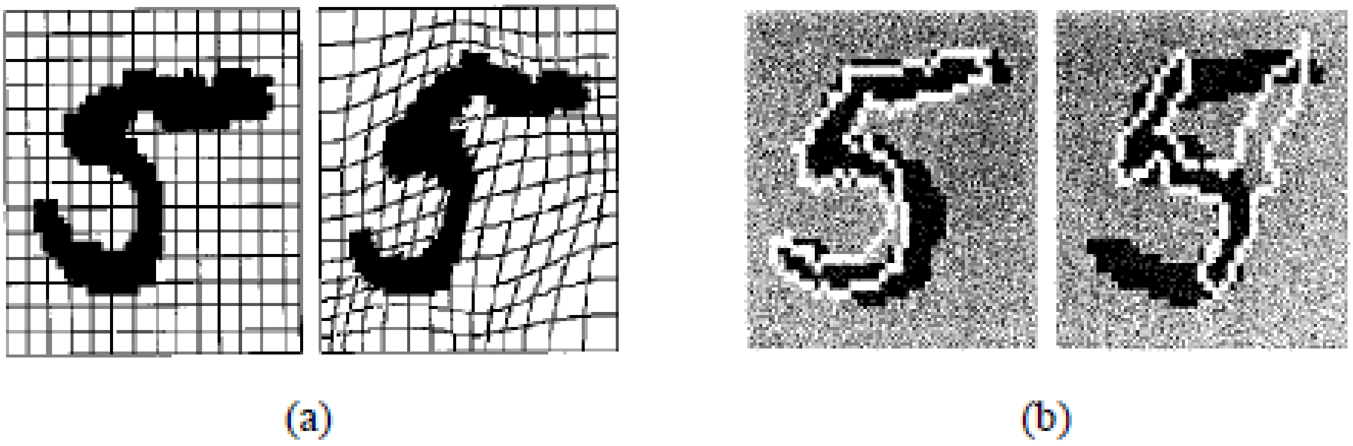
Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)

ABSTRACT: Given the ubiquity of handwritten documents in human transactions, Optical Character Recognition (OCR) of documents have invaluable practical worth. Optical character recognition is a science that enables to translate various types of documents or images into analyzable, editable and searchable data. During last decade, researchers have used artificial intelligence / machine learning tools to automatically analyze handwritten and printed documents in order to convert them into electronic format. The objective of this review paper is to summarize research that has been conducted on character recognition of handwritten documents and to provide research directions. In this Systematic Literature Review (SLR) we collected, synthesized and analyzed research articles on the topic of handwritten OCR (and closely related topics) which were published between year 2000 to 2019. We followed widely used electronic databases by following predefined review protocol. Articles were searched using keywords, forward reference searching and backward reference searching in order to search all the articles related to the topic. After carefully following study selection process 176 articles were selected for this SLR. This review article serves the purpose of presenting state of the art results and techniques on OCR and also provide research directions by highlighting research gaps.

DATA EXTRACTION AND SYNTHESIS

During this phase, metadata of selected studies (176) was extracted. As stated earlier, we used Mendeley and MS Excel to manage the metadata of these studies. The main objective of this phase was to record the information that was obtained from the initial studies [22]. The data containing study ID (to identify each study), study title, authors, publication year, publishing platform (conference proceedings, journals, etc.), citation count, and the study context (techniques used in the study) were extracted and recorded in an excel sheet. This data was extracted after a thorough analysis of each study to identify the algorithms and techniques proposed by the researchers. This also helped us to classify the studies according to the languages on which the techniques were applied. Table 2 shows the fields of the data extracted from research studies.



LANGUAGE SPECIFIC RESEARCH

The distributions/number of selected studies with respect to investigated scripting languages. A total number of selected studies are 176, and out of these 172 studies, the English language has the highest contribution of 53 studies in the domain of handwritten character recognition, 44 studies related to the Arabic language, 37 studies are on the Indian scripts, 23 on the Chinese language, 118 on the Urdu language, while 14 studies were conducted on the Persian language. Some of the selected articles discussed multiple languages.

CLASSIFICATION METHODS OF HANDWRITTEN OCR

In handwritten OCR an algorithm is trained on a known dataset, and it discovers how to accurately categorize/classify the alphabets and digits. Classification is a process to learn a model on a given input data and map or label it to predefined category or classes [17]. In this section, we have discussed the most prevalent classification techniques in OCR research studies beginning from 2000 till 2019.

ARTIFICIAL NEURAL NETWORKS (ANN)

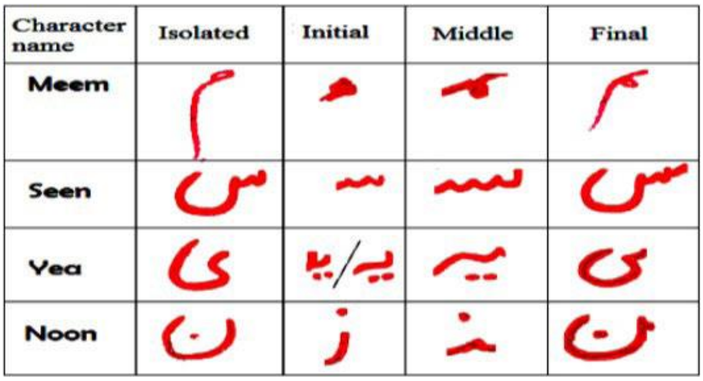
Biological neuron inspired architecture, Artificial Neural Networks (ANN) consists of numerous processing units called neurons [56]. These processing elements (neurons) work together to model given input data and map it to predefined class or label [57]. The main unit in neural networks is nodes (neuron). Weights associated with each node are adjusted to reduce the squared error on training samples in a supervised learning environment (training on labelled samples/data). Figure 8 presents a pictorial representation of Multi-Layer Perceptron (MLP) that consists of three layers, i.e. (input, hidden and output). Feedforward networks / Multi-Layer Perceptron (MLP) achieved renewed interest of research community in the mid 1980s as by that time ``Hopfield network'' provided the way to understand human memory and calculate the state of a neuron [59]. Initially, the computational complexity of finding weights associated with neurons hindered the application of neural networks. With the advent of deep (many layers) neural architectures, i.e. Recurrent Neural Network (RNN) and Convolutional Neural Networks (CNN), neural networks have established itself as one of the best classification technique for recognition tasks including OCR [60]fi[63]. Refer Sections VIII and IX-B for current and future research trends. The early implementation of MLP for handwritten OCR was done by Shamsher *et al.* [64] on the Urdu language. The researchers proposed feed-forward neural network algorithm of MLP (Multi-Layer Perceptrons) [65]. Liu and Suen [66] used MLP on Farsi and Bangla numerals. One hidden layer was used with the connecting weights estimated by the error backpropagation (BP) algorithm that minimized the squared error criterion. On the other hand, Cirecsan *et al.* [30] trained five MLPs with two to nine hidden layers and varying numbers of hidden units for the recognition of English numerals. Recently, Convolutional Neural Network (CNN) has reported great success in character recognition task [67]. A convolutional neural network has been widely used for classification and recognition of almost all the languages that have been reviewed for this systematic literature review [68]fi[74].

STRUCTURAL PATTERN RECOGNITION

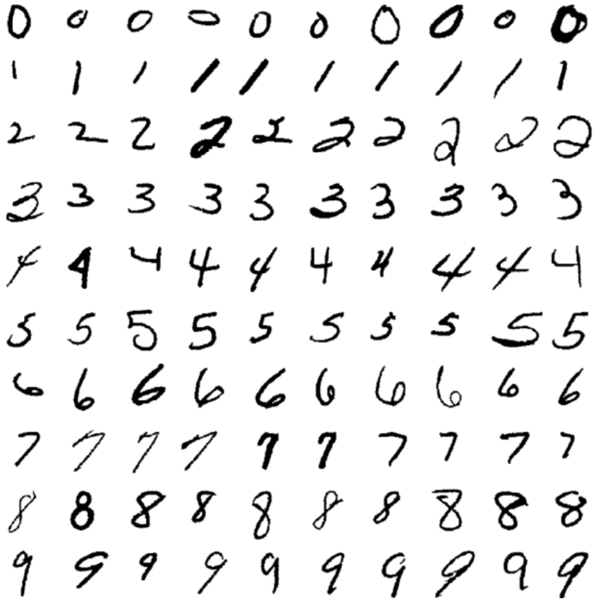
Another classification technique that was used by OCR research community before the popularization of kernel methods and neural networks / deep learning approach was structural pattern recognition. Structural pattern recognition aims to classify objects based on a relationship between its pattern structures and usually structures are extracted using pattern primitives (refer Figure 11 for an example of pattern primitives), i.e. edge, contours, connected component geometry etc. One of such image primitive that has been used in OCR is Chain Code Histogram (CCH) [98], [99]. CCH effectively describes image / character boundary / curve, thus helping in classify character [57], [75]. Prerequisite condition to apply CCH for OCR is that image should be in binary format, and boundaries should be well defined. Generally, for handwritten character recognition, this condition makes CCH difficult to use. Thus, different research studies and publicly available datasets use/provide binarized images [87]. In research studies of OCR, structural models can be further subdivided on the basis of the context of structure, i.e. graphical methods and grammar-based methods. Both of these models are presented in the next two sub-sections.

DATASETS

Generally, for evaluating and benchmarking different OCR algorithms, standardized databases are needed/used to enable a meaningful comparison [55]. Availability of a dataset containing enough amount of data for training and testing purpose is always a fundamental requirement for a quality research [110], [111]. Research in the domain of optical character recognition mainly revolves around six different languages, namely, English, Arabic, Indian, Chinese, Urdu and Persian / Farsi script. Thus, there are publicly available datasets for these languages such as MNIST, CEDAR, CENPARMI, PE92, UCOM, HCL2000 etc. Following subsections presents an overview of most used datasets for the above mentioned languages.



Character name	Isolated	Initial	Middle	Final
Meem	م	مـ	ـمـ	ـم
Seen	س	سـ	ـسـ	ـس
Yea	ى	يـ/يـ	ـيـ	ـى
Noon	ن	نـ	ـنـ	ـن



ARABIC SCRIPT

Research on handwritten Arabic OCR systems has passed through various stages over the past two decades. Studies in the early 2000s focused mainly on the neural network methods for recognition and developed variants of databases [165]. In 2002, Pechwitz *et al.* [37] developed the first IFN/ENIT-database to allow for the training and testing of Arabic OCR systems. This is one of the highly cited databases and has been cited more than 470 times. Another database was developed by Mozaffari *et al.* [166] an Mozaffari and Soltanizadeh [167] in 2006. It stores grey-scale images of isolated offiine handwritten 17,740 Arabic / Farsi numerals and 52,380 characters. Another notable dataset containing Arabic handwritten text images was introduced by Mezghani *et al.* [168]. The dataset has an open vocabulary written by multiple writers (AHTID/ MW). It can be used for word and sentence recognition, and writer identification [169]. A survey by Lorigo and Govindaraju [18] provides a comprehensive review of the Arabic handwriting recognition methodologies and databases used until 2006. This includes research studies carried out on IFN/ENIT database. These studies mostly involved artificial neural networks (ANNs), Hidden Markov Models (HMM), holistic and segmentationbased recognition approaches. The limitations pointed out by the review included restrictive lexicons and restrictions on the text appearance. In 2009, Graves and Schmidhuber [24] introduced a globally trained offiine handwriting recognizer based on multi-directional recurrent neural networks and connectionist temporal classification. It takes raw pixel data as input. The system had an overall accuracy of 91.4%, which also won the international Arabic recognition competition. Another notable attempt for Arabic OCR was made by Lutf *et al.* [170] in 2014, which primarily focused on the speciality of the Arabic writing system. The researcher proposed a novel method with minimum computation cost for Arabic font recognition based on diacritics. Flood-fill based and clustering-based algorithms were developed for diacritics segmentation. Further, diacritic validation is done to avoid misclassification with isolated letters. Compared to other approaches, this method is the fastest with an average recognition rate of 98.73% for 10 most popular Arabic fonts. An Arabic handwriting synthesis system devised by Elarian *et al.* [171] in 2015 synthesizes words from segmented characters. It uses two concatenation models: ExtendedGlyphs connection and the Synthetic-Extensions connection. The impact of the results from this system shows significant improvement in the recognition performance of an HMM-based Arabic text recognizer. Akram *et l.* [172] discussed an analytical approach to develop a recognition system based on HMM Toolkit (HTK). This approach requires no priori segmentation. Features of local densities and statistics are extracted using a vertical sliding windows technique, where each line image is transformed into a series of extracted feature vectors. HTK is used in the training phase, and Viterbi algorithm is used in the recognition phase. The system gave an accuracy of 80.26% for words with ``Arabic-numbers'' database and 78.95% with IFN / ENIT database. In a study conducted in 2016 by Elleuch *et al.* [173], convolutional neural network (CNN) based on support vector machine (SVM) is explored for recognizing offiine handwritten Arabic. The model automatically extracts features from raw input and performs classification. In 2018, researchers applied the technique of DCNN (deep CNN) for recognizing the offiine and handwritten Arabic characters [174]. An accuracy of 98.86% was achieved when the strategy of DCNN using transfer learning was applied to two datasets. In another similar study [175] an OCR technique based on HOG (Histograms of Oriented Gradient) [176] for feature extraction and SVM for character classifi- cation was used on the handwritten dataset.

ENGLISH SCRIPT

The English Language is the most widely used language in the world. It is the official language of 53 countries and articulated as a first language by around 400 million people. Bilinguals use English as an international language. Character recognition for the English language has been extensively studied throughout many years. In this systematic literature review, the English language has the highest number of publications, i.e. 45 publications after concluding the study selection process (refer Section II-D and Section III-D). The OCR systems for the English language occupy a significant place as a large number of studies have been done in the era of 2000-2018 on the English language. The English language OCR systems have been used successfully in a wide array of commercial applications. The most cited study for English language handwritten OCR is by Plamondon and Srihari [35] in 2000, which have more than 2900 citations, refer Table 3. The objective of the research by Plamondon *et al.* was to present a broad review of state of the art in the field of automatic processing of handwriting. This paper explained the phenomenon of pen-based computers and achieved the goal of automatic processing of electronic ink by mimicking and extending the pen-paper metaphor. To identify the shape of the character, structural and rule-based models like (SOFM) self-organized feature map, (TDNN) time-delay neural network and (HMM) hidden Markov model was used. Another comprehensive overview of character recognition presented in [36] by Arica *et al.* has more than 500 citations. Arica *et al.* concluded that characters are natural entities, and it is practically impossible for character recognition to impose a strict mathematical rule on the patterns of characters. Neither the structural nor the statistical models can signify a complex pattern alone. The statistical and structural information for many characters pattern can be combined by neural networks (NNs) or harmonic markov models (HMM). Connell and Jain [9] demonstrated a template-based system for online character recognition, which is capable of representing different handwriting styles of a particular character. They used decision trees for efficient classification of characters and achieved 86% accuracy. Every language has specific way of writing and have some diverse features that distinguished it with other language. We believe that to efficiently recognize handwritten and machine printed text of the English language, researchers have used almost all of the available feature extraction and classification techniques. These feature extraction and classification techniques include but not limited to HOG [130], bidirectional LSTM [131], directional features [132], multilayer perceptron (MLP) [119], [133], [134], hidden Markov model(HMM) [26], [52], [54], [62], Artificial neural network (ANN) [135]fi[137] and support vector machine (SVM) [29], [67]. Recently trend is shifting away from using handcrafted features and moving towards deep neural networks. Convolutional Neural Network (CNN) architecture, a class of deep neural networks, has achieved classification results that exceed state-of-the-art results specifically for visual stimuli/input [138]. LeCun [20] proposed CNN architecture based on multiple stages where each stage is further based on multiple layers. Each stage uses feature maps, which are basically

arrays containing pixels. These pixels are fed as input to multiple hidden layers for feature extraction and a connected layer, which detects and classifies object [55]. A recent study by [69] used fully convolutional neural network(FCNN) on IAM and RIMES datasets. Results were promising, and researchers achieved the character error rate(CER) and word error rate(WER) of 4.7%, 8.22%, 2.46%, 5.68% respectively. Jayasundara [139] proposed a novel technique called capsule networks(CapsNet) for the handwritten character recognition with very small datasets.