Journal of Engineering Sciences
Vol 11, Issue 7,July/2020
ISSN NO:0377-9254

Pen-script To Spiel

#Karthik P, #Mugil Kumar S, #Poopalakannan M, *Suganya B
#Student, *Assistant Professor
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College, Coimbatore – 641 022, India.

***Abstract* – Pen-Script character recognition is an area of pattern recognition which defines an ability of a machine to analyse patterns and identify the character. Pattern recognition is the science of making inferences from perceptual data based on either a priori knowledge or on statistical information. The project introduced an innovative, efficient and real-time cost beneficial technique that enables user to hear the contents of text images instead of reading through them. It combines the concept of Optical Character Recognition (OCR) and Text to Speech Synthesizer (TTS). Now-a-days physically challenged persons were achieving lots of things. They are even better than us with the help of some innovative projects. Similarly, we were trying to help these people with our project. Blind people can't able to read the content of notes written by others. They need someone's help to dictate the content of the notes. But, Pen-Script to Spiel system allows these users to hear those content without any intermediate person to dictate it. People who can't read Handwritten prescription given by doctor's can also use this Pen-Script to Spiel System for hearing it to buy those products.**

***Keywords* – Pen-script to spiel, Handwritten to speech, Deep Learning, Convolution Neural Network, Recurrent Neural Network,**

1. INTRODUCTION

Now-a-days physically challenged persons were achieving lots of things. They are even better than us with the help of some innovative projects. Similarly, we were trying to help these people with our project. Blind people can't able to read the content of notes written by others. They need someone's help to dictate the content of the notes. But, Pen-Script to Spiel system allows these users to hear those content without any intermediate person to dictate it. People who can't read Handwritten prescription given by doctor's can also use this Pen-Script to Spiel System for hearing it to buy those products.

Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised. Deep learning architectures such as deep neural networks, deep belief networks, recurrent neural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced results comparable to and in some cases surpassing human expert performance.

Optical character recognition or optical character reader (OCR) is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example from a television broadcast).

In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analyzing visual imagery. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics. They have applications in image and video recognition, recommender systems, image classification, medical image analysis, natural language processing, and financial time series.

A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a

temporal sequence. This allows it to exhibit temporal dynamic behaviour. Derived from feedforward neural networks, RNNs can use their internal state (memory) to process variable length sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition[2] or speech recognition.

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech computer or speech synthesizer, and can be implemented in software or hardware products. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech.

2. Drawbacks of Existing System Model

On-line handwriting recognition is the task of determining what letters or words are present in 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 the Optical Character Recognition (OCR), whose domain can be machine-print only. The introduced system is a character-based recognition and it is a writer independent system. The recognition responsibility of the proposed system is for 52-character classes [uppercases (A-Z) and the lowercases (a-z)]. The suggested system includes the essential stages needed for most of the pattern recognition systems. These stages are the preprocessing stage, the features extraction stage, the pattern matching and classification stage and the postprocessing stage.

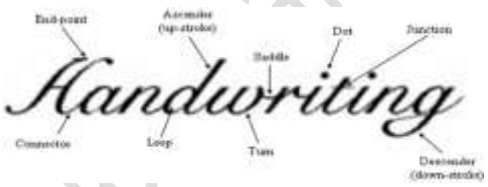

Figure 1 Proposed segments using CNN

The proposed method employs the 3-Dimensional Multiwavelet transform 3D-DMWTCS using multiresolution image decomposition techniques working together with multiple classification methods as a powerful classifier. The classification stage is designed by using a minimum distance classifier depending on Euclidean Distance which has a high-speed performance. The system design also includes a modest postprocessing stage that makes a consistency between the recognized characters within the same word in relation to their upper and lower cases. The overall classification accuracy of proposed systems can be obtained are 95.305 percent with 3D-DMWTCS based on the Rimes database.

3. Methodology

The proposed method comprises of convolutional neural network in which Pen-Script Recognition (PS) systems transcribe text contained in scanned images into digital text. We will build a Neural Network (NN) which is trained on word-images from the IAM dataset. As the input layer (and therefore also all the other layers) can be kept small for word-images, NN-training is feasible on the CPU (of course, a GPU would be better). This implementation is the bare minimum that is needed for PS using TF.

We use a NN for our task. It consists of convolutional NN (CNN) layers, recurrent NN (RNN) layers and a final Connectionist Temporal Classification (CTC) layer.

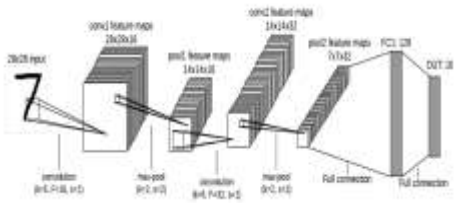

Figure 2 Classified Layers of CNN

We can also view the NN in a more formal way as a function which maps an image (or matrix) M of size W×H to a character sequence (c1, c2, …) with a length between 0 and L. As you can see, the text is recognized on character-level, therefore words or texts not contained in the training data can be recognized too (as long as the individual characters get correctly classified).

CNN: The input image is fed into the CNN layers. These layers are trained to extract relevant features from the image. Each layer consists of three operation. First, the convolution operation, which applies a filter kernel of size 5×5 in the first two layers and 3×3 in the last three layers to the input. Then, the non-linear RELU function is applied. Finally, a pooling layer summarizes image regions and outputs a downsized version of the input. While the image height is downsized by 2 in each layer,

feature maps (channels) are added, so that the output feature map (or sequence) has a size of 32×256.

RNN: The feature sequence contains 256 features per time-step, the RNN propagates relevant information through this sequence. The popular Long Short-Term Memory (LSTM) implementation of RNNs is used, as it is able to propagate information through longer distances and provides more robust training-characteristics than vanilla RNN. The RNN output sequence is mapped to a matrix of size 32×80. The IAM dataset consists of 79 different characters, further one additional character is needed for the CTC operation (CTC blank label), therefore there are 80 entries for each of the 32 time-steps.

CTC: While training the NN, the CTC is given the RNN output matrix and the ground truth text and it computes the loss value. While inferring, the CTC is only given the matrix and it decodes it into the final text. Both the ground truth text and the recognized text can be at most 32 characters long.

Input: It is a Gray-value image of size 128×32. Usually, the images from the dataset do not have exactly this size, therefore we resize it (without distortion) until it either has a width of 128 or a height of 32. Then, we copy the image into a (white) target image of size 128×32. Finally, we normalize the gray-values of the image which simplifies the task for the NN. Data augmentation can easily be integrated by copying the image to random positions instead of aligning it to the left or by randomly resizing the image.

$$\underset{W \times H}{NN: M} \rightarrow \underset{0 \le n \le L}{(C_1, C_2, \ldots, C_n)}$$

Figure 3 The mathematical function of a NN.

CNN output: The output of the CNN layers which is a sequence of length 32. Each entry contains 256 features. Of course, these features are further processed by the RNN layers, however, some features already show a high correlation with certain high-level properties of the input image: there are features which have a high correlation with characters, or with duplicate characters, or with character-properties such as loops.

RNN output: The matrix shown in the top-most graph contains the scores for the characters including the CTC blank label as its last (80th) entry. The other matrix-entries, from top to bottom, correspond to the following characters: " !"#&'()*+,./0123456789:;?ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz". It can be seen that most of the time, the characters are predicted exactly at the position they appear in the image. Only the last character "e" is not aligned.

Pyttsx: This application invokes the pyttsx.init() factory function to get a reference to a pyttsx.Engine instance. During construction, the engine initializes a pyttxs.driver.DriverProxy object responsible for loading a speech engine driver implementation from the pyttxs.drivers module. After construction, an application uses the engine object to register and unregister event callbacks; produce and stop speech; get and set speech engine properties; and start and stop event loops.

4. DATASET

The IAM Handwriting Database contains forms of handwritten English text which can be used to train and test handwritten text recognizers and to perform writer identification and verification experiments.

The database was first published at the ICDAR 1999. Using this database an HMM based recognition system for handwritten sentences was developed and published at the ICPR 2000. The segmentation scheme used in the second version of the database is documented and has been published in the ICPR 2002. The IAM-database as of October 2002 is described. We use the database extensively in our own research, see publications for further details.

The database contains forms of unconstrained handwritten text, which were scanned at a resolution of 300dpi and saved as PNG images with 256 Gray levels. The figure below provides samples of a complete form, a text line and some extracted words.

All forms and also all extracted text lines, words and sentences are available for download as PNG files, with corresponding XML meta-information included into the image files. All texts in the IAM database are built using sentences provided by the LOB Corpus.

The IAM Handwriting Database 3.0 is structured as follows:

- 657 writers contributed samples of their handwriting
- 1'539 pages of scanned text
- 5'685 isolated and labeled sentences
- 13'353 isolated and labeled text lines
- 115'320 isolated and labeled words

The words have been extracted from pages of scanned text using an automatic segmentation scheme and were verified manually. The segmentation scheme has been developed at our institute.

All form, line and word images are provided as PNG files and the corresponding form label files, including segmentation information and variety of estimated parameters (from the pre-processing steps described), are included in the image files as meta-information in XML format which is described in XML file and XML file format (DTD).

5. Conclusion

We discussed a NN which is able to recognize text in images. The NN consists of 5 CNN and 2 RNN layers and outputs a character-probability matrix. This matrix is either used for CTC loss calculation or for CTC decoding. An implementation using TF is provided and some important parts of the code were presented. Finally, hints to improve the recognition accuracy were given.

The project introduced an innovative, efficient and real-time cost beneficial technique that enables user to hear the contents of text images instead of reading through them. It combines the concept of Optical Character Recognition (OCR) and Text to Speech Synthesizer (TTS). With accuracy of 94% in Pen-script prediction and it is been converted to spiel of desired language.

Neural Network Model	Accuracy Rate
BST	72.45%
Beam Search	84.73%
Word Beam Search	94.08%

Table 1 Accuracy Rate of Neural Network Model

This work can be extended by making a complete application for blind people who can't able to read the content of notes written by others. They need someone's help to dictate the content of the notes. So, Pen-Script to Spiel system allows these users to hear those content without any intermediate person to dictate the content of the notes.

The project introduced an innovative, efficient and real-time cost beneficial technique that enables user to hear the contents of text images instead of reading through them. It combines the concept of Optical Character Recognition (OCR) and Text to Speech Synthesizer (TTS). With accuracy of 94% in Pen-script prediction and it is been converted to spiel of desired language.

People who can't read Handwritten prescription given by doctor's can also use this Pen-Script to Spiel System for hearing it to buy those products.

Thus, it can be further improved to have an accuracy over 95% - 99% and some extra features were being planned to add to this system.

6. References

[1] Archana A, Shinde D. (2012) Text pre-processing and text seg-mentation for OCR. International Journal of Computer Science Engineering and Technology. 2012:810–12.

[2] Ciresan .D .C, U. Meier, J. Schmidhuber. (2012) Multi-column Deep Neural Networks for Image Classification. IEEE Conf. on Computer Vision and Pattern Recognition CVPR 2012.

[3] Graves .A, M. Liwicki, S. Fernandez, R. Bertolami, H.Bunke, J.Schmidhuber.(2009) A Novel Connectionist System for Improved Unconstrained Handwriting Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 5, 2009.

[4] Marti .U and H. Bunke. (1999) A full English sentence database for off-line handwriting recognition. In Proc. of the 5th Int. Conf. on Document Analysis and Recognition, pages 705 - 708, 1999.

[5] Mithe R, Indalkar S, Divekar N. Optical character recog-nition. (2013) International Journal of Recent Technology and Engineering. 2013 Mar; 2(1).

[6] Smith R. (2007) An overview of the Tesseract OCR engine, USA: Google Inc; 2007.

[7] Zimmermann. M and H. Bunke. (2000) Automatic Segmentation of the IAM Off-line Database for Handwritten English Text. In Proc. of the 16th Int. Conf. on Pattern Recognition, Volume 4.