Converting Non-Digitzed Health Data to Digital Format

Sarita Rathod[1]

[1]Department of Information Technology,K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai, India.

[1]sarita.r@somaiya.edu

***Abstract*—The main aim of our project is to convert medical records in traditional non-digitized format to a more useful digitized format using Optical Character Recognition (HCR). Using HCR software to recognize medical documents has many benefits. Manual method for data entry was used previously to capture data from medical records, when HCR was not present. With this handwritten character recognition method, it shows the ability of a computer to receive and recognize handwritten data in medical records. This paper mainly focuses on the recognition of handwritten English characters. Deep learning is the method used for identity recognition which will depend on the neural network. The records digitized using neural network will be stored on the cloud for future use.**

Keywords— HCR (Handwritten Character Recognition), Deep learning, Neural Network, Keras etc.

1. Introduction

HCR for medical documents is useful because this software provides invaluable benefits in terms of cost savings and even increases productivity. When used in hospitals or health centres such software can be used to perform a variety of functions. A use of HCR can be scanning patient identity cards and capturing data from them. After this the identity card is stored as a scanned image that is fully text searchable. This means that when processed with HCR for medical documents the resultant scanned images along with the text searchable version can be stored easily in databases, and located with a simple text search[2].When HCR for medical documents software is used many other medical documents such as patient records, test results and active and historical medical records can all be scanned and converted into text searchable format[8].This results in faster conversion of documents into text searchable format, as well as makes them easier to find in databases, which is not possible in the case of scanned documents[7].Before the use of HCR for medical documents manual data entry was used throughout the world when capturing data from medical records. This was a time consuming and costly process as data entry teams spent thousands of hours on capturing data and ensuring its accuracy[1].This manual process also meant that large data entry teams had to be maintained and paid. Most manual data entry work has been done away with since HCR for medical documents software has come into use. In fact, this software can provide the same accuracy rates as manual data entry but in a fraction of the time. HCR for medical documents operates at high speeds and often can process batches of medical documents in different formats. The installation of this software can sometimes even process documents faster than an entire data entry team, and since it can operate 24/7 data documents can enter a database faster.

In healthcare, the efficacy of a treatment plan mainly depends on the information a patient can divulge on his own medical history, but our documentation is largely dependent on our memories and at best on previous medical reports, which largely compromises the accuracy of the history. When HCR for medical documents software is used, many medical documents such as patient records, test results and active and historical medical records can be scanned and converted into searchable text format[1].This ultimately results in faster conversion of documents into searchable text format, and also makes it easier to find in databases, which is not possible in the case of today's system. Such is the power of Deep Learning, which we are going to implement in our project. extraction. To develop the system, Python and openCV libraries for programming purposes are utilized. The classifiers which are useful for the classification are K-NN, CNN etc[3].

A. Scope of Project

- Improving the accuracy of recognition of the characters during document processing compared to various existing available character recognition methods.
- HCR technique derives the meaning of the characters, from their bit-mapped images.
- Increasing the character recognition speed in document processing.
- The system can process a huge number of documents with-in less time and saves time.
- Provide an efficient and enhanced software tool for the users to perform Medical record Analysis, document processing by reading and recognizing

the characters in medical organizations that are having large pools of documented, scanned images.

- The product will recognize them, search them and process them faster according to the needs of the environment.

2. Related Work

Below In [1] Datt and Amin, analyzed Gujarati handwritten characters using the Artificial neural network. Errors were corrected by RBPNN (Radial Basis Probabilistic Neural Network) method. Also, the HCR is the technology which provides the fast and automated data capture which helps to save time as well as cost too. Image acquisition, processing, segmentation, feature extraction, classification and recognition, post-processing these are the basic steps which are involved in the HCR.

In [2] Chandarana and Kapadia compared handwritten optical characters of different fonts with standard ones by using the techniques Feature extraction, Segmentation, Template matching and correlation. Recognition accuracy observed was 91.16% when the number of images was tasted under the experiment which was better as compared to previous ones.

In [3] Patel and Desai used cross types approach predicated on tree classifier and k-Nearest Neighbour (k - NN) for identification of handwritten Gujarati characters. Combination of structural features along with statistical features is used for classification and identification of characters. The features are relatively simple to derive. By studying the appearance of various handwritten characters the structural features are selected. Combination of structural features and statistical features are used for classification and identification of characters. The structural features were selected by studying the appearance of various handwritten characters of Gujarati script. A success rate of 63% was achieved using this approach.

In [4] Sojitra and Dhakad preferred the algorithm Neural Network, Self-organizing map and the classifier used is a nearest neighbour classifier. The analysis was carried out with ten input samples and five different fonts. 50 samples are collected for a single character of Gujarati script. Accuracy found was 97.78% and recognition rate for slight changes in the 3 training set was found 98.83%. The results state that pre-processing of data before giving input to K-NN has given the highest recognition rate. Merging Neural with existing methods for recognition has given optimum results and best recognition rate. The comparison is performed on a pixel by pixel basis.

In [5] Prasad and Kulkarni have stated that the recognition rate image is highly affected by the similarity of various characters, and these characters which can be verified with some recognition rates of Class 2, 3 and 4. In these classes, there are more similar characters which in turn degrade and recognition rate. It classified the Gujarati Character set into 6 classes, which is based on their geometrical shapes. An average overall Recognition Efficiency of 71.66 %. Each letter was evaluated in a learning set and efficiency of recognition was evaluated.

Adobe Scan can quickly and easily turn physical identity, business cards etc. into digital contacts on your mobile phone. Adobe scan uses image processing techniques, to scan the digital text on a identity or business card extractable, reusable, and searchable in a secure, reliable PDF. It can even automatically remove unwanted objects from your business or identity card scanned.

The motivation from Adobe Scan was that it's fast and easy to capture multi-page documents too. Adobe Scan is optimized for capturing multi-page documents without sacrificing quality. After user has captured a scan, the user will be prompted to save it as a PDF. The motivation from Google HCR API is that it can perform feature detection on a local image file by sending the contents of the image file as a base64 encoded string in the body of the request.

3. Proposed System

The Architecture of the handwritten character recognition system on a grid infrastructure consists of following components:-

- Scanning hardware: It can be used to capture the image of the file that is to be converted to digitized format. The scanning hardware can be a camera or scanner.
- Image pre-processing module: This will help convert the image into the form suitable for the handwriting recognition module. In the case of this system, it will transform the image into grayscale format. This grayscale format of image causes the recognition module to process the image more efficiently than the raw colored image.
- Text area extraction module: This particular module marks the position of handwritten text on the image which helps in segmentation of image to feed it into the recognition module or engine.
- Handwriting recognition module: It is the most important part of the system. This module will recognize the handwritten text extracted from the input image. This module uses the deep learning model to identify the handwritten text from the image.
- Document storage unit: This part of the system will help to store the documents converted into text format for further use of the user of this system.

Below figure shows the System architecture for proposed system.

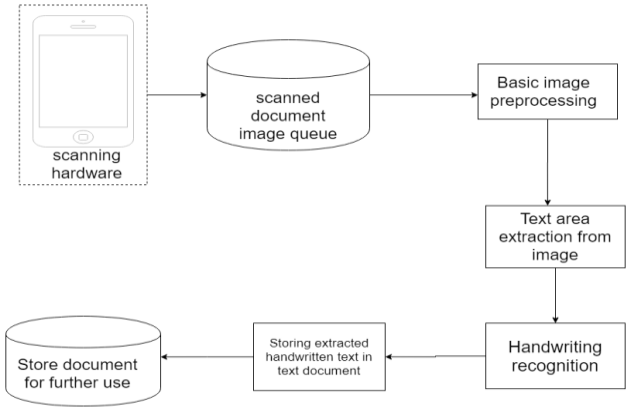

Fig.1 System Architecture for the Proposed System

3A.Methodology, Techniques and Algorithms

Traditionally to establish features like curvature of each type of letter, spacing between letters, etc. the feed them into a classifier like SVM to distinguish between the characters. Here we have used a Deep learning approach to identify such features. We'll break images into small parts and feed them to a Convolutional Neural Network and train using a softmax classification loss function.

a) Data Gathering

The IAM dataset which we have used contains 1539 pages of scanned text sentences written by 600+ writers. This project uses the top 50 writers with the most amount of data. These pages are shortened and then uploaded in the IAM Handwriting dataset in the sentences directory. This data is so elaborate that it has all formats and for all purposes. As Neural Networks don't need much preprocessing of raw data, we have kept the images unchanged rather we make few parts of the image and pass them.

b) Preprocessing

For CNN to understand the writing style, language is not a restriction, so we pass parts of text having image size 113x133 from each sentence. We break them into smaller image sets rather than into words or sentences. For serving the purpose, a generator function is implemented to scan through each sentence and generate random patches with the same patch size. CNN doesn't even need to take the full data, so we have limited the number of patches to be 30% of the total patches which could've been generated from the function and we have shuffled the dataset for preparing a better training model.

c) Self-designed CNN Model

We have used Keras with TensorFlow backend. A standard CNN Model has multiple convolution and maxpool layers, a few dense layers and a final output layer with the softmax activation. ReLU activation was also used between the convolution and dense layers. The resulting model was optimized using Adam Optimizer.

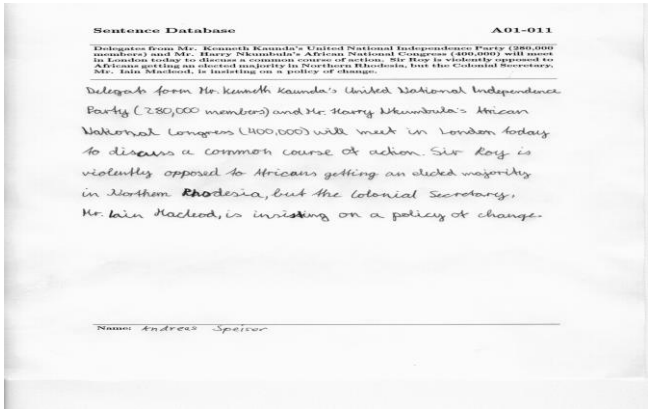

Fig. 2 Data Sample

4. Implementation

4A. Design Details

Our software system Handwritten Character Recognition for medical records can be divided into three modules based on its functionality. The modules classified are as follows:

- Document Processing Module
- System Training Module.
- Document Searching and Editing Module.

a) Document Processing Module

This module is accessed by the administrator and performs certain activities such as scanning documents, storing them as images, recognizing characters in images to transfer them into word format. During the recognition process, this module uses the HCR methodology. The module supports the following services:

- Scanning printed documents.
- Storing the documents as snapshots or images.
- Processing those image-based documents.
- Converting these image-based documents into editable and searchable files.

b) System Training Module

This module can be accessed by both the administrator and the end-user of the system. Before converting the printed documents into editable and search able documents, the first and mandatory step is providing training to the system. Here training is done so that the handwriting in the scanned document should be identified by the system. The image file should be provided as an input during the training process. The user then clicks the training button provided in the recognition module. This module supports:

- Training the system with the handwriting present in the dataset collected from IAM.

- Training the system with the new handwriting that are not present in the system and that cannot be identified by the system.

c) Document Editing Module

This module can be accessed by both the administrator and the end-user. Once the scanned documents are stored, they reside in computer memory ready to upload in the recognition system. This data resides in the form of an image that is viewable using any image viewer software. The desired form of the document may be MS-Word, Text as given by the user. The user queries the system to search for a particular document. Then the system finds the documents based on HCR methodology and returns the result of the search to the user. The objective of this module is to let the user perform:-

- Addition of specific content to the documents
- Deletion of certain content from documents

4B. Results

The implementation of handwriting recognition module tested for some image inputs is given below:

Fig 3: input sample1

```
Init with stored values from ../model/snapshot-38
Recognized: "offered"
Probability: 0.834434
```

Fig 4: output of input sample1

Fig 5: input sample2

```
Init with stored values from ../model/snapshot-38
Recognized: "Gaurar"
Probability: 0.22041999
```

Fig 6: output of input sample2

Conclusion:

We have proposed a design for Handwriting Character Recognition (HCR) Technique which will help in converting handwritten text into a digital text. The Proposed Design will help in detecting the text which is present in the given input and help that input to convert to the digital format which can be stored in a desired format. With the help of Deep Learning, we are able to give the efficient design for the HCR technique.

References:

[1]. Sagar S.Dutt, Prof. Jay D.Amin Student of Gujarat Technological University "Handwritten Gujarati text recognition using artificial neural network and Error correction using Probabilistic Neural network in recognized text " ,IEEE Fourth International Conference on Multimedia Big Data (BigMM) ,2016.

[2]. Jagruti Chandaran, Mayank Kapadia, "Optical Character Recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[3]. Chhaya Patel, Anand, Apurva Desai, Veer Narmad, "Gujarati Handwritten Character Recognition Using Hybrid Method Based On Binary Tree-Classifier And K-Nearest Neighbour" IEEE Transactions on Pattern Analysis and Machine Intelligence ,2014.

[4]. S. Rishi Kumar, G.Madhavan, M. Naveen, S.Subash, U. Selvamalar Beulah Ponrani "Image Processing based Multilingual Translator for Travellers using Raspberry Pi", IEEE Global Humanitarian Technology Conference (GHTC) ,2017 .

[5]. Abin M Sabu, Anto Sahaya Das ,Vimal Jyothi Engineering," A Survey on various Optical Character Recognition Techniques", Proc. IEEE Conference on Emerging Devices and Smart Systems (ICEDSS 2018) ,2-3 March 2018.

[6]. Kian Peymani, Mohsen Soryani" Machine Generated To Handwritten Character Recognition; A Deep Learning Approach", 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA 2017) April 19-20, 2017.

[7]. Build a Handwritten Text Recognition System using TensorFlow, "https://towardsdatascience.com/build-a-handwritten-text-recognition-system-using-tensorflow-2326a3487cd5", accessed on 10 Feb 2019.

[8]. Scheidl - Handwritten Text Recognition in Historical Documents,"https://repositum.tuwien.ac.at/obvutwhs/download/pdf/2874742", accessed on 15 Feb 2019.

[9].https://arxiv.org/pdf/1507.05717.pdf

[10].https://repositum.tuwien.ac.at/obvutwoa/download/pdf/2774578

[11]. Marti - The IAM-database: an English sentence database for offline handwriting recognition, "http://www.fki.inf.unibe.ch/databases/iam-handwriting-database", accessed on 17 Feb 2019.