International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

CORTES (Characters Optically Recognized as Text and Speech)

Kedar Deshmukh[1], Mihir Yeole[2], Aditya Kshirsagar[3], Sarita Deshpande[4]

[1,2,3]*Student, Dept. of Information Technology, P.E.S's Modern College of Engineering, Pune, Maharashtra, India*
[4] *Head of Dept., Dept. of Information Technology, P.E.S's Modern College of Engineering, Pune, Maharashtra, India*

Abstract - *According to Dale's Cone of Experience, Human Beings, after 2 weeks, recall only 10 percent of what they read, but 50 percent of what they hear and see. Audio learning is a portal into new subjects and areas of expertise, and it is useful for getting the broader strokes of an idea. We aspire to develop a tool/technique that enables book enthusiasts and educational facilities around the world to recognize handwritten text on paper with minimal errors to present important non-digital information in the form of textual and voice output. We use Fully Convolutional Neural Networks (CNN) to implement the Intelligent Character Recognition phase and Text to Speech API for its conversion to Speech, since Deep Learning approaches have recently demonstrated efficient performances using recurrent neural networks. CORTES would help Educational Facilities, Authors, and Book Readers all over the world to convert handwritten notes into the downloadable digital text as well as audio output formats. This tool with these integrated features will be deployed as a web application, which would allow it to be used across multiple platforms and devices.*

Key Words: **OCR, Optical Character Recognition, Handwriting Recognition, CNN, Convolutional Neural Networks, Text to Speech, Deep Learning, Text Detection, Image Processing.**

1. INTRODUCTION

We aim, through the means of this project, to develop an application that enables book enthusiasts and educational facilities around the world to recognize handwritten text on paper with minimal errors and high accuracy to present important non-digital information in the form of textual and voice output.

Authors, may they be aspiring to be one or are already well-acclaimed generally prefer the pen and paper method to put down their creativity in the form of words. When the eventual first draft is completed, they need feedback from their acquaintances, friends or family members. The bit of a hassle here seems to be that people will usually find it difficult to go through the process of recognizing someone's handwritten points and try to make sense of it, or the other situation will be the author himself will have to spend his valuable time in converting this data into a more readable format. Visually impaired individuals also have to compromise on the situation which is, they can only perceive information through the means of braille.

A lot of data retrieved throughout history has shown the world the importance and impact of handwritten data. Papers however in the long term are perishable and no means of backup is available when information is stored in this manner. Researchers throughout history have shown tendencies of achieving breakthroughs in various fields through the means of scribbling and jotting down key points. With the already existing problem of every single person having distinct handwriting, the fact that papers break down over time has made it difficult for us to decipher important information. This has thus made us realize the need to preserve handwritten information in digital format in the new age of technology.

1.1 Application Specifications

The application aims to serve the following disciplinary areas:

- Book Readers.
- Aspiring Authors.
- Primary to University Level Education (Assignments and Notes).
- Those who wish to learn the English language in terms of pronunciation.
- Everyday Life: Letters, Diaries, Minutes of a meeting, etc.
- Banking Sectors.

CORTES (abbreviation: **C**haracters **O**ptically **R**ecognized as **TE**xt and **S**peech) is a web-based application that can thus provide its services to these disciplines. English being a widely recognized language in the world, foreigners should be able to use the app to use their handwritten notes that they have made themselves or those that have been provided to them to work on their pronunciation and accents.

The user will have to navigate to the website in order to avail these features. From their side, only a photograph of the handwritten English language will serve as an input, and the procedure that follows it will be handled by the application itself. The output will be a PDF file directly downloadable from the website, with the option to convert it to speech on the website itself. CORTES will use means of OCR (Optical Character Recognition) using Fully Convolutional Neural Networks (Deep Learning) with Text To Speech Synthesis in order to ensure today's upcoming technology helps reform the traditional means of conveying information along with its storage and retrieval.

1.2 Requirement Specifications

- Ability to scan unique handwritten characters as opposed to printed text in Traditional OCR systems.
- Ability to scan the English language characters handwritten on paper that includes (A-Z), (a-z) and Numeric values (0-9) and Punctuation marks.
- The automatic and accurate distinction between ambiguous figures such as the way '9' and 'g' can be written in a similar manner.
- Proper sentence formations in the digital text produced, dependent on the provision of punctuation marks by the user to whom the scanned image of paper belongs.
- Ability to download the file and store it on the user's system for future use and backup, thereby discouraging the need to preserve paper.
- Voice output feature on the web browser itself in various accents of the English language such as Russian, Indian, German, Arabic, etc. using a Web Browser's HTML5 feature.
- Cross-platform device compatibility.
- Easy to use, visually appealing GUI with adequate instructions to guide the user throughout the entire process.
- Eliminating the need for a Multisyn unit selection approach and the need for building a TTS Synthesizer from scratch due to the use of well-developed and researched API. [6]
- Future provisions of Spell checkers and Word Suggestions that can allow a user to improve their content.

2. DESIGN AND ARCHITECTURE

CORTES will undergo the following modules step by step in order to achieve its objectives:

I. IMAGE PRE-PROCESSING (OCR)

II. CNN APPLICATION AND CREATION OF TEXT FROM HANDWRITTEN SOURCE (POST-PROCESSING)

III. TEXT TO SPEECH SYNTHESIS

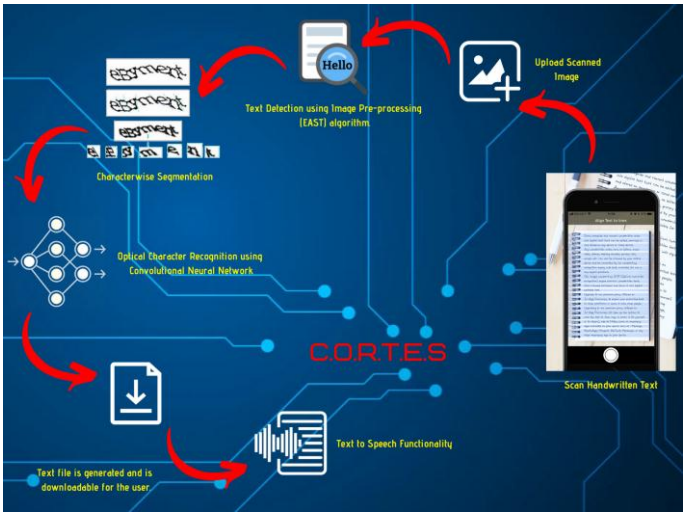


Fig -1: Design and Flow of the Application

3. IMAGE PRE-PROCESSING

This is the first step in Optical Character Recognition. An Image for us is filled with information that can be perceived by a human easily just by seeing it. However, for a Machine, Images are just 2-Dimensional arrays of numbers, and for extracting information from this array, complex mathematical algorithms are required.[8]

When we take an image, it has an innumerable amount of impurities in the form of blurry edges and due to the noise from the sensors. This makes it difficult for the algorithms to detect edges. In OCR, the main problem is to detect written words in an image full of information. Various algorithms have been devised for the detection of text in an Image like RATD algorithm and EAST algorithm.

Even if we manage to detect the text, the problem is that the Convolutional Neural Networks cannot detect whole words by itself. That is because of the sheer number of words, we cannot train the algorithm to know each word. So we also need to divide these words character by character.

The recent proliferation of cheap digital cameras in today's time and the fact that nearly every smartphone now has a camera, has led to concerns with the conditions that the image was captured under, and furthermore, what assumptions we can and cannot make. The summarized version of the challenges that need to be overcome are: [2,3]

- Noisy Images: Sensor noise from a handheld camera is typically higher than that of a traditional scanner. Low-priced cameras can also interpolate the pixels of raw sensors in an attempt to produce real colors.
- Blurring: Uncontrolled environments tend to have blurs, especially if the end-user is utilizing a smartphone that does not have some form of stabilization.

- Lighting conditions: Assumptions regarding lighting conditions in natural scene images cannot be made. It may be in the dark, the flash on the camera may be on, or the sun may be shining brightly, thus saturating the entire image.
- Resolution: Not all cameras are created equal and we may be dealing with cameras with sub-par resolution.

To factor in these problems, we first analyze images for light and dark areas in order to identify each alphabetic letter or a numeric digit. For that purpose, we need to provide preliminary image pre-processing. The pre-processing algorithm includes a few necessary steps:

3.1 Edge Enhancement Diffusion

Edge enhancement is an image processing filter that enhances the edge contrast of an (in this case: images) in an attempt to improve its apparent sharpness. The filter can identify sharp edge boundaries in the image, such as the edge between a subject and a background of a contrasting color, and increasing the image contrast in the area immediately around the edge. This has the effect of creating subtle bright and dark highlights on either side of edges present in the image, termed as overshoot and undershoot, leading the edge to look more defined. [3]

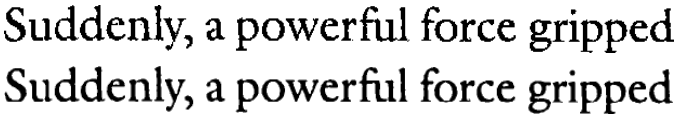


Fig -2: Before and After Edge Enhancement Diffusion

3.2 Image Binarization

Binarization is defined as the process of converting a pixel image into a binary image. A Grayscale image pixel value can vary from 0-255. Binarization of an image converts it to either 0 or 1. We implement Image Binarization using the OpenCV library method called Image Thresholding. It gives a binarized image by converting pixel values above the threshold value to 1 and those below to 0. [8]

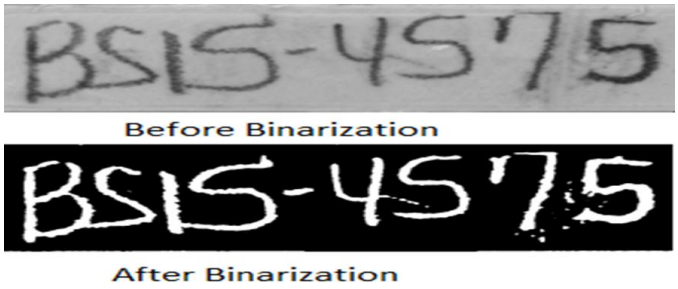


Fig -3: Binarization of Image

3.3 Segmentation

A good segmentation of characters is necessary for a high recognition accuracy. Segmentation of words into characters becomes very difficult because handwritten scripts have unconstrained nature and presence of cursive styling. Off-line handwritten word segmentation is a subject of much attention due to the presence of many difficulties such as: [2]

- Cursive nature of handwriting i.e. two or more alphabets in a word written connected to each other.
- Words may be written by a pen having ink of different colors.
- Some characters (e.g. 'u' and 'v') in a handwritten word image can have similar contours.

After the pre-processing of the input handwritten word image, the height and width of the word image are calculated for the analysis of the ligatures. The word image is scanned vertically, from top to bottom, column-wise and the number of foreground pixels in the inverted word image is counted in each column. The positions of all these columns are saved for which the sum of foreground black pixels is either 0 or 1. Many consecutive PSC is present in various groups in the whole word image where the sum of foreground pixels is 0 or 1. This situation can be termed as over-segmentation. [2]

When there is a clear vertical space between two consecutive characters in a word image, the problem of over-segmentation is eliminated by taking an average of all the PSC (Potentially Segmented Column) present in that area as the sum of foreground pixels for all these PSC columns is 0. The threshold value is supposed to be the minimum distance along the width of the word between consecutive PSCs and is chosen such that its value must be less than the width of the thinnest character possible in a word image. By experimenting several times, the value of the threshold is set to a value, e.g. 7. This means that all those PSC's which are separated by a distance of 7 pixels or less by another PSC will be merged to a single SC (Segmentation Column). [1,2]

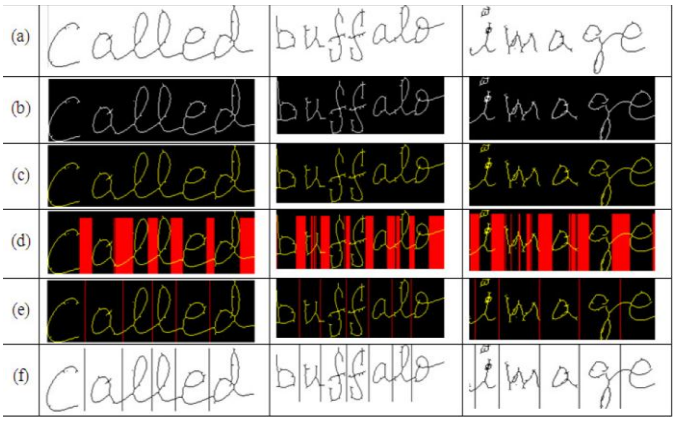


Fig -4: Character Segmentation [2]

4. CNN PHASE

The next step after Optical Character Recognition is the usage of Convolutional Neural Network (CNN, or ConvNet), which is a class of deep neural networks most commonly applied for analyzing visual imagery. The architecture of deep neural networks comes under the broader family of machine learning known as 'Deep Learning'.

Deep learning is a machine learning technique that is used to convey to the computers for doing what comes naturally to humans: 'learning by example'. In deep learning, a computer model learns to perform classification tasks directly from media such as images, text, or sound. Models are trained by using a large set of labeled data and neural network architectures that contain multiple layers. This provides us with an opportunity to use this feature for the prediction of the handwritten character that has been passed to a neural network from the Image Pre-processing phase. One of the most popular types of deep neural networks is known as the Convolutional Neural Network, which can serve to be a major component in Text Recognition. [1,4]

A CNN convolves learned features with input data, and uses 2D convolutional layers, thus making this architecture well suited and reliable to process 2D data, such as images for our scenario. CNN has its usage based in image recognition and processing, and is specifically designed to process pixel data. It makes the explicit assumption that the inputs are images, which allows us to encode certain properties into the architecture. These then make the forward function more efficient, vastly reducing the number of parameters in the network

The layers of a CNN consist of an Input layer, an Output layer and a Hidden layer that includes multiple Convolutional layers, ReLU layers, Pooling layers and Fully Connected layers. [9]

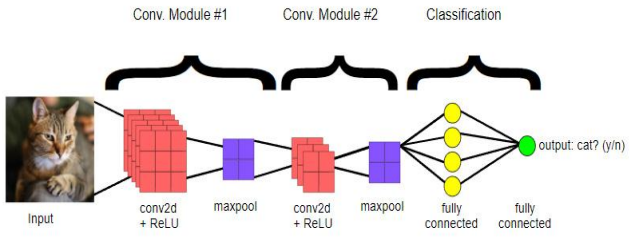


Fig -5: Convolutional Neural Network

5. TEXT TO SPEECH SYNTHESIS

The successful execution of Pre-Processing and CNN Layers will result in the creation of a downloadable PDF file containing the converted handwritten source text in a digital format. The user can now move on to its conversion into speech as well. Voice synthesis, defined as TTS (an acronym for Text-To-Speech), is a computer system that should be able to read aloud any text, regardless of its origin and primarily aims to produce human-voice in an artificial manner. [5] A TTS synthesizer is a computer-based system that should generate artificial speech waveforms from another data format. [5]

Previous research deals with the realization of a speech synthesis system for languages using the multisyn unit selection approach developed in Java working on the basis of: [6]

1. An NLP (Natural Language Processing) module responsible for the production of the phonetic transcription of an input text and, in some cases, the generation of prosody (intonation, duration, intensity or power). [6]
2. A DSP (Digital Signal Processing) module, which converts the data output from the NLP module into speech. [6]

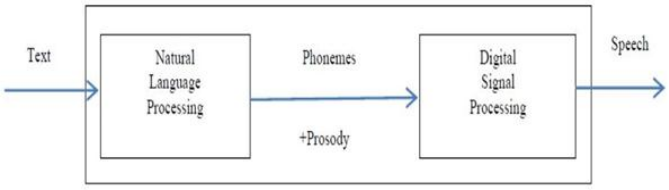


Fig -6: Working behind a TTS Synthesizer

Our approach, however, involves web application integration which provides us with much more flexibility and functionalities. This is accomplished by the means of a Text to Speech API, which more feasible to use instead of building a Text-to-Speech Synthesizer from scratch and demonstrates it's usage for indirectly enabling voice for handwritten text.

5.1 ResponsiveVoice.org

ResponsiveVoice.org is an organization that provides 'on-the-fly' text-to-speech functionality to websites by providing their API to developers who wish to integrate it into their applications. HTML5 introduces the Speech API for Speech Synthesis and Speech Recognition. Text To Speech (TTS) is now available in most modern browsers.[7] It uses Smart Chunking with large blocks of text. Preference is given to splitting at a full stop, question mark, colon or semicolon, and after that split is performed by the nearest comma and falling back from the nearest space between words. [7]

Some basic features enabled by using the API: [7]

- cancel(): Stops playing the speech.
- voiceSupport(): Checks if the browser supports native TTS.
- setDefaultVoice(): Detects if native TTS or TTS audio element is producing output.

- pause() and resume()
- Welcome message on page load.
- Enable speech by highlighting onscreen text.
- Multiple accents available for the English language.

6. IMPLEMENTATION

6.1 Image Pre-Processing for Uploaded Image

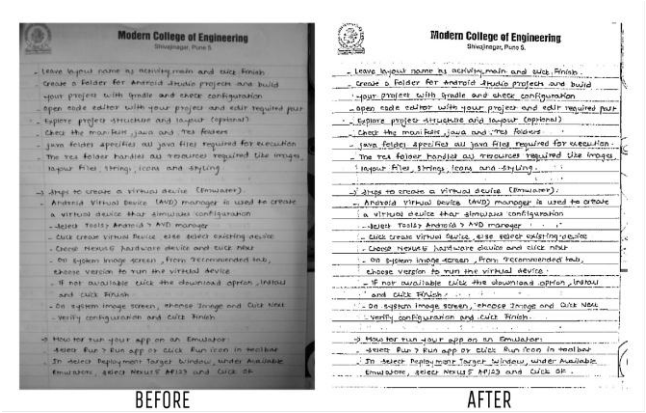


Fig -7: Pre-Process Testing for better Character Visibility

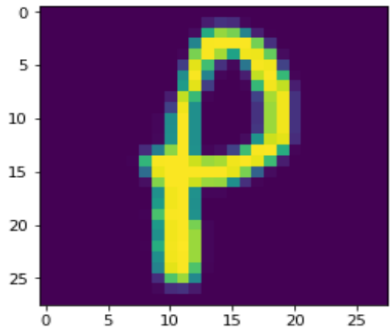
6.2 CNN Model

For the Convolutional Neural Network module, the EMNIST-character database was chosen due to its reputation of being a popular handwritten character database. [4] The neural network model was created using the Keras library in Python, which uses a TensorFlow backend. During compilation of the Keras model, the 'Adam' Optimizer seemed to converge well, and the 'Cross-Entropy' loss was used.

The neural network model can be tested by using the images provided in the EMNIST database. Users can select any index pertaining to the bounds of the data and can see the image for themselves. A predict function is used to pass this image to the neural network which then classifies the input image into an alphabet.

```
ind = 2

showImage(X[ind])
```




```
predict(X[ind])
```

```
'p'
```

Fig -8: CNN predicts handwritten character 'p' from the image passed to it from EMNIST Database.

7. RESULTS AND EVALUATION

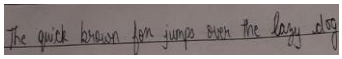
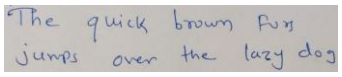
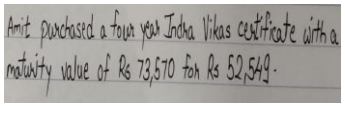
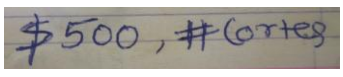
Handwritten Sample Conversion using CORTES		
Text Photo	Converted Text	Analysis
	The quick brown for jumps over the lazy dog	Minor anomalies due to unclear difference between 'x' and 'r'
	The quick brown furs Jumps over the lazy dog	Inaccurate conversion due to ambiguous alphabets and style of writing.
	Amit purchased a four year Indra Vikas certificate with a maturity value of RS 73, 570 for RS 52, 549.	Successful Numerical Value Detection.
	$ 500 # Cortes	Special Character Detection.

Table -1: Some Handwritten samples and their conversions

8. CONCLUSION

CORTES is a web application built using the Flask Micro-framework in Python along with Bootstrap 4.0 to ensure an intuitive GUI and cross-device and platform compatibility. The application is successfully able to convert handwritten text from a single user uploaded image into a PDF file containing digital text for the user, with the added functionality of Text-to-Speech that also has provisions of different accents.

CORTES gets us one step closer to reform the means of information storage, retrieval and conversion, while keeping in mind the abstraction of backend technologies to offer a seamless and easy experience for the end-user. It also offers an open end towards future enhancements since as spell-checker integrations and eventually achieving even higher accuracies for ambiguous & unique handwritings.

REFERENCES

[1] Raymond Ptuchaa, Felipe Petroski Sucha, Suhas Pillai, Frank Brockler, Vatsala Singh, Paul Hutkowski. "Intelligent character recognition using fully convolutional neural networks". Elsevier. April 2019. https://doi.org/10.1016/j.patcog.2018.12.017

[2] Amit Choudhary, Rahul Rishib, Savita Ahlawat. "A New Character Segmentation Approach for Off-Line Cursive Handwritten Words". Elsevier. https://doi.org/10.1016/j.procs.2013.05.013

[3] Sachin Kumar S, Parvathy Rajendran, Prabaharan P, K P Soman. "Text/Image Region Separation for Document Layout Detection of Old Document Images using Non-linear Diffusion and Level Set". Elsevier. https://doi.org/10.1016/j.procs.2016.07.235

[4] Tao Wang, David J. Wu, Adam Coates, Andrew Y. Ng. "End-to-End Text Recognition with Convolutional Neural Networks". IEEE 2012.

[5] Nwakanma Ifeanyi, Oluigbo Ikenna, Okpala Izunna. "Text – To – Speech Synthesis (TTS)". International Journal of Research in Information Technology (IJRIT).

[6] Théophile K. Dagba, Charbel Boco. "A Text To Speech system for Fon language using Multisyn algorithm". Elsevier. https://doi.org/10.1016/j.procs.2014.08.125

[7] ResposiveVoice Text – To - Speech. https://responsivevoice.org/api

[8] Ivo Vynckier. *A Close Look at Optical Character Recognition*. http://www.how-ocr-works.com/OCR/OCR.html

[9] Sumit Saha. *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way.* Towards Data Science. 15th Dec 2018. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

BIOGRAPHIES

Kedar Deshmukh
Student,
Department of Information Technology, P.E.S.'s Modern College of Engineering, Pune, Maharashtra, India.

Mihir Yeole
Student,
Department of Information Technology, P.E.S.'s Modern College of Engineering, Pune, Maharashtra, India.

Aditya Kshirsagar
Student,
Department of Information Technology, P.E.S.'s Modern College of Engineering, Pune, Maharashtra, India.

Sarita D. Deshpande
Head of Department,
Department of Information Technology, P.E.S.'s Modern College of Engineering, Pune, Maharashtra, India.