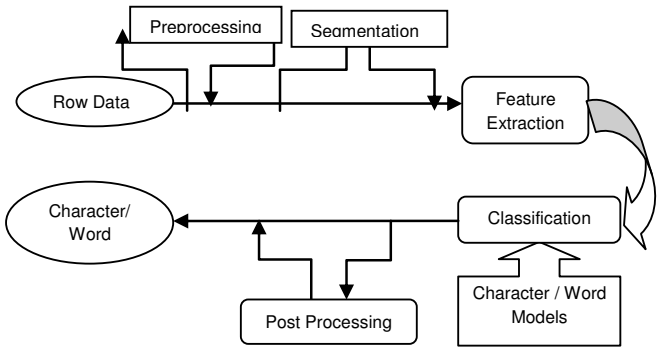

Fig. 1. OCR Working Process in English Recognition

Preprocessing: The preprocessing stage is a collection of operations that apply successive transformations on an image. It takes in a raw image and enhances it by reducing noise and distortion, and hence simplifies segmentation, feature extraction, and consequently recognition. The quality of input text depends on many factors.

Document history: A document that has been faxed or copied several times is harder to read than the original. Text gets thinner or thicker, salt and pepper noise appears, and contrast diminishes.

Printing process: A typeset document is clearer than a typewritten one, which in turn is clearer than the output of a dot-matrix printer. Other deformations that relate to the printing process include ink spreading, and ink chipping.

Font clarity: Exotic fonts, small font sizes, italic and bold characters, subscripts and superscripts, and using multiple font sizes (e.g., drop caps) and styles complicate recognition.

Paper quality: Opaque, heavy-weight, smooth, uniform grain paper is easier to read than lightweight, transparent paper (e.g., newspapers).

Document condition: The presence of extraneous markings and stains make reading harder.

Image *acquisition:* The digitization of on-line script is limited by tablet resolution and sampling rate, and often introduces distortions like small zigzags. The quality of scanned text is compromised by positioning variations (skew, translations, stretching, etc.), defocusing, unclean document glass, and the limited resolution.

Once text is acquired, either on-line or off-line, it should be preprocessed to simplify recognition. Preprocessing operations are usually specialized image processing

operations that transform the image into another with reduced noise and variation.

Those operations include binarization, filtering and smoothing, thinning, alignment, normalization, and baseline detection. Ideally, preprocessing should remove all variations and detail from a text image that are meaningless to the recognition method. As that goal are still illusive, preprocessing attempts to reduce noise and data variations as much as possible.

Segmentation: The *segmentation* stage takes in a page image and separates the different logical parts, like text from graphics, lines of a paragraph, and characters (or parts thereof) of a word. After the preprocessing stage, most OCR systems isolate the individual characters or strokes before recognizing them. Segmenting a page of text can be broken down into two levels: page decomposition and word segmentation, When working with pages that contain different object types like graphics, headings, mathematical formulas, and text blocks, page decomposition separates the different page elements, producing text blocks, lines, and sub-words. While page decomposition might identify sets of logical components of a page, word segmentation separates the characters of a sub-word.

Feature Extraction: The *feature extraction* stage analyzes a text segment and selects a set of features that can be used to uniquely identify the text segment. These features are extracted and passed in a form suitable for the recognition phase.

Once an OCR system has an isolated pattern (character or primitive), its next step is to extract the features of the pattern and pass them along to the classifier to classify it. Feature extraction is one of the most difficult and important problems of pattern recognition [14]. The selected set of features should be a small set whose values efficiently discriminate between patterns of different classes, but are similar for patterns within the same class. The feature extraction step is closely related to classification because the type of features extracted here must match what the classifier expects. The two main control approaches for feature extraction and classification are interleaved control versus one-step control. In interleaved control, an OCR system alternates between feature extraction and classification. In one such realization, the OCR system extracts a set of features from a pattern, and based on the feature values, classifies the pattern into a (small) number of categories. The system then extracts another set of features that are specific to each category and classifies the pattern [15, 16, and 17].

Classification: The *classification* stage is the main decision-making stage of an OCR system. The classification stage uses the features extracted in the previous stage to identify the text segment according to preset rules. This stage may use feature models obtained in an (off-line) training (modeling) phase to classify the test data. Classification in an OCR system is the main decision making stage in which the features extracted from a pattern are compared to those of the model set. Based on the features, classification attempts to identify the pattern as a member of a certain class. When classifying a pattern, classification often produces a set of hypothesized solutions instead of generating a unique solution. The (subsequent) post-processing stage uses higher level information to select the correct solution.

Historically, classification followed two main paradigms: syntactic (or structural) and statistical (or decision theoretic) classification. Recently, recognition using neural networks has provided a third paradigm.

Post Processing: The *post-processing* stage, which is the final stage, improves recognition by refining the decisions taken by the previous stage and recognizes words by using context. It is ultimately responsible for outputting the best solution and is often implemented as a set of techniques that rely on character frequencies, lexicons, and other context information.

The final stage in the recognition process is post-processing. One of the objectives of post-processing is to improve word recognition rate (as opposed to character recognition rate). Post-processing is often implemented as a set of techniques that rely on character frequencies, lexicons, and other contextual information. As classification, sometimes, produces a set of possible solutions instead of a unique solution, post-processing is responsible for selecting the right solution using higher level information that is not available to the classifier. Post-processing also uses that higher level information to check the correctness of the solutions returned by the classifier. The most common post-processing operations are spell checking and correction. Spell checking can be as simple as looking up words in a lexicon.

III. PROPERTIES OF ENGLISH SCRIPTS

Fig. 2. Samples of Handwritten English Characters A to Z

The modern English alphabet is a Latin alphabet consisting of 26 letters (each having an upper case and a lower case form).

Characters	Group
O, Q, Q and o,	Circle
A, L, T, V, Y, v, r, and y	Triangle
	Line
I, i, and l	Shape
	Rectangle
U, u, n, U	Ellipse
E, F, M, N, W, H, K, Z, z,	Diamond
m, w, S, s	Cone
X, x, t	Semi Circle
P, b, d, f, J, j, p and q	Other
D, G, e, B, C, c	
R, a, h, k	

Fig.3. Characters and their geometric shape groups

Fig. 4. Different zones Of English Text Line

In the English language, a text line may be partitioned into three zones. The upper-zone denotes the portion above the head-line, the middle-zone covers the portion of basic (and compound) characters below head-line and the lower-zone is the portion below base-line. Those text where script lines do not have head-line, the mean-line separates upper- and middle-zone, while the base-line separates middle- and lower-zone. An imaginary line, where most of the uppermost (lowermost) points of characters of a text line lie, is referred as mean-line (base-line). Examples of zoning are shown in Fig. 2. In this case, the head or mean-line along with base-line partition the text line into three zones [18].

IV. METHODS FOR OFFLINE HANDWRITTEN ENGLISH CHARACTERS RECOGINITION

Anshul Gupta, Manisha Srivastava and Chitralekha Mahanta [19] focuses on developing a CR system for recognition of handwritten English words. They first segment the words into individual characters and then represent these characters by features that have good discriminative abilities. They also explore different neural network classifiers to find the best classifier for the CR system. They combine different CR techniques in parallel so that recognition accuracy of the system can be improved. This Research work done is as follows: Firstly deals with segmentation of words into individual characters where a heuristic algorithm is used to first over segment the word followed by verification using neural network. In this

A. Segmentation using a heuristic algorithm,
B. Manual marking of segmentation points
C. Training of the Artificial Neural Network (ANN)

D. Testing phase of the segmentation technique

Feature extraction of handwritten characters is discussed in Secondly. Next Phase describes selection procedure of a suitable classifier. This is done by

A. Testing multilayer perceptron (MLP),
B. Radial basis function (RBF) and
C. Support vector machine (SVM)

and selecting the one that has the maximum accuracy. In next post processing is discussed where different character recognition techniques are combined in parallel by using a variation of the Borda count.

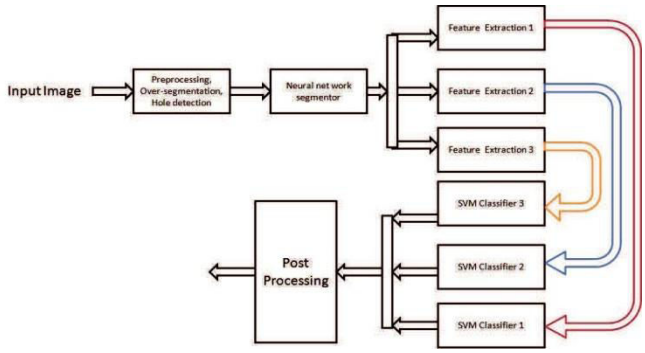

Fig. 5. Block diagram of the proposed CR system

This Research carries out a study of various feature based classification techniques for offline handwritten character recognition. After experimentation, it proposes an optimal character recognition technique. The proposed method involves segmentation of a handwritten word by using heuristics and artificial intelligence. Three combinations of Fourier descriptors are used in parallel as feature vectors. Support vector machine is used as the classifier. Post processing is carried out by employing lexicon to verify the validity of the predicted word. The results obtained by using the proposed CR system are found to be satisfactory.

This research applies Convolutional Neural Networks (CNNs) for offline handwritten English character recognition. *Aiquan Yuan, Gang Bai, Lijing Jiao, Yajie Liu* [20] use a modified *LeNet-5* CNN model, with special settings of the number of neurons in each layer and the connecting way between some layers. Outputs of the CNN are set with error-correcting codes, thus the CNN has the ability to reject recognition results. For training of the CNN, an error-samples-based reinforcement learning strategy is developed.

In 1995, Convolutional Neural Networks (CNNs) was brought about by LeCun and caused huge attention immediately [21]. In a CNN recognition system, 2-D image can be directly input and feature extraction is thus avoided. Many experiments with the CNN have seen moderately good performance. This Research focuses mainly on offline HECR on UNIPE dataset [22], with 26 characters for uppercase and lowercase, respectively.

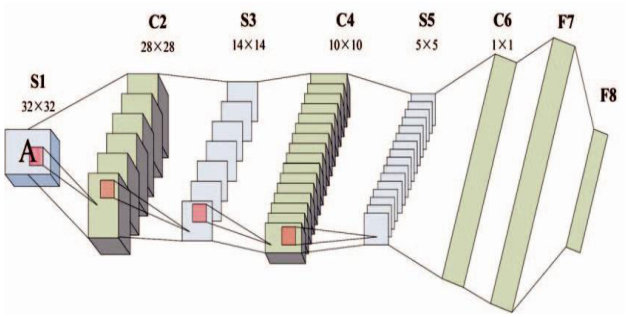

Fig.6. the basic architecture of *LeNet-5*

A common model of CNNs is the *LeNet-5* model [23], as shown in Figure 6. Each unit in the *LeNet-5* model is connected to a local neighborhood in the previous layer, thus it can be seen as a local feature detector. Insensitivity to local transformations is built into the network architecture and the same features on different parts of the input are detected.

This research shows the solution for HECR with CNNs. The output of the CNN are set with EC codes, thus the CNN has the ability for rejection in recognition. Comparing with other state-of-the-art methods, the CNN has provided an encouraging solution for offline HECR.

In this research, *J.Pradeep, E.Srinivasan and S.Himavathi* [24] used a hybrid feature extraction scheme by combining two zonal based approaches, namely, diagonal and directional is proposed. The features obtained are used to train neural network based classifiers such as feed forward network and radial basis function network. A comparison is carried out with the nearest neighbor classifier. The hybrid features are used to train to Neural Network based classifiers and the results obtained are presented. The best recognition system is identified and the experimental results are presented and discussed.

A Hybrid feature extraction based off-line handwritten character recognition system with different classifiers namely, Feed forward NN, radial basis function NN and nearest neighbour network for recognizing handwritten English alphabets is proposed. A hybrid feature extraction technique, combining two different approaches namely, diagonal based feature extraction and directional based feature extraction is used. The different classifiers have been trained with 200 sets of 26 alphabets and tested extensively. Experimental results show that the feed forward neural network is distinctly superior to the other classifiers in recognizing the handwritten English alphabets.

The aim of this research is to recognize upper-case English alphabets from handwritten documents. An OCR system based on four-sided projections of an alphabet is proposed. The projection points are approximated by polygons and feature points are extracted subsequently for notch elimination and segmentation of the polygon. The resulting segments are smoothed by Bézier approximation. Statistical matching technique is successfully applied for character recognition. In this research *Sarbajit Pal, Jhimli Mitra and Paromita Banerjee* [25] propose an offline handwritten character recognition method for 26 upper-case English alphabets. A character can be best characterized by the scans obtained from four sides. But, different orientations of handwritten characters can yield various scan results. Thus, the proposed system has been developed to accommodate the diversities in handwritten characters. A model dataset is created resembling the Canadian standard [26], where the characters are handwritten with ideal orientation. A matching technique has been developed with maximum level of accuracy to match handwritten characters with variations.

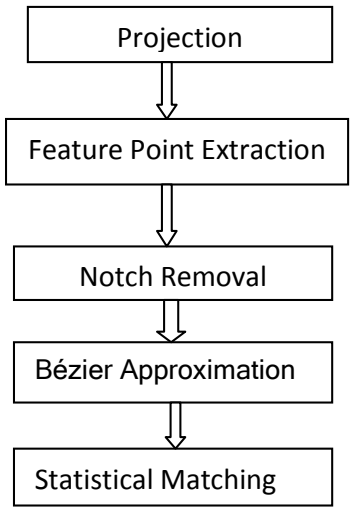

Fig. 7. Flowchart of the proposed approach

It involves four side projections of the characters. Then a simplistic feature point based notch-elimination technique is applied. However, the threshold level used in the notch-elimination technique is a heuristic approach, which can be modified using a dynamic threshold value. The feature extraction algorithm is applied on the resulting set of curves [27]. The feature points are first polygonally approximated, and then a Bézier approximation using a fast algebraic method is applied [28]. The Bézier curves are used in the matching technique, which is a simple statistical method based on correlation factors. A correlation factor of 0.5 to 1 represents a good match and below 0.5 as a poor match. The uniform sub-sampling of number of sample points to a fixed number of points over the curve does not deteriorate the match result. However, a preferential sub sampling may be applied where more number of sample points could be taken around the change of curvature of the curve.

In this research, *Rajib Lochan Das, Binod Kumar Prasad and Goutam Sanyal* [29] propose a recognition model based on multiple Hidden Markov Models (HMMs) followed by few novel feature extraction techniques for a single character to tackle its different writing formats. They also propose a post-processing block at the final stage to enhance the recognition rate further.

The tool to train the system with the obtained feature vectors is taken to be HMM because OHR systems based on HMM have been shown to outperform segmentation based approaches [30]-[32]. With the usage of HMM models for the pattern recognition or character recognition, a HMM model keeps information for a character when the model is trained properly and the trained model can be used to recognize an unknown character. The advantage with HMM based systems is that they are segmentation free that is no pre- segmentation of word/line images into small units such as sub-words or characters is required [31]. On the other hand, HMM based approaches have been found to possess some limitations also. These limitations are due to two reasons-(a) the assumptions of conditional independence of the observations given the state sequence and (b) the restriction on feature extraction imposed by frame based observations [33].

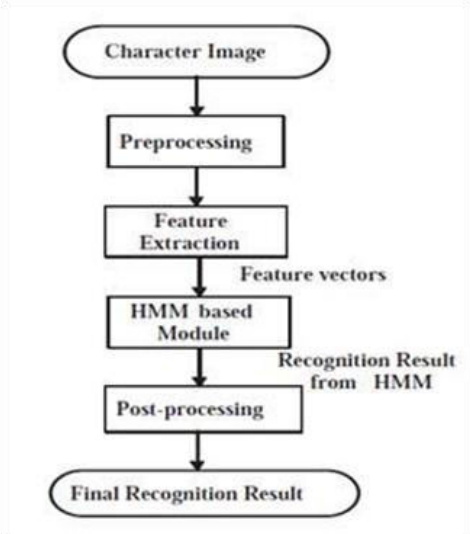

Fig.8. System Overview of proposed approach

In this research, an approach has been made to increase the rate of recognition of handwritten character by finding both local and global features. Multiple level HMM model is designed for some specific letters having wide range of variations from writer to writer. In the last section, a trial has been made to put a line of demarcation between similar looking characters.

All this specialty of this research has made us obtain an average accuracy of 98.26%. For a significant number of letters, the accuracy rate is even close to 100%.

Rakesh Kumar Mandal and N R Manna [34] give a concept of recognizing hand written character pattern has been developed and implemented called Row-wise segmentation technique. RST helps in minimizing errors in pattern recognition due to different handwriting styles to great extent. In this method input pattern matrix is segmented row-wise into different groups. Target pattern is also grouped where each group is the numeric equivalent of the chronological position of each English alphabet. Each input segment is fully interconnected with each target group. Number of target groups is equal to the number of rows in the input matrix.

In general, the overall program has been divided into two parts, training and testing. Training requires the net to read segmented input patterns and testing requires the net to read any test character pattern, to read the produced target samples and to count the majority of samples and to find out the numeric equivalent of the sample to identify the character. RST gives very good accuracy, if the characters were written in boxed sheets. In this research, the method applied used the logic of encaging the characters without using the boxed sheets but the logic provides static encaging. Problems in identifying the characters arises when the characters gets fully deviated from their positions on the sheet. Efficient algorithm is still to be explored to encage the characters written on any position of the paper. Generalizations among variations in sizes of the characters in the box also produces problem.

Supriya Deshmukh, Leena Ragha [36] proposes a method on offline isolated English character. The method is also applied to Marathi vowels. The image acquired is preprocessed to remove all unwanted details from the image so that the image is suitable for feature extraction. Feature extraction plays an important role in handwritten recognition. The two feature extraction methods based on directional features are considered. The first method uses stroke distribution of a character. The second method uses contour extraction. The Two directional features are compared with two different correlation techniques separately to check the suitability of the recognition method. First correlation technique calculates the dissimilarity between reference pattern and test pattern, and the other calculates the similarity between reference pattern and test pattern. The result of the comparison is to classify the character under consideration to a class if hit. If miss, the confusion information is extracted for the analysis. They observed that the Stroke length method give good performance on a character that has straight lines while contour method behaves well on a character curves.

Hiromitsu N. and Takehiko T. [37] research examines effective recognition techniques for deformed characters, extending conventional recognition techniques using on-line character writing information containing writing pressure data. That study extends conventional recognition techniques using on-line character writing information containing writing pressure information. A recognition system using simple pattern matching and HMM was made for evaluation experiments using Common Hand printed English character patterns from the ETL6 database to determine effectiveness of the proposed extending recognition method. Character recognition performance is increased in both expansion recognition methods using on-line writing information. On-line character writing information comprises vector patterns of pen movement at writing scene. Offline character patterns are merely pixel patterns that include no vector information. Although some off-line character recognition systems use some correlation of stroke information [38], all proposed off-line character recognition methods use only off-line character pattern information. However, an effective system for improving recognition performance could use some on-line writing information for off-line recognition.

TABLE: 1I CHARACTERS RECOGNITION TECHNIQUES

Sr	Authors	Method	Classifier	Data Set	Accuracy (%)
1.	Anshul Gupta, Manisha Srivastava and Chitralekha Mahanta [19]	Support Vector Machine (SVM)	Neural Network	74K	98.86
2.	Aiquan Yuan, Gang Bai, Lijing Jiao and Yajie Liu [20]	LeNet-5	Convolutional Neural Network	61K	90.20
3.	J.Pradeep, E.Srinivasan and S.Himavathi [24]	Hybrid-Feature Extraction	Feed-Forward Neural Network	2494	95.96
4.	Sarbajit Pal, Jhimli Mitra and Paromita Banerjee [25]	Statistical Matching	Projection Based Notch-Elimination and segmentation	Conceptual	99
5.	Rajib Lochan Das, Binod Kumar Prasad and Goutam Sanyal [29]	Local and Global Feature Extraction	HMM	13K	98.26
6.	Rakesh Kumar Mandal and N R Manna [34]	Row-Wise Segmentation (RST)	Single Layer ANN	Conceptual	~99
7.	Supriya Deshmukh, Leena Ragha [36]	Stroke length	Dissimilarity correlation	100 sample	88.04
			Similarity correlation	100 sample	94.27
		Contour directional angle	Dissimilarity correlation	100 sample	87.54
			Similarity correlation	100 sample	93.69
8.	Hiromitsu NISHIMURA and Takehiko TIMIKAWA [37]	Conventional recognition in online data	Pattern matching and HMM	600 sample	96.4 (1D)
					99.0 (MD)

V. METHODS FOR OFFLINE HANDWRITTEN ENGLISH WORDS RECOGINITION

In this research, *Oliever C., Avila M., Courtellemont P., Paquet T. Lecourtier Y.* [40] propose a method for the recognition of handwritten literal amount on various bank checks. They present the pre-processing of the original 256 gray-levels image, containing in homogeneous background, and the locating of the handwritten information. For recognition of the amount, they choose a Markovian approach, which is first applied to the sequences of words. The first results allow considering an extension of the method to the sequences of graphemes' in words in order to improve the recognition rate. The data base used to test the algorithm is made of about l000 check amounts, which correspond to about 3000 words. This base is formed with elements stemming from the *Service de Recherche Technique de la Poste (SRTP)* and the *Matra CAP Systems society* in France.

Alessandro Vinclarelli, Samy Bengio, Horst Bunke [39] presents a system for the offline recognition of large vocabulary unconstrained handwritten texts. The only assumption made about the data is that it is written in English. This allows the application of Statistical Language Models in order to improve the performance of our system. Several experiments have been performed using both single and multiple writer data. Lexica of variable size (from 10,000 to 50,000 words) have been used. The use of language models is shown to improve the accuracy of the system (when the lexicon contains 50,000 words, the error rate is reduced by \sim 50 percent for single writer data and by\sim 25 percent for multiple writer data). An experimental setup to correctly deal with unconstrained text recognition is proposed.

Matthias Zimmermann, Jean-Ce´dric Chappelier, and Horst Bunke [41] proposes a sequential coupling of a Hidden Markov Model (HMM) recognizer for offline handwritten English sentences with a probabilistic bottom-up chart parser using Stochastic Context-Free Grammars (SCFG) extracted from a text corpus. Based on extensive experiments, they conclude that syntax analysis helps to improve recognition rates significantly. In experimental Setup and System Optimization, Two different recognition tasks are defined. In the Multiwriter Task (MWT) the recognizer is trained on handwritten texts from a large set of known writers. For the Writer Independent Task (WIT), writing styles are not known in advance, i.e., the writers represented in the training set are not represented in either the validation or the test set of this task. This research was partly supported by the Swiss National Science Foundation NCCR program "Interactive Multimodal Information Management" (IM2) in the individual Project "Scene Analysis."

Anne-Laure Bianee-Bernard, Fare's Menasri, Rami Al-Hajj Mohamad, Chafic Mokbel, Christopher Kermorvant, and Laurence Likforman-Sulem [42] proposed a research which aims at building an efficient word recognition system resulting from the combination of three handwriting recognizers. The main component of this combined system is an HMM based recognizer which considers dynamic and contextual information for a better modeling of writing

units. For modeling the contextual units, a state-tying process based on decision tree clustering is introduced. Decision trees are built according to a set of expert-based questions on how characters are written. Questions are divided into global questions, yielding larger clusters, and precise questions, yielding smaller ones. Such clustering enables us to reduce the total number of models and Gaussians densities by 10. We then apply this modeling to the recognition of handwritten words. They introduced a novel approach to build efficient context dependent word models based on the HMM framework. The key features of such approach are the use of dynamic features and state-based clustering.

Salvador E.B., M.J. Castro-Bleda, Jorge Gorbe-Moya, and Francisco Z.-M.,[43] proposed the use of hybrid Hidden Markov Model (HMM)/Artificial Neural Network (ANN) models for recognizing unconstrained offline handwritten texts. The structural part of the optical models has been modeled with Markov chains, and a Multilayer Perceptron is used to estimate the emission probabilities. That research also presents new techniques to remove slope and slant from handwritten text and to normalize the size of text images with supervised learning methods. Slope correction and size normalization are achieved by classifying local extrema of text contours with Multilayer Perceptrons. Slant is also removed in a nonuniform way by using Artificial Neural Networks. Experiments have been conducted on offline handwritten text lines from the IAM database, and the recognition rates achieved, in comparison to the ones reported in the literature, are among the best for the same task.

The key features of the recognition system are the novel approach to preprocessing and recognition, which are both based on ANNs. The preprocessing is based on using MLPs:

- to clean and enhance the images,
- to automatically classify local extrema in order to correct the slope and to normalize the size of the text lines images, and
- to perform a nonuniform slant correction.

The recognition is based on hybrid optical HMM/ANN models, where an MLP is used to estimate the emission probabilities.

Aiquan Yuan, Gang Bai, Po yang, yanni Guo, Xinting Zgao [44] presents a novel segmentation-based and lexicon-driven handwritten English recognition systems based. For the segmentation, a modified online segmentation method based rules are applied. Then, convolutional neural networks are introduced for offline character recognition. Experiments are evaluated on UNIPEN lowercase data sets, with the word recognition rate of 92.20%. That word recognition system is segmentation dependent, exploring segmentation methods with better performances is considerately critical.

Sr.	Authors	Method	Classifier		Dataset (Word)	Accuracy (%)
1.	*OLIVER C., AVILA M., COURTELLEMONT P., PAQUET T., LECOURTIER Y.* [40]	Image Segmentation	HMM		3000	52.25
2.	*ALESSANDRO VINCIARELLI, SAMY BENGIO, HORST BUNKE* [39]	Language Resources	HMM and Statistical Model	Multi Writers Task	50000	75.00
				Single Writer Task	50000	50.00
3.	*MATTHIAS Z., JEAN-C. CHAPPELIER, HORST BUNKE* [41]	Stochastic Context-Free Grammars (SCFG)	HMM	Multi Writers Task	2000	75.60
				Single Writer Task	2000	54.40
4.	*Anne-Laure Bianne-Bernard, Fare`s Menasri, Rami Al-Hajj Mohamad, Chafic Mokbel, Christopher Kermorvant, and Laurence Likforman-Sulem* [42]	Rimes feature extraction	HMM		2130	74.1-78.0
		Sliding Window System	HMM	Context Independent	5334	68.57
				Context Independent	5334	75.53
		Rimes feature extraction and Sliding Window	NN and HMM		7742	89.10
5.	*Salvador E.B., M. J. Castro-Bleda, Jorge Gorbe-Moya, and Francisco Z.-M.* [43]	MLP Based	Hybrid HMM/ANN Model		30000	84.39
6.	*Aiquan Yuan, Gang Bai, Po Yang, Yanni Guo, Xinting Zhao* [44]	Lexicon-driven and Segmentation Based	CNN		1791	92.20

TABLE: II WORDS RECOGNITION TECHNIQUES

TABLE: I OFFLI

IV. CONCLUSION AND FUTURE WORK

In this paper, we presented a review of OCR work done on English language scripts. Here, at First, we briefly discussed Different methodologies applied in OCR development in international scenario and then different work done for English language scripts recognition. Finally, we discussed Offline Handwritten English Characters and Words recognition Techniques performance needed for better English script OCR development. We believe that our survey will strongly encourage activities of automatic handwritten document processing and OCR of English language scripts.

From the survey, it is noted that the errors in recognizing handwritten English characters are mainly due to incorrect character segmentation of touching or broken characters. Because of upper and lower modifiers of English text, many portions of two consecutive lines may also overlap and proper segmentation of such overlapped portions are needed to get higher accuracy. Many authors suggest that the post processing of classifier outputs by integrating a dictionary with the OCR system can significantly reduce the misclassifications in printed as well as handwritten word recognitions.

REFERENCES

[1] M. Liwicki, H. Bunke, *Recognition of Whiteboard Notes online, offline and combination*, World Scientific Publishing, 2008.

[2] U. V. Marti and H. Bunke, *A Full English sentence database for offline handwriting recognition*, In Proc. of the 5th Int. Conf. on Document Analysis and Recognition, pages 705 - 708, 1999.

[3] U. Marti and H. Bunke. *The IAM-database: An English Sentence Database for Off-line Handwriting Recognition*. Int. Journal on Document Analysis and Recognition, Volume 5, pages 39 - 46, 2002.

[4] C. Viard-Gaudin, P. M. Lallican, S. Knerr, and P. Binter, *The IRESTE On/Off (IRONOFF) Dual Handwriting Database*, Proc. Intl. Conference of Document Analysis and Recognition (ICDAR), Bangalore, India, 1999.

[5] Ashish Chaturvedi and Yusuf Perwej, *Neural Networks for Handwritten English Alphabet Recognition,* International Journal of Computer Applications (0975 – 8887) Volume 20– No.7, April 2011.

[6] S. Mori, C.Y. Suen, K. Yamamoto, "Historical review of OCR research and development", pp1029–1058, Proc. IEEE 80 (1992).

[7] S. Mori, K. Yamamoto, M. Yasuda, Research on machine recognition of hand-printed characters, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1984) 386–405.

[8] G. Nagy, At the frontiers of OCR, Proc. IEEE 80 (7) (1992) 1093–1100.

[9] R. Plamondon, S.N. Srihari, On-line and o8-line handwritten recognition: a comprehensive survey, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 62–84.

[10] S.N. Srihari, J.J. Hull, in: S.C. Shariro (Ed.), Character Recognition, Encyclopedia of ArtiHcial Intelligence, Wiley, New York, 1992, pp. 138–150.

[11] Nafiz Arica and Fatos T. Yarman-Vural "*An Overview of Character Recognition Focused on Off-Line Handwriting"* IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS —PART C: APPLICATIONS AND REVIEWS, VOL. 31, NO. 2, MAY 2001.

[12] Ming Ma, Dong-Won Park, Soo Kyun Kim and Syungog An, "*Online Recognition of Handwritten Korean and English Characters*", Journal of Information Processing Systems, Volume 8, No 4, December 2012.

[13] Badr Al-Badr, Sabri A. Mahmoudb, "Survey and bibliography of Arabic optical text recognition", Signal Processing, Volume 41, Page no.49-77, 1995.

[14] S. Impedovo, L. Ottaviano and S. Occhinegro, "Optical character recognition - A survey", *Internat. J. Pattern Recognition and Artificial Intekgence,* Vol. 5, No. 1, pp. l-24, 1991.

[15] S.S. Hyder and A. Koujah, "Character recognition of cursive scripts", *Proc. 1st Internat. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems IEA/AIE - 88,* Tullahoma, TN, pp. 1146-1150, June 1988.

[16] V.P. Conception, M.P. Grzech and D.P. D'Amato, "Using morphology in document image processing", *Visual Communications and Image Processing '91: Image Processing,* Vol. SPIE-1606, pp. 132-140, 1991.

[17] M. Maier, "Separating characters in scripted documents", *Proc. 8th Internat. Joint Conf. on Pattern Recognition,* Paris, France, pp. 1056-1058, October 1986.

[18] U. Pal, B.B. Chaudhuri, Indian script character recognition: a survey, www.elsevier.com, Pattern Recognition 37 (2004) 1887 – 1899.

[19] Anshul Gupta, Manisha Srivastava and Chitralekha Mahanta"*Offline Handwritten Character Recognition Using Neural Network*" 2011 International Conference on Computer Applications and Industrial Electronics (ICCAIE 2011).

[20] Aiquan Yuan, Gang Bai, Lijing Jiao, Yajie Liu " *Offline Handwritten English Character Recognition based on Convolutional Neural Network*" 10th IAPR International Workshop on Document Analysis Systems 2012.

[21] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361, 1995.

[22] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet. Unipen project of on-line data exchange and recognizer benchmarks. In *Pattern Recognition 1994. Vol. 2-Conference*

B: Computer Vision & Image Processing, Proceedings of the 12th IAPR International. Conference on, volume 2, pages 29–33. IEEE, 1994.

[23] D. Bouchain. Character recognition using convolution neural networks. *Institute for Neural Information Processing*, 2007, 2006.

[24] J.Pradeep, E.Srinivasan and S.Himavathi "*Performance Analysis of Hybrid Feature Extraction Technique for Recognizing English Handwritten Characters***"** 978-1-4673-4805-8/12/ 2012 IEEE.

[25] Sarbajit Pal, Jhimli Mitra, Soumya Ghose and Paromita Banerjee "*A Projection Based Statistical Approach for Handwritten Character Recognition"* International Conference on Computational Intelligence and Multimedia Applications 2007.

[26] Suen, C. Y., et al., 1983. Canadian standard alphanumeric character set for hand printing, Z243.34-M1983, Canadian Standards Association, and Toronto.

[27] Pal, S., Mitra, J., Ghose, S. (2006). Segmentation of Noisy Digital curve for Approximation with Bézier. IEEE Inc. BVBCET. 1st Intl. Conf . on Signal and Image Processing, Hubli, India. Part-I, 484-488.

[28] Gangly, P., Ganguly, C. And Pal, S. (2006). A new approach for constructing Digital Curves. Intl. Conf. on Systematics, Cybernetics and Informatics Proc. 440-443.

[29] Rajib Lochan Das, Binod Kumar Prasad and Goutam Sanyal *"HMM based Offline Handwritten Writer Independent English Character Recognition using Global and Local Feature Extraction"* International Journal of Computer Applications (0975 – 8887) Volume 46– No.10, May 2012.

[30] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition", Proceedings of The IEEE, vol. 77, no. 2, pp. 257-286, Feb. 1998.

[31] C. Mokbel, H. Abi Akl, and H. Greige, "Automatic speech recognition of Arabic digits over Telephone network" ,Proc.Research Trends in Science and Technology, 2002.

[32] R. El-Hajj, L. Likforman-Sulem, and C. Mokbel, "Arabic Hand- writing Recognition Using Baseline Dependent Features and Hidden Markov Modeling, Proc. Eighth Intl Conf. Document Analysis and Recognition, pp. 893-897, 2005. [19] H. El Abed and V. Margner, "ICDAR 2009 -Arabic handwriting recognition competition", Inter. Journal on Document Analysis and Recognition, vol. 1433-2833, 2010.

[33] Zhang Hong lin, "Visiual C++Digital image pattern recognition technology and engineering practice," Beijing: Posts & Telecom Press, 2008,pp. 52-58.

[34] Rakesh Kumar Mandal, N R Manna "*Hand Written English Character Recognition using Row- wise Segmentation Technique (RST)"* International Symposium on Devices MEMS, Intelligent Systems & Communication (ISDMISC) 2011 Proceedings published by International Journal of Computer Applications® (IJCA).

[35] Nafiz Arica and Fatos T. Yarman-Vural *"An Overview of Character Recognition Focused on Off-Line Handwriting"* IEEE Transactions on systems, man, and cybernetics—part c: Applications and reviews, VOL. 31, NO. 2, MAY 2001.

[36] Supriya Deshmukh, Leena Ragha, "Analysis of Directional Features - Stroke and Contour for Handwritten Character Recognition", *2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009.*

[37] Hiromitsu NISHIMURA and Takehiko TIMIKAWA, "Off-line Character Recognition using On-line Character Writing Information", *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'2003).*

[38] M. Yasuda and H. Fujisawa, "An Improvement of Correlation Method for Character Recognition," IEICE J62-D, pp. 217-224, 1979.

[39] ALESSANDRO VINCIARELLI, SAMY BENGIO, HORST BUNKE. "*Offline Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Language Models"* IEEE Transactions on pattern Analysis and Machine Intelligence, Volume 26 Issue 6, June 2004 Page 709-720.

[40] OLIVER C., AVILA M., COURTELLEMONT P., PAQUET T., LECOURTIER Y. "*Handwritten Word Recognition by Image Segmentation and Hidden Markov Models"* IEEE 0-7803-0891-3/93.

[41] Matthias Zimmermann, Jean-Ce´dric Chappelier, and Horst Bunke. *"Offline Grammar-Based Recognition of Handwritten Sentences"* IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 28, NO. 5, MAY 2006.

[42] Anne-Laure Bianne-Bernard, Fare`s Menasri, Rami Al-Hajj Mohamad, Chafic Mokbel, Christopher Kermorvant, and Laurence Likforman-Sulem. "*Dynamic and Contextual Information in HMM Modeling for Handwritten Word Recognition"*, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 33, NO. 10, OCTOBER 2011.

[43] Salvador Espan˜a-Boquera, Maria Jose Castro-Bleda, Jorge Gorbe-Moya, and Francisco Zamora-Martinez. *"Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models",* IEEE TRANSACTIONS

ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 33, NO. 4, APRIL 2011.

[44] Aiquan Yuan, Gang Bai, Po Yang, Yanni Guo, Xinting Zhao. *"Handwritten English Word Recognition based on Convolutional Neural Networks"*, International Conference on Frontiers in Handwriting Recognition, 2012.