OPTICAL CHARACTER RECOGNITION USING DEEP LEARNING – A TECHNICAL REVIEW

Ms. Preeti P. Bhatt, Ms. Isha Patel

Abstract: OCR is used to identify the character from human written text. To recognize the text segmentation of character is important stage. So here, we addressed different techniques to recognize the character. This document also presents comparison of different languages for character and numeral recognition with its accuracy achieved by different writer. Segmentation problem of each language were different also handwritten character was also varied user to user, so it is necessary to make OCR systems more effective and accurate for segmentation. Comparative study concludes that deep learning technique gives good segmentation and gives better result in case with large dataset compares to other techniques.

KEYWORD - OCR, deep learning, CNN.

I INTRODUCTION

OCR is the automated translation of images of typed, printed or handwritten text into coded text, whether from a scanned document, a photo of a document, from subtitle text overlying on an image.[1] There are many methods of OCR to recognise character like thinning, thickening, pre-processing, feature extraction, feature vector, segmentation and so on. Handwriting recognition is a pivotal issue in machine learning. A methodologies and techniques have been proposed but it still an unresolved issue. [1] Using neural network, it is done but it has some problem with it, so one can use deep learning over it for better results.

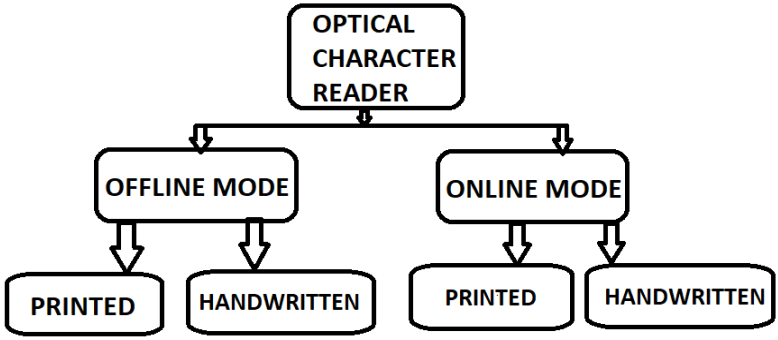

Figure 1 classification of OCR

In recognition system involves major five steps especially, Data collection, pre-processing, segmentation, features extraction and classification.

- **Data Collection:** For data acquisition many researchers used standard data sets available globally and some of researches uses own data sets for processing.
- **Pre - processing:** In pre-processing phase many mathematical and morphological operations are applied on input document image for gay scale conversion, normalization, finalization, baseline detection, skew correction, slant detection, slant correction, noise removal and etc.
- **Segmentation:** If the input document image consists of a sentence than that will be segmented into words than each word is consider as distinct substance.
- **Feature Extraction:** The word is taken for further process, features like density features, structure based features, hierarchical features and other features are extracted from the input image. Which are the features extracted those are comparing with the training image features if the matching class is presented than that image is recognized.
- **Classification:** for classification process classifiers are used like support Vector Machine (SVM), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) , Neural Networks, K Nearest Neighbour (KNN), Hidden Markov Model (HMM) and etc.[3]

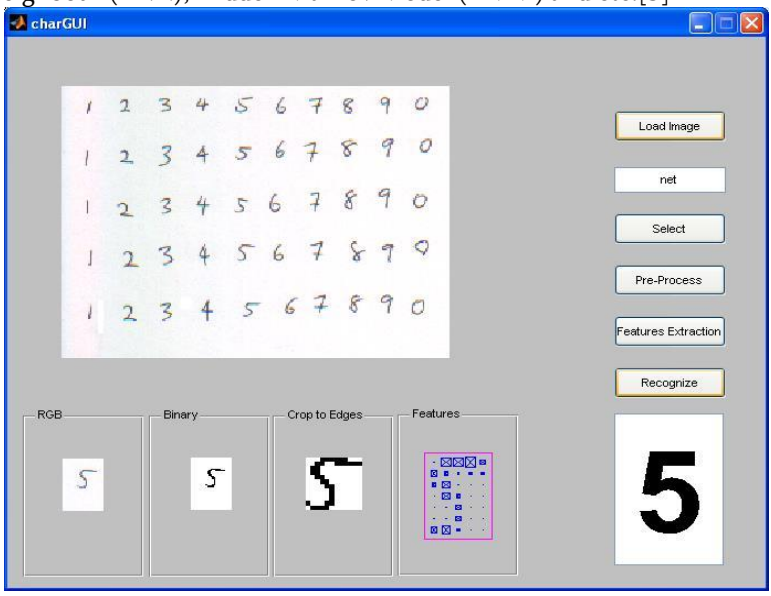

Figure 2 processing of OCR

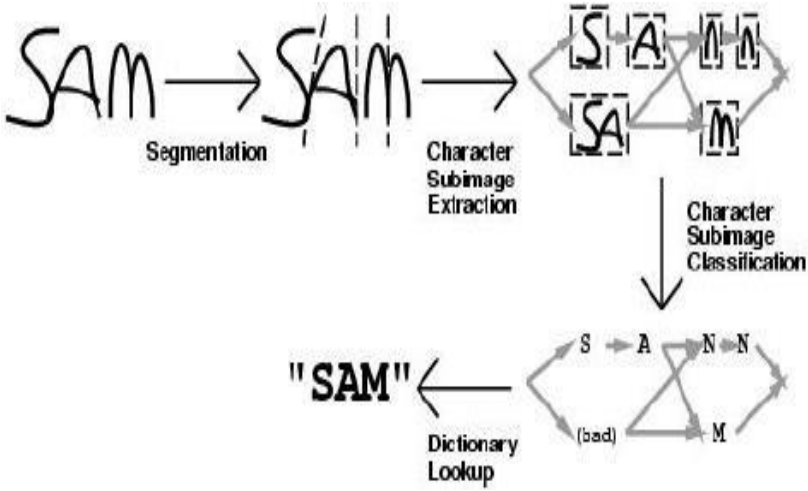

Figure 3 processing of OCR (2)

There are 2 types of recognition: online recognition and offline recognition.

- Online recognition refers to the process of recognizing handwriting recorded with a digitizer on an electronic notepad, tab and etc by a special pen and at that time temporal information is stored as a time sequence of pen coordinates.

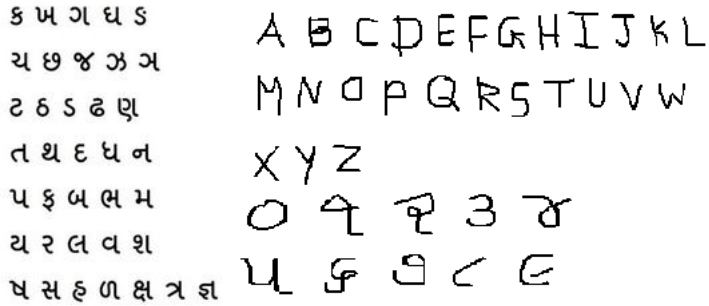

Figure 4 Online Handwritten Text

- Offline Recognition is a process of scan paper and stored the character written in the documents by identifying the text and process on it.

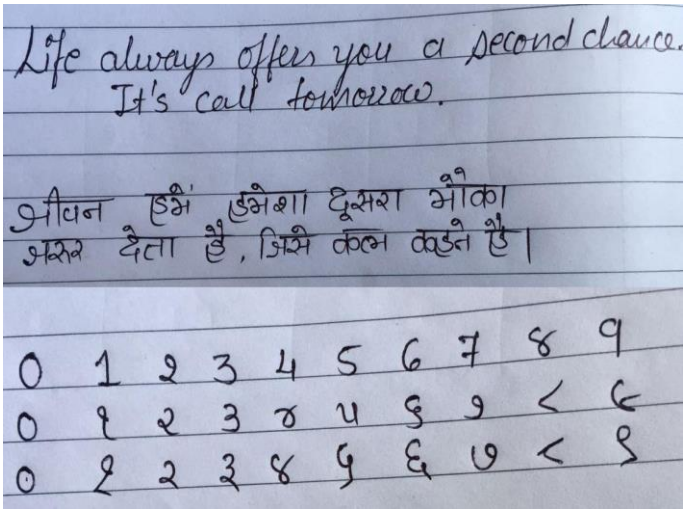

Figure 5 Offline Handwritten character

II DEEP LEARNING

Deep learning is a powerful feature extraction method applied to extract the feature of the handwritten characters. Deep learning provides the task specific method, which inherits features from machine learning methods based on learning data representation.

Learning can be supervised, semi supervised or unsupervised. It architectures such as Deep Neural Network, deep belief networks and RNN have been applied to fields natural language processing, computer vision, speech recognition, drug design, social network filtering, machine translation, bioinformatics, audio recognition, and where they have produced results comparable to and in some cases superior to human experts.[1]

There are many nonlinear hidden layers in deep neural networks, so the number of connections and parameters are very large. Apart from this, the train is also getting very difficult, too CNN is a class of deep neural network with relatively small parameters set and easy to train. The ability to model CNN correctly, the input data can be replaced with the number of layered layers and trainable parameters at each level and they also make the correct assumptions on the nature of the images. [2]

A CNN is a class of deep, feed-forward artificial neural networks that has successfully been applied to analysing visual imagery. CNNs also have been applied to acoustic modelling for automatic speech recognition (ASR).A RNN is a

class of artificial neural network where connections between units form a directed cycle. This allows it to exhibit dynamic temporal behaviour. In which data can flow in any direction, are used for applications such as language modelling. Long short-term memory is particularly effective for this use. [1]

III LITERATURE REVIEW

In [2], around 90000 images of more than 40 different classes of characters of Devanagari script were segmented from handwritten documents. Used deep learning architecture for recognition and CNN for superior result to traditional shallow networks in many recognition tasks and focus the use of Idler and dataset increment approach to improve test accuracy. The base form of consonant characters can be combined with vowels to form additional characters which is not explored in that research. So, they used Deep CNNs with additional Dataset increment techniques and Dropout layer which results in very high test accuracy even for a various and challenging dataset.

In [3], they describes same problem of handwriting recognition. They used holistic approach to identify the handwritten words, each word take is an individual entity so holistic approach is better and used such methods namely density features, long run features and structural features for extraction in the input handwritten document image. After that they apply classification by using Support Vector Machines (SVM). They achieved 88.13% of recognition rate.

In [4], they researched on digit recognition in Arabic with help of CNN. They proposed a novel algorithm based on deep learning neural networks using appropriate activation function and regularization layer, which shows significantly improved accuracy compared to the existing Arabic numeral recognition methods. In the Multi-Layer Perceptron (MLP) model, they implement dropout regularization to reduce over fitting in between fully connected layers. The output layer, consisting of 10 neurons with softmax activation, predicts the probability for 10 individual digit classes (0-9). They apply two method in it, MLP and CNN but they achieved high accuracy in CNN.

In [5], they propose a workflow and a machine learning model for recognizing handwritten characters on form document. It is based on CNN as a powerful feature extraction and Support Vector Machines (SVM) as a high-end classifier. Based on the experiment results using data, both for training and testing, the proposed method achieves an accuracy rate better than only CNN method. The proposed method was also validated using ten folds cross-validation, and it shows that the recognition rate for this proposed method is still able to be improved.

In [6], they investigate the applicability of Deep Convolutional Neural Network (DCNN) using the transfer learning strategies on two datasets; they demonstrated the abilities of satisfactory recognition and done enhanced methods in the field of handwritten Arabic character recognition (HACR). He examined and discussed the use of CNN in the field of off-line HACR. We used the same architecture as Alex-Net, without the pre-processing phase and with a three learning strategy.

In [7], they provide a new system for DCNN DropSample, and apply it to a large number of online handwritten Chinese letter identification (HCRR). It is connected to the Quota Function, which is dynamically DCNN. Based on confidence expressed by softmax output. It is with a variety of domain-specific knowledge, the accuracy of HCCR can be improved effectively.

In [8], they recognize handwritten Devanagari digits. They use density and background direction distribution facilities for zones. They use common images of different sizes of 32 * 32, 40 * 40 and 48 * 48. For the purpose of classification, they used the SMS classifier with the RBF kernel. Documents used for handwritten Devanagari numbers are given by the Indian Statistical Institute (ISI), Kolkata. They recommend a 144-size feature vector to identify the test sample. 32 * 32 normally tested, the accuracy of the test by the 144 quake is 98.51%, which is prominent and cost-efficient.

In [9], they used two classification methods to identify handwritten Devanagari numbers. They used two classifiers HMM and ANN to introduce the recognition system. The digital image is classified according to the maximum score obtained by ANN Classifier.

In [10], they introduced a novel offline strategy for recognition of online handwritten characters written in Devnagari entered in an unrestricted manner. They experiment different classifiers like SVM, HMM, ANN and trained on statistical, structural or spectral features but they used CNN because it allows writers to enter characters in any number or arrangement of strokes and is also strong to certain amount of overwriting. They tests with 10 different arrangements of CNN and for both Exponential Decay and Inverse Scale Annealing approaches to convergence, show highly promising results. Using a hybrid approach they conclude that character level data is extracted from the collected words and covers all probable variants owing to the different writing styles and varied parent word structures.

In [11], they recognize offline handwritten Guajarati character based on water reservoir and radial histogram. They presented use of structural features based on water reservoir principle and radial histogram for Gujarati character recognition.

They used two-layer feed-forward neural network for classification to train the recognition system. They conclude a two-layer feed-forward network is obtained overall 74.16% of accuracy.

IV COMPARATIVE ANALYSIS

This section provides summarized representation of Indian languages and several techniques used.

Table 1: Information of scripts

Script	Languages	Consonants	Vowels	Existence of modifiers
Devanagari[13]	Many	33	14	Yes
Gujarati[13]	Gujarati	34	12	Yes
Bangla[13]	Assamese, Bengali	36	21	Yes
Oriya[13]	Oriya	34	14	Yes
Gurumukhi[13]	Punjabi	35	9	Yes
Chinese	Chinese	22	9	Yes
Arabic	Arabic	25	3	Yes

Table 1 shows the languages and composition of vowels and consonants and status of modifiers.

Table 2: work done on handwritten recognition

Script	Data type	Classifier	Accuracy	Reference
Devanagari	Numeral	SVM	98.62%	[8]
Devanagari	Numeral	1) ANN 2) HMM	1) 92.83%, 2) 87.69%	[9]
Devanagari	Character	CNN	98.19%	[10]
Devanagari	Character	DCNN	98.47%	[2]
Gujarati	Character	Neural Network	74.16%	[11]
Gujarati	Character	Binary tree and K-NN	63.1%	[12]
Chinese	Character	DCNN	97.737%	[7]
Arabic	Numeral	DCNN	97.4%	[4]
Arabic	Character	DL technique : 1) CNN from scratch 2) CNN as fixed feature extractor 3) CNN fine-tuned	With pre-processing 1) 99.97% 2) 70.77% 3) 96.19%	[6]

Table 3: pros and cons of different techniques

Technique	Advantage	Disadvantage
CNN	• Accuracy in image recognition problems	• Need a lot of training data
Neural networks	• Neural networks can be trained with any number of inputs and layers. • Neural networks work best with more data points.	• As the number of neurons increase the network becomes complex • Requires high processing time for large neural networks
ANN	• Large amount of data sets. • The output performance will depend upon the trained parameters and the data set relevant to the training.	• It cannot extrapolate the results. • Extracting the knowledge (weights in ANN) is very difficult (few papers are available and they have classified ANN as grey box model rather than a black-box model)
KNN	• Naturally handles multi-class cases • Flexible to feature / distance choices • Can do well in practice with enough representative data	• Large search problem to find nearest neighbours • Storage of data • Must know we have a meaningful distance function
HNN	• Allowing more sequences to be significantly found	• Needs to be trained on a set of seed sequences and generally requires a larger seed than the simple Markov models
SVM	• Easy to integrate different distance function • Fast classification of new object • Very small training sample sizes	• Slow training • Must combine two-classes

V CONCLUSION

This paper includes the discussion of different scripts. The challenges and problems of the review paper are discussed such that it is helpful to the researcher working in this field. In India, regional languages have problems like line segments, half characters, same characters, different thicknesses, discontinuation or distortion of characters.

The challenging task about character recognition is that it needs big data set to use deep learning techniques. In comparative study (in table 2), by comparing deep learning techniques and other techniques, we reach to a conclusion that using deep learning technique leads to achievement of better result and high accuracy.

In future, for more accuracy we can do research on OCR using deep learning in Guajarati language.

VI REFERENCES

[1] Shailesh Acharya, Ashok Kumar Pant, Prashnna Kumar Gyawali," Deep Learning Based Large Scale Handwritten Devanagari Character Recognition", 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), 2015.

[2] Jenis J. Macwan, Mukesh M. Goswami, Archana N. Vyas, "A Survey on Offline Handwritten North Indian Script Symbol Recognition", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016.

[3] Shruthi A , M S Patel," Offline Handwritten Word Recognition using Multiple Features with SVM Classifier for Holistic Approach" , International Journal of Innovative Research in Computer and Communication Engineering, June 2015.

[4] Akm Ashiquzzaman and Abdul Kawsar Tushar," Handwritten Arabic Numeral Recognition using Deep Learning Neural Networks", 2011.

[5] Darmatasia and Mohamad Ivan Fanany," Handwriting Recognition on Form Document Using Convolutional Neural Network and Support

Vector Machines", 2017 Fifth International Conference on Information and Communication Technology (ICoICT), 2017.

[6] Chaouki Boufenar , Adlen Kerboua, Mohamed Batouche," Investigation on deep learning for off-line handwritten Arabic character recognition", Cognitive Systems Research, 2017.

[7] WeixinYang , LianwenJin , DachengTao , ZechengXie , ZiyongFeng, "A new training method to enhance deep convolutional neural networks for large-scale unconstrained handwritten Chinese character recognition".

[8] Jangid, Mahesh, et al. "SVM classifier for recognition of handwritten Devanagari numeral." Image Information Processing (ICIIP), 2011 International Conference on. IEEE, 2011.

[9] Bhattacharya, U., et al., "Neural combination of ANN and HMM for Handwritten Devanagari numeral recognition." Tenth International Workshop on Frontiers in Handwriting Recognition. Suvisoft, 2006.

[10] Mehrotra, Kapil, et al., "Unconstrained handwritten Devanagari character recognition using convolutional neural networks." Proceedings of the 4th International Workshop on Multilingual OCR. ACM, 2013.

[11] Jitendra V. Nasriwala, Dr. Bankim C. Patel, "Offline Handwritten Gujarati Character Recognition Based on Water Reservoir and Radial Histogram", International Journal of Electrical Electronics & Computer Science Engineering, 2016.

[12] C. Patel, and A. A. Desai, "Gujarati Handwritten Character Recognition Using Hybrid Method Based On Binary Tree-Classifier And K-Nearest Neighbor". International Journal of Engineering Research and Technology, ESRSA Publications. (Vol. 2, No. 6). June-2013.

[13] Jenis J. Macwan, Mukesh M. Goswami, Archana N. Vyas, "A Survey on Offline Handwritten North Indian Script Symbol Recognition", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016.

AUTHORS' PROFILE

Ms. Preeti P. Bhatt is working as an Assistant Professor at BMIIT. She has more than seven years of experience in academics. Her interest area includes artificial intelligence. She has attended many seminar and workshops and also publishes research papers in national and international journal. She can be contacted at preeti.bhatt@utu.ac.in

Ms. Isha Patel is studying in 5 Years Integrated M.Sc. (IT) programme, semester 8 of Babu Madhav Institute of Information Technology affiliated to Uka Tarsadia University-Bardoli.