A PROJECT REPORT
ON

IMAGE2SPEECH

TEXT RECOGNITION IN IMAGES AND CONVERTING RECOGNIZED TEXT TO SPEECH

Submitted in partial fulfillment of the requirementsFor

the degree of BACHELOR

OF TECHNOLOGY in
Computer Science And Engineering

Submitted by

Anmol Tyagi (1902310100016)

Vasu Chaudhary (190231010018)

Ajay Pal (1902310100007)

Rahul Singh (1902310100078)

Under the Guidance of

Prof. Hemant Bhardwaj
Department of Computer Science & Engineering

R.D. Engineering College, Technical Campus, Ghaziabad

Dr. A.P.J. Abdul Kalam Technical University, Lucknow (UP), India

DECLARATION

We hereby declare that this submission is our own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person normaterial which to a substantial extent has been accepted for the award of any other degree of theuniversity or other institute of higher learning, except where due acknowledgement has been in the text.

Signature:
Name: Anmol Tyagi
Roll No.: 1902310100016
Date: 13/05/2023

Signature:
Name: Vasu Chaudhary
Roll No.: 190231010018
Date: 13/05/2023

Signature:
Name: Ajay Pal
Roll No.: 1902310100007
Date: 13/05/2023

Signature:
Name: Rahul Singh

Roll No.: 1902310100078
Date: 13/05/2023

CERTIFICATE

This is to certify that the project entitled “**Image2Speech -** Text Recognition in Image and Converting Recognized Text to Speech” submitted by Anmol Tyagi (1902310100016), Vasu Chaudhary (190231010018), Ajay Pal (1902310100007), Rahul Singh (1902310100078) in the partial fulfillment of the requirements for award of Bachelor of Technology in Computer Science and Engineering from Dr. A.P.J. Abdul Kalam Technical University, U.P., Lucknow under my supervision. The project embodies result of original result of original work and studies carried out by the student’s their self and the contents of the project do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other university/Institution.

Project Guide
Mr. Hemant Bhardwaj
(Associate Professor, CSE Department)
RDEC, Ghaziabad

Date: 13/05/2023

ABSTRACT

This project consolidates the idea of Image Text to Speech Synthesizer (TTS) and Optical Character Recognition (OCR). This sort of framework assists visually impaired people byconnecting with computers successfully through vocal interface. Text-to-Speech conversion is astrategy that scans and reads 38+ languages and numbers that are in the image utilizing OCR method and transforming it to voices. This project implements two modules, voice processing module and image processing module. There are many techniques, for example, Edged Based method, connected component Method, texture Based Method, Mathematical Morphology Method is been utilized previously, yet they have some restrictions when estimated by exactness, f-score and review. Image Text is the content data installed or written in Image of various structure. Image text can be found in magazines, captured images, newspapers, banners and so on These image texts are exceptionally accessible these days and they are vital in addressing, describing and moving data which help people in communication, accessibility, making of new sorts of jobs, cost viability, efficiency, tackling issues, globalization and so on.

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report on the B. Tech Project undertaken during B. Tech Final Year. We owe a special debt of gratitude to our guide Prof. Hemant Bhardwaj, Department of CSE, R.D. Engineering College, Ghaziabad for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen the light of the day.

We also take the opportunity to acknowledge the contribution of Prof. Mohd. Vakil, Department of CSE, R.D. Engineering College, Ghaziabad, for his full support and assistance during the development of the project.

We also do not want to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not least, we acknowledge our friends for their contribution to the completion of the project.

Anmol Tyagi (1902310100016)
Vasu Chaudhary (190231010018)
Ajay Pal (1902310100007)
Rahul Singh (1902310100078)

TABLE OF CONTENTS

Page No.

Contents

INTRODUCTION…………………………………………………………………….…...i

DECLARATION ii

CERTIFICATE iii

ABSTRACT iv

ACKNOWLEDGEMENT... v

TABLE OF CONTENTS vi

LIST OF FIGURES viii

Chapter 1 Introduction 1

1.1. Motivation and Objective 1

1.2. Structure of Project Report 3

CHAPTER 2: 7

CHAPTER 3: 17

3.1. Artificial Intelligence 17

3.1. Machine Learning...18

3.1.Deep Learning...19

3.1. Text Recognition 21

3.2. Speech Synthesis 22

3.3.Google Trans 32

3.3.Grey Scale Images..34

3.3.Neural Network..36

3.3.Digital Image Processing 40

3.3.OCR...43

CHAPTER 4: 50

RESULT AND DISCUSSIONS..50

4.1 .Introduction 50

4.1.Methodology..50

4.1.Result and Discussion 51

CHAPTER 5: 52

CONCLUSIONS AND FUTURE SCOPE 52

5.1.Future Scope 53

REFERENCES 54

LIST OF FIGURES

Figure No.	Figure Name	Page No.
1	Text Recognition Steps	11
2	Classification of Text Detection Techniques	12
3	Structure of Project Report	15
4	A Simple Neural Network	45

ABBREVATIONS

Abbreviations	Full Form	Page No.
TTS	Text To Speech	10
MSER	Maximally Stable Extremal Regions	11
CNN	Convolution Neural Network	12
FPN	Future Pyramid Network	12
RNN	Recurrent Neural Network	12
PR	Public Relation	16
OCR	Optical Character Recognition	16
ANN	Artificial Neural Network	18
AI	Artificial Intelligence	26
NLP	Natural Language Processing	27
ML	Machine Learning	27
DSP	Digital Signal Processing	30

Chapter 1

Introduction

1.1. Motivation and Objective

Languages are the oldest way of communication between human beings whether they are in spoken or written forms. In the recent era, visual text in natural or manmade scenes might carry very important and useful information. Text that appears in images contains important, useful and rich semantic information. This information is of great value for image interpretation. Text localization and recognition of natural scene images are based on the analysis and identification of scanned documents and images. Therefore, the scientists have started to digitize these images, extract and interpret the data by using specific techniques, and then perform text-to-speech synthesis (TTS). It is done in order to read the information aloud for the benefit and ease of the user. Text extraction and TTS can be utilized together to help people with reading disabilities and visual impairment to listen to written information by a computer system. In this work, a novel text detection framework is proposed which is based on connected component analysis and MSER algorithms are employed for extraction of CCs, which are taken as letter candidates. CCs that are likely to be characters are selected on the basis of their geometric properties and stroke width variation. Afterwards, the selected objects are grouped into detected text sequences, which are then fragmented into isolated words. Optical character recognition isemployed to recognize and extract the words and finally the extracted text is converted to appropriate speech using text-to-speech synthesizer. The proposed algorithm is tested on images representing different scenes ranging from documents to natural scenes. Promising results have been reported which prove the accuracy and robustness of the proposed algorithm and encourage its practical implementation in real world scenarios. Reading text from natural images is a challenging problem mainly in complex backgrounds. Deep learning is a type of machine learning and it is a main part of data science including statistics and predictive modeling. A neural network is a series of algorithms that are designed to recognize different patterns. It seeksto identify the underlying relationships in a set of data through a process that mimics the way the human brain works. It currently provides excellent solutions to many problems such as speech recognition, image recognition and natural language processing.

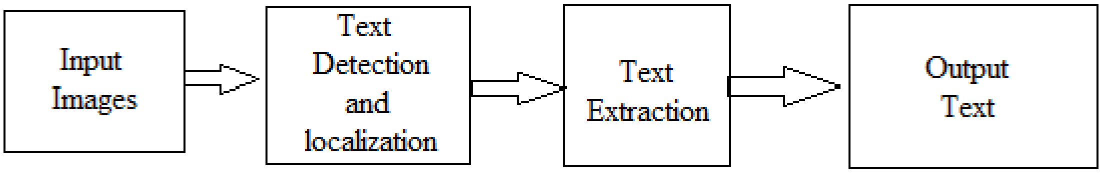

Fig. 1. Text Recognition Steps

Image processing is a method of performing certain actions on an image to obtain a modified image or to extract some meaningful information from images. Different image processingmethods like morphological operations and MSER (Maximally Stable Extremal Regions) detectors are used for text extraction from scene images. There are many applications for locating and recognizing text from images, such as vehicle number plate recognition, keyword based image exploration, objects recognition, and visually impaired assistance.

The text recognition process is mainly divided into the following steps (fig. 1)

- Input Image: It is the natural scene image contains text with complex backgrounds.
- Text Detection and Localization: Text detection means detecting the text in input image and localization means locating the text areas by eliminating the background regions.
- Text Extraction: The detected text in natural scene images is extracted to words or strings.
- Output Text: Output is in the form of words, strings or characters.

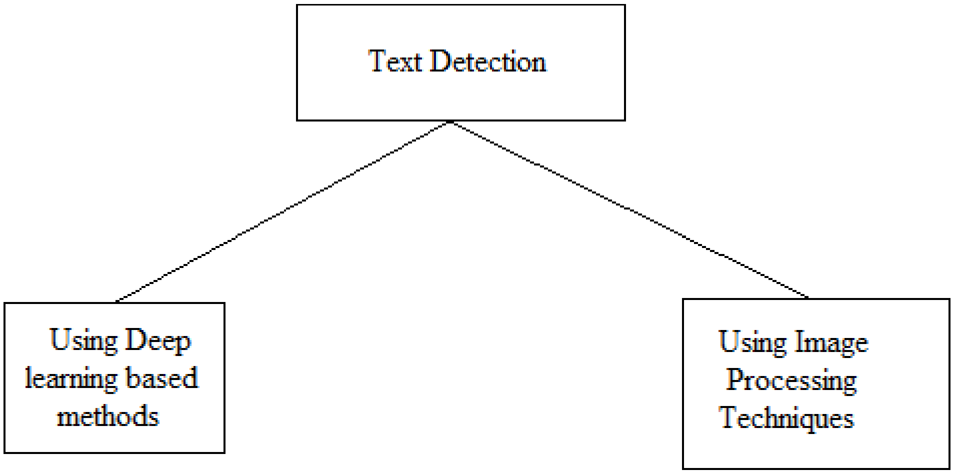
Text Detection
Using Deep learning based methods
Using Image Processing Techniques

Fig. 2. Classification of Text Detection Techniques

For better comparative study, the papers considered in this survey are mainly divided into two categories that is the text detection methods are categorized into Neural Network in deep learning based methods and Image processing based methods (fig. 2).

• Neural Networks in Deep Learning: Different deep learning based methods are used for text detection from natural scene images. Convolutional Neural Networks (CNN), Feature Pyramid Networks (FPN) and Recurrent Neural Networks (RNN) are some examples.

• Image Processing based methods: Image binarization, morphological operations like erosion and dilation, MSER detector are some image processing methods used for detection and extraction of text from natural scene images.

1.2. Structure of Project Report

The entire report can be divided into three broad sections that are mentioned below:

1) Pre-factory information

a) Title Page

- This includes the title of the report, name and address of the organization that is conducting the research.
- The name of the client to whom the report is to be submitted.
- And the date of submission

b) Declaration

- It states that the project group has submitted the results of their own thought, research or self-expression and there is no plagiarism.

c) Certificate

- The certificate states that the work has been carried out by this project group andhas not been submitted by any other project group of the institute for award or degree.

d) Acknowledgments

- Acknowledgments provide a way to thank those who supported or encouragedyou in research, writing and other parts of developing reports and paper.

e) Table of Contents

- This covers the list of all the topics with their page numbers.

f) Abstract

- It is a very important part of this section which summarizes the problem, research design and the major findings and conclusions.

2) Main Body

a) Introduction

- This chapter discusses the introduction, motivation and need of a project in thereal world.

b) Literature Survey

- This chapter discusses all the past research work related to the field the projecthas been made and mentions the working, shortcomings, improvements etc. ofpast research works.

c) Fundamentals of Text Recognition and Speech

- This chapter discusses about the project made by team members including working, basics, implementation, designing etc.

d) Results and Discussions

- This chapter discusses the conclusions and research paper published while makingthe project.

e) Conclusions and Future Scope

- This chapter interprets the conclusion and future scope that need to be done.

3) End Section

a) References

- A references documents the sources used by researchers in writing the report.

b) Certificates

- It contains Certificates that the researchers have got for their research paper in the course of there study.

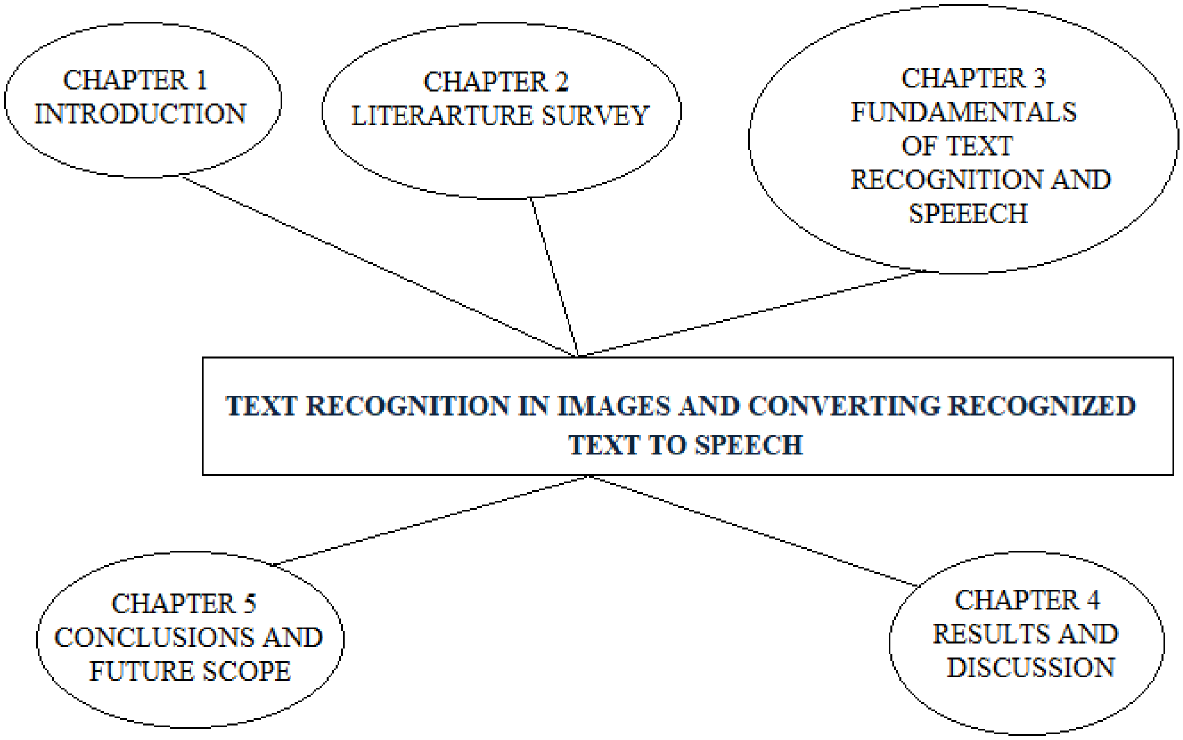

Fig. 3. Structure of Project Report

CHAPTER 2:

Literature Survey

In [1] Author proposed that Detection and recognition of text in images and video frames plays an important role, to access images and videos from multimedia data base containing huge amount of imagery and video data. This paper gives a brief review of methods and techniques used in the detection, recognition and analysis of text in images and video frames. This paper also highlights the various challenges facing in detecting and extracting text and the application of text in various fields. Text may suffer from various complex problems, for detection and extraction need to increase resolution and remove blur in images or video frames, so super resolution and deblurring techniques have been addressed. This paper gives information about data sets, compares result of existing techniques and gives brief idea to conduct research in future directions.

In [2] Mobile phone camera-based document video scanning is an interesting research problem which has entered into a new era with the emergence of widely used, processing capable and motion sensors equipped smartphones. We present our ongoing research on mobile phonecamera-based document image mosaic reconstruction method for video scanning of paperdocuments. In this work, we have optimized the classic key point feature descriptor-based imageregistration method, by employing the accelerometer and gyroscope sensor data. Experimental results are evaluated using optical character recognition (OCR) on the reconstructed mosaicfrom mobile phone camera-based video scanning of paper documents.

In [3] The purpose of this literature is to summarize the pattern recognition (PR) and deep learning (DL) artificial intelligence methods developed for the management of data in the lastsix years. The methodology used for the study of documents is a content analysis. For this study, 186 references are considered, from which 120 are selected for the literature review. First, a general introduction to artificial intelligence is presented, in which PR/DL methods are studied and their relevance to data management evaluated. Next, a literature review is provided of the most recent applications of PR/DL, and the capacity of these methods to process large volumes of data is evaluated. The analysis of the literature also reveals the main applications, challenges, approaches, advantages, and disadvantages of using these methods. Moreover, we discuss the main measurement instruments; the methodological contributions by study areas and research domain; and major databases, journals, and countries that contribute to the field of study. Finally, we identify emerging research trends, their limitations, and possible futureresearch paths.

In [4] Author is analyzes the problems of document image recognition and the existing solutions. Document recognition algorithms have been studied for quite a long time, but despite this, currently, the topic is relevant and research continues, as evidenced by a large number of associated publications and reviews. However, most of these works and reviews are devoted to individual recognition tasks. In this review, the entire set of methods, approaches, and algorithms necessary for document recognition is considered. A preliminary systematization allowed us to distinguish groups of methods for extracting information from documents of different types: single-page and multi-page, with text and handwritten contents, with a fixed template and flexible structure, and digitalized via different ways: scanning, photographing, video recording. Here, we consider methods of document recognition and analysis applied to a wide range of tasks: identification and verification of identity, due diligence, machine learning algorithms, questionnaires, and audits. The groups of methods necessary for the recognition of asingle page image are examined: the classical computer vision algorithms, i.e., key points, local feature descriptors, Fast Hough Transforms, image binarization, and modern neural network models for document boundary detection, document classification, document structure analysis, i.e., text blocks and tables localization, extraction and recognition of the details, post-processingof recognition results. The review provides a description of publicly available experimental datapackages for training and testing recognition algorithms. Methods for optimizing the performance of document image analysis and recognition methods are described.

In [5] author is proposed that device text line recognition framework that is designed for mobile or embedded systems. We consider per-character segmentation as a language independentproblem and individual character recognition as a language-dependent one. Thus, the proposed solution is based on two separate artificial neural networks (ANN)and dynamic programming instead of employing image processing methods for the segmentation step or end-to-end ANN. To satisfy the tight constraints on memory size imposed by embedded systems and to avoid overfitting, we employ ANNs with a small number of trainable parameters. The primary purpose of our framework is the recognition of low-quality images of identity documents with complex backgrounds and a variety of languages and fonts.

In [6] Author uses of automated document recognition has extended and as a result, recognition techniques that do not require specialized equipment have become more relevant. Among such techniques, document recognition using mobile devices is of interest. However, it is not always possible to ensure controlled capturing conditions and, consequentially, high quality of input images. Unlike specialized scanners, mobile cameras allow using a video stream as an input, thus obtaining several images of the recognized object, captured with various characteristics. In this case, a problem of combining the information from multiple input frames arises. In this paper, we propose a weighing model for the process of combining the per-frame recognition results, two approaches to the weighted combination of the text recognition results, and two weighing criteria. The effectiveness of the proposed approaches is tested using datasets of identity documents captured with a mobile device camera in different conditions, including perspective distortion of the document image and low lighting conditions. The experimental results show that the weighting combination can improve the text recognition result quality in the video stream, and the per-character weighting method with input image focus estimation asa base criterion allows one to achieve the best results on the datasets analyzed.

In [7] literature having based on hard AI problems, CAPTCHA(Completely Automated Public Turing test to tell the Computers and Humans Apart) is a hot research topic in the field of computer vision and artificial intelligence. CAPTCHA is a challenge-response test conducted to single out humans and bots.

It is ubiquitously implemented on the web since its introduction. As text-based CAPTCHAs are successfully broken by various researchers therefore several design variants have been proposed and implemented in order to further strengthen it. Animated Text-based CAPTCHAsare one of the design variant of it and are based on the difficulty of reading the moving text. They are based on zero knowledge per frame principle. Although it’s still easy for humans toread animated text but it’s a challenge for machines. As proposals for animated CAPTCHAs are on the rise so there is a strong need to scrutinize their strength against automated attacks. In this research, such CAPTCHAs are investigated to verify their robustness against automated attacks.

The proposed methods proved that these CAPTCHAs are vulnerable and they do not guarantee the robustness against automated attacks. The proposed frame selection, noise removal, segmentation and recognition methods have successfully decoded these CAPTCHAs with an overall precision, segmentation accuracy and recognition rate of up to 53.8%, 92.9% and93.5%respectively.

In [8] Author present a learning-free method for text line segmentation of historical handwritten document images. This method relies on automatic scale selection together with second derivative of anisotropic Gaussian filters to detect the blob lines that strike through the textlines. Detected blob lines guide an energy minimization procedure to extract the text lines. Historical handwritten documents contain noise, heterogeneous text line heights, skews and touching characters among text lines. Automatic scale selection allows for automatic adaption to the heterogeneous nature of handwritten text lines in case the character height range is correctly estimated. In the extraction phase, the method can accurately split the touching characters among the text lines. We provide results investigating various settings and compare the model with recent learning-free and learning-based methods on the c BAD competition dataset.

In [9] Author is proposed that robust algorithms for character segmentation and recognition are presented for multilingual Indian document images of Latin and Devanagari scripts. These documents generally suffer from their layout organizations, local skews, and low print quality and contain intermixed texts(machine-printed and handwritten). In the proposed character segmentation algorithm, primary segmentation paths are obtained using structural property of characters, whereas overlapped and joined characters are separated using graph distance theory. Finally, segmentation results are validated using highly accurate support vector machine classifier. For the proposed character recognition algorithm, three new geometrical shape-based features are computed. First and second features are formed with respect to the center pixel of character, whereas neighborhood information of text pixels is used for the calculation of third feature.

For recognizing the input character, k-Nearest Neighbor classifier is used, as it has intrinsically zero training time. Comprehensive experiments are carried out on different databases containing printed as well as handwritten texts. Benchmarking results illustrate that proposed algorithms have better performances compared to other contemporary approaches, where highest segmentation and recognition rates of 98.86%and 99.84%, respectively, are obtained.

In [10] Author have focused their attention on improving accuracy that provide significant advances. However, if they were limited to classification tasks, nowadays with contributions from Scientific Communities who are embarking in this field, they have become very useful in higher level tasks such as object detection and pixel-wise semantic segmentation. Thus, brilliant ideas in the field of semantic segmentation with deep learning have completed the state of the art of accuracy, however this architectures become very difficult to apply inembedded systems as is the case for autonomous driving. We present a new Deep fully Convolutional Neural Network for pixel-wise semantic segmentation which we call Squeeze- Seg Net. The architecture is based on Encoder-Decoder style. We use a Squeeze Net-like encoder and a decoder formed by our proposed squeeze-decoder module and up sample layer using down sample indices like in Seg Net and we add a deconvolution layer to provide final multi- channel feature map. On datasets like Cam vid or City-states, our net gets Seg Net-level accuracywith less than 10 times fewer parameters than Seg Net.

In [11] Natural image has many features, and the text in natural scene image has different meanings. In complex environments such as light, low visual acuity, dim fonts, font distortion, and multiple colors, it is very difficult to distinguish text from a natural scene image. Thispaper gives a survey on different text recognition methods in complex backgrounds. Different types of methods are used to extract text from complex natural scenes. This survey also includesa comparative study of different text recognition methods based on the accuracy and data sets used. There are mainly two types of classification used in this paper: image processing-based methods and deep learning-based methods. This comparative study is helpful for beginners in research fields.

In [12] Text recognition plays an important role in recognizing texts presented in the images as they provide important information. Scene text recognition has been an active research topic with rapid growth of development to improve the performance of text recognition with better reliability and accuracy. However, scene text recognition is challenging due to images containing inconsistent lighting, low resolution and blurriness. In addition, scene texts areusually taken from outdoor signboards, signage and road signs, which contain various orientation and fancy font styles to attract attention. Various researchers have proposed methodsfor recognizing different orientations of scene texts, such as horizontal texts, curved texts and rotated texts. However, to data there is a lack of research in recognizing vertical texts in natural scene images. In this research, a model for effective automatic recognition of vertical texts in natural scene images has been proposed, consisting of two major processes which are text localization and segmentation and text recognition. This proposed model recognizes threedifferent types of vertical scene texts, which are top-to-bottom vertical texts, bottom-to-top vertical texts and horizontal-stacked vertical texts.

In [13] Author do a lot of research has been devoted to identity documents analysis andrecognition on mobile devices. However, no publicly available datasets designed for this particular problem currently exist. There are a few datasets which are useful for associated subtasks but in order to facilitate a more comprehensive scientific and technical approach to identity document recognition more specialized datasets are required. In this paper we present aMobile Identity Document Video dataset (MIDV-500) consisting of 500 video clips for 50 different identity document types with ground truth which allows to perform research in a wide scope of document analysis problems. The paper presents characteristics of the dataset and evaluation results for existing methods of face detection, text line recognition, and document fields data extraction. Since an important feature of identity documents is their sensitiveness as they contain personal data, all source document images used in MIDV-500 are either in public domain or distributed under public copyright licenses. The main goal of this paper is to presenta dataset. However, in addition and as a baseline, we present evaluation results for existing methods for face detection, text line recognition, and document data extraction, using the presented dataset.

In [14] Text detection in natural scene image is challenging due to text variation in size, orientation, color and complex background, contrast, and resolution. In this paper, we focus on the long text detection in complex background. In order to deal with multi-scale text variation and exploit the recognition result to enhance the detection performance, we propose a detection and verification model based on SSD and encoder-decoder network for scene text detection. First, we present a text localization neural network based on SSD, which incorporates a text detection layer into the standard SSD model and can detect horizontal texts, especially long and dense Chinese texts in natural scenes more effectively. Second, a text verification model based on the encoder-decoder network is designed to recognize and verify the initial detection results, in order to eliminate non-text areas that are falsely detected as text areas. A series of experiments have been conducted on our constructed horizontal text detection dataset, which is composed of the horizontal text images in ICDAR 2017 Competition on Reading Chinese Text in the Wild (RCTW 2017) and some scene images taken by cameras. Compared with previous approaches, experimental results show that our method as achieved the highest recall rate of 0.784 and competitive precision rate in text detection, indicating the effectiveness of our proposed method.

In [15] Author is proposed that scene text detection is to detect the position of a text in the natural scene, the quality of which will directly affect the subsequent text recognition. It plays animportant role in fields such as image retrieval and autopilot. How to perform multi-scale and multi-oriented text detection in the scene still remains as a problem. This paper proposes an effective scene text detection method that combines the convolutional neural network (CNN)and recurrent neural network (RNN). In order to better adapt to texts in different scales, feature pyramid networks (FPN) have been applied in the CNN part to extract multi-scale features of theimage. We then utilize bidirectional long–short-term memory (Bi-LSTM) to encode thesefeatures to make full use of the text sequence characteristics with the outputs as a series of text proposals.

The generated proposals are finally linked into a text line through a well-designed text connector, which can be flexibly adapted to any oriented texts. The proposed method is evaluated on three public datasets: ICDAR2013, ICDAR2015, and USTB-SV1K. For ICDAR2013 and USTB-1K, we have reached 92.5% and 62.6% F-measure, respectively. Our method has reached 72.8% F-measure on the more challenging ICDAR2015 which demonstrates the effectiveness of our method.

In [16] Image recognition and optical character recognition technologies have become anintegral part of our everyday life due in part to the ever-increasing power of computing and the ubiquity of scanning devices. Printed documents can be quickly converted into digital text files through optical character recognition and then be edited by the user. Consequently, minimal time is required to digitize documents; this is particularly helpful when archiving volumes of printed materials. This study demonstrates how image-processing technologies can be used in combination with optical character recognition to improve recognition accuracy and to improve the efficiency of extracting text from images. Two software systems are developed and tested during this study: a character recognition system applied to cosmetic-related advertising images and a text detection and recognition system for natural scenes. The results of the experiment demonstrate that the proposed systems can accurately recognize text in images.

In [17] In this paper, author introduce an ‘‘on the device’’ text line recognition framework thatis designed for mobile or embedded systems. We consider per-character segmentation as a language-independent problem and individual character recognition as a language-dependent one. Thus, the proposed solution is based on two separate artificial neural networks (ANN) and dynamic programming instead of employing image processing methods for the segmentation step or end-to-end ANN. To satisfy the tight constraints on memory size imposed by embedded systems and to avoid overfitting, we employ ANNs with a small number of trainable parameters.The primary purpose of our framework is the recognition of low-quality images of identity documents with complex backgrounds and a variety of languages and fonts. We demonstrate that our solution shows high recognition accuracy on natural datasets even being trained on purely synthetic data. We use MIDV-500 and Census 1961 Project datasets for text line recognition. The proposed method considerably surpasses the algorithmic method implemented in Tesseract 3.05, the LSTM method (Tesseract4.00), and unpublished method used in the ABBYY FineReader 15 system. Also, our framework is faster than other compared solutions. We show the language-independence of our segmented with the experiment with Cyrillic, Armenian, and Chinese text lines.

In [18] A novel robust natural text recognition network (RNTR-Net) is proposed based on a combination of convolutional neural network (CNN) (for feature extraction) and a recurrent neural network (RNN) (for sequence recognition). The pipeline design comprises an improved block of residual learning combined with a general residual block to extract feature maps. Two bidirectional Long Short Term Memory (LSTM) networks are used for sequence recognition, and a transcription layer is used for decoding.

The proposed network can handle text images suffering from distortion or other degradations. Compared with previous algorithms, we achieve superior results in general datasets, including the IIIT-5K, Street View Text and ICDAR datasets. Moreover, the performance of the presented network is either highly competitive or even state-of-the-art regarding the highly challenging SVT-Perspective and CUTE80 datasets. We obtain considerable performance of 84.7% and 62.6% on lexicon-free IIIT-5K and CUTE80 datasets, respectively.

In [19] Author proposed project framework is the implementation of the image catching technique in MATLAB software. Most low visual people using Braille for reading a book and text is difficult to make and limited easily available. The OCR can be simulated in MATLABfor conversion. The object of attention from the background or other objects in the camera view is effectively distinguishing.

In [20] In this paper, the ultrasonic sensor gives vibration sensing for visually impaired people to easily recognize the bus name and bus number in the bus stop with help of an audioprocessing algorithm.

In [21] Author do the Raspberry Pi 3 camera to capture the image, and tesseract OCR is the engine that extracts the recognized text. The good accuracy gives Tesseract and Festival with compared classical techniques of image processing and Optical character recognition.

In [22] Author proposed, the MATLAB using for the image is transformed to text and then the text is converted to speech. We can translate text from a document and generate synthesized through a computers speaker.

In [23] This paper using image acquisition, image preprocessing, TTS conversion for automatic text detection and recognition based on image processing techniques. The preprocessing is used for noise removal. The approved median filter using for noise removal, the nonlinear filter used due to its excellent denoising power and computational performance.

In [24] The image of text print is taken from camera. The raspberry pi is connected to camera and converts into text. The preprocessing, segmentation are using for noise removal. The audio amplifier using for amplified the speech output.

In [25] In this paper TTS synthesis a simple character to voice translation. The alphabets (a-z, A-Z) and digits (0-9) are recorded in the order of wave data (.wav) for the database. Each character has a unique pronunciation. We play the wave data corresponding to each character read, in character to voice translation, we can also play the wave data for each word read. Once the text is read, for every word the corresponding wave file are concatenated and played. The MOS value for a female is 4 and a male is 3.5. the delay is reduced and mismatching of words also reduce.

In [26] In this paper author proposed that the raspberry pi model for scanning the images from stereotypical forms such as bus numbers, store signs, product signs, and door numbers. The AdaBoost algorithm is used for processing the visual information converting into audio speech.

CHAPTER 3:
FUNDAMENTALS OF TEXT RECOGNITION & SPEECHSYNTHESIS

3.1.Artificial Intelligence

Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to intelligence of humans and other animals. Example tasks in which this is done include speech recognition, computer vision, translation between (natural) languages, as well as other mappings of inputs.

AI applications include advanced web search engines (e.g., Google Search), recommendation systems (used by YouTube, Amazon, and Netflix), understanding human speech (such as Siri and Alexa), self-driving cars (e.g., Waymo), generative or creative tools (Chat GPT and AI art), automated decision-making, and competing at the highest level in strategic game systems (such as chess and Go).

As machines become increasingly capable, tasks considered to require "intelligence" are often removed from the definition of AI, a phenomenon known as the AI effect. For instance, optical character recognition is frequently excluded from things considered to be AI, having become a routine technology.

Artificial intelligence was founded as an academic discipline in 1956 and in the years since it has experienced several waves of optimism, followed by disappointment and the loss of funding(known as an "AI winter"), followed by new approaches, success, and renewed funding. AI research has tried and discarded many different approaches, including simulating the brain, modeling human problem solving, formal logic, large databases of knowledge, and imitating animal behavior. In the first decades of the 21st century, highly mathematical and statistical machine learning has dominated the field, and this technique has proved highly successful, helping to solve many challenging problems throughout industry and academia.

The various sub-fields of AI research are centered around particular goals and the use of particular tools. The traditional goals of AI research include reasoning, knowledge representation, planning, learning, natural language processing, perception, and the ability to move and manipulate objects. General intelligence (the ability to solve an arbitrary problem) is among the field's long-term goals. To solve these problems, AI researchers have adapted and integrated a wide range of problem-solving techniques, including search and mathematical optimization, formal logic, artificial neural networks, and methods based on statistics, probability, and economics. AI also draws upon computer science, psychology, linguistics, philosophy, and many other fields.

The field was founded on the assumption that human intelligence "can be so precisely describedthat a machine can be made to simulate it". This raised philosophical arguments about the mind and the ethical consequences of creating artificial beings endowed with human-like intelligence;these issues have previously been explored by myth, fiction, and philosophy since antiquity. Computer scientists and philosophers have since suggested that AI may become an existential risk to humanity if its rational capacities are not steered towards beneficial goals. The term artificial intelligence has also been criticized for overhyping AI's true technological capabilities.Natural language processing (NLP) allows machines to read and understand human language. Asufficiently powerful natural language processing system would enable natural-language user interfaces and the acquisition of knowledge directly from human-written sources, such asnewswire texts. Some straightforward applications of NLP include information retrieval, question answering and machine translation.

Symbolic AI used formal syntax to translate the deep structure of sentences into logic. This failed to produce useful applications, due to the intractability of logic and the breadth ofcommonsense knowledge. Modern statistical techniques include co-occurrence frequencies (how often one word appears near another), "Keyword spotting" (searching for a particularword to retrieve information), transformer-based deep learning (which finds patterns in text), and others. They have achieved acceptable accuracy at the page or paragraph level, and, by 2019, could generate coherent text.

3.1. Machine Learning

Machine learning (ML) is a field of inquiry devoted to understanding and building methods that "learn" – that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of artificial intelligence.

Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, agriculture, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. A subset of machine learning isclosely related to computational statistics, which focuses on making predictions using computers, but not all machine learning is statistical learning. The study of mathematicaloptimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Some implementations of machine learning use data and neural networksin a way that mimics the working of a biological brain. In its application across business problems, machine learning is also referred to as predictive analytics.

3.1.Deep Learning

Deep-learning architectures such as deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks, convolutional neural networks and transformers have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection and board game programs, where they have producedresults comparable to and in some cases surpassing human expert performance.

Artificial neural networks (ANNs) were inspired by information processing and distributed communication nodes in biological systems. ANNs have various differences from biological brains. Specifically, artificial neural networks tend to be static and symbolic, while the biological brain of most living organisms is dynamic (plastic) and analog.

The adjective "deep" in deep learning refers to the use of multiple layers in the network. Early work showed that a linear perceptron cannot be a universal classifier, but that a network with a nonpolynomial activation function with one hidden layer of unbounded width can. Deep learning is a modern variation that is concerned with an unbounded number of layers ofbounded size, which permits practical application and optimized implementation, while retaining theoretical universality under mild conditions. In deep learning the layers are also permitted to be heterogeneous and to deviate widely from biologic.

Deep learning is closely related to a class of theories of brain development (specifically,neocortical development) proposed by cognitive neuroscientists in the early 1990s. These developmental theories were instantiated in computational models, making them predecessors of deep learning systems. These developmental models share the property that various proposedlearning dynamics in the brain (e.g., a wave of nerve growth factor) support the self- organization somewhat analogous to the neural networks utilized in deep learning models. Like the neocortex, neural networks employ a hierarchy of layered filters in which each layer considers information from a prior layer (or the operating environment), and then passes its output (and possibly the original input), to other layers. This process yields a self-organizing stack of transducers, well-tuned to their operating environment. A 1995 description stated,"...the infant's brain seems to organize itself under the influence of waves of so-called trophic- factors ... different regions of the brain become connected sequentially, with one layer of tissue maturing before another and so on until the whole brain is mature."

A variety of approaches have been used to investigate the plausibility of deep learning models from a neurobiological perspective. On the one hand, several variants of the backpropagation algorithm have been proposed in order to increase its processing realism.

Deep learning is a class of machine learning algorithms that[8]:199–200 uses multiple layers to progressively extract higher-level features from the raw input. For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces.

From another angle to view deep learning, deep learning refers to ‘computer-simulate’ or‘automate’ human learning processes from a source (e.g., an image of dogs) to a learned object (dogs). Therefore, a notion coined as “deeper” learning or “deepest” learning makes sense. The deepest learning refers to the fully automatic learning from a source to a final learned object. A deeper learning thus refers to a mixed learning process: a human learning process from a source to a learned semi-object, followed by a computer learning process from the human learned semi-object to a final learned object.

Large-scale automatic speech recognition is the first and most convincing successful case of deep learning. LSTM RNNs can learn "Very Deep Learning" tasks that involve multi-second intervals containing speech events separated by thousands of discrete time steps, where one time step corresponds to about 10 ms. LSTM with forget gatesis competitive with traditional speech recognizers on certain tasks.

The initial success in speech recognition was based on small-scale recognition tasks based on TIMIT. The data set contains 630 speakers from eight major dialects of American English, where each speaker reads 10 sentences. Its small size lets many configurations be tried. More importantly, the TIMIT task concerns phone-sequence recognition, which, unlike word- sequence recognition, allows weak phone bigram language models. This lets the strength of the acoustic modeling aspects of speech recognition be more easily analyzed. The error rates listed below, including these early results and measured as percent phone error rates (PER), have beensummarized since 1991.

3.1.Text Recognition

There are two basic types of core OCR algorithm, which may produce a ranked list of candidate character.

- Matrix matching involves comparing an image to a stored glyph on a pixel-by- pixel basis; it is also known as "pattern matching", "pattern recognition", or "image correlation". This relies on the input glyph being correctly isolated from the rest of the image, and on the stored glyph being in a similar font and at the same scale. This technique works best with typewritten text and does not work well when new fonts are encountered. This is the technique the early physical photocell-based OCR implemented, rather directly.
- Feature extraction decomposes glyphs into "features" like lines, closed loops, line direction, and line intersections. The extraction features reduces the dimensionality of the representation and makes the recognition process computationally efficient. These features are compared with an abstract vector- like representation of a character, which might reduce to one or more glyph

prototypes. General techniques of feature detection in computer vision are applicable to this type of OCR, which is commonly seen in "intelligent" handwriting recognition and indeed most modern OCR software. Nearest neighbour classifiers such as the k-nearest neighbors algorithm are used to compare image features with stored glyph features and choose the nearest match.]

Software such as Cuneiform and Tesseract use a two-pass approach to character recognition.The second pass is known as "adaptive recognition" and uses the letter shapes recognized with high confidence on the first pass to recognize better the remaining letters on the second pass. This is advantageous for unusual fonts or low-quality scans where the font is distorted (e.g. blurred or faded).

Modern OCR software include Google Docs OCR, ABBYY FineReader and Tran sym. Others like OCR opus and Tesseract uses neural networks which are trained to recognize whole lines of text instead of focusing on single characters.

A new technique known as iterative OCR automatically crops a document into sections based on page layout. OCR is performed on the sections individually using variable character confidence level thresholds to maximize page-level OCR accuracy. A patent from the United States Patent Office has been issued for this method

The OCR result can be stored in the standardized ALTO format, a dedicated XML schema maintained by the United States Library of Congress. Other common formats include OCR andPAGE XML.

3.2. Speech Synthesis

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a 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. The reverse process is speech recognition.

Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, thestorage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood clearly. An intelligible text-to-speech program allows people with visual impairments or reading disabilities to listen to written words on a home computer. Many computer operating systems have included speech synthesizers since the early 1990s.

A text-to-speech system (or "engine") is composed of two parts: a front-end and a back-end. The front-end has two major tasks. First, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words. This process is often called text normalization, pre-processing, or tokenization. The front-end then assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text- to-phoneme or grapheme-to-phoneme conversion. Phonetic transcriptions and prosody information together make up the symbolic linguistic representation that is output by the front- end. The back-end—often referred to as the synthesizer—then converts the symbolic linguistic representation into sound. In certain systems, this part includes the computation of the target prosody (pitch contour, phoneme durations), which is then imposed on the output speech.

Synthesizer technologies

The most important qualities of a speech synthesis system are naturalness and intelligibility. Naturalness describes how closely the output sounds like human speech, while intelligibility is the ease with which the output is understood. The ideal speech synthesizer is both natural and intelligible. Speech synthesis systems usually try to maximize both characteristics.

The two primary technologies generating synthetic speech waveforms are concatenative synthesis and formant synthesis. Each technology has strengths and weaknesses, and the intended uses of a synthesis system will typically determine which approach is used.

Concatenation synthesis

Concatenative synthesis is based on the concatenation (stringing together) of segments ofrecorded speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech. However, differences between natural variations in speech and the nature ofthe automated techniques for segmenting the waveforms sometimes result in audible glitches in the output. There are three main sub-types of concatenative synthesis.

Unit selection synthesis

Unit selection synthesis uses large databases of recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, diphones, half-phones, syllables, morphemes, words, phrases, and sentences. Typically, the division into segments is done using a specially modified speech recognizer set to a "forced alignment" modewith some manual correction afterward, using visual representations such as the waveform and spectrogram. An index of the units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency (pitch), duration, position in the syllable, and neighboring phones. At run time, the desired target utterance is created by determining the best chain of candidate units from the database (unit selection). This process is typically achieved using a specially weighted decision tree.

Unit selection provides the greatest naturalness, because it applies only a small amount of digital signal processing (DSP) to the recorded speech. DSP often makes recorded speech soundless natural, although some systems use a small amount of signal processing at the point of concatenation to smooth the waveform. The output from the best unit-selection systems is often indistinguishable from real human voices, especially in contexts for which the TTS system has been tuned. However, maximum naturalness typically require unit-selection speech databases tobe very large, in some systems ranging into the gigabytes of recorded data, representing dozens of hours of speech. Also, unit selection algorithms have been known to select segments from a place that results in less than ideal synthesis (e.g. minor words become unclear) even when a better choice exists in the database. Recently, researchers have proposed various automated methods to detect unnatural segments in unit-selection speech synthesis systems.

Diphone synthesis

Diphone synthesis uses a minimal speech database containing all the diphones (sound-to-sound transitions) occurring in a language. The number of diphones depends on the phonotactics of the language: for example, Spanish has about 800 diphones, and German about 2500. In diphone synthesis, only one example of each diphone is contained in the speech database. At runtime, the target prosody of a sentence is superimposed on these minimal units by means of digital signal processing techniques such as linear predictive coding, PSOLA or MBROLA. or more recent techniques such as pitch modification in the source domain using discrete cosine transform. Diphone synthesis suffers from the sonic glitches of concatenative synthesis and the robotic-sounding nature of formant synthesis, and has few of the advantages of either approach other than small size. As such, its use in commercial applications is declining,[citation needed] although it continues to be used in research because there are a number of freely available software implementations. An early example of Diphone synthesis is a teaching robot, that was invented by Michael J. Freeman contained information regarding class curricular and certain biographical information about the students whom it was programmed to teach. It was tested in a fourth grade classroom in the Bronx, New York.

Domain-specific synthesis

Domain-specific synthesis concatenates prerecorded words and phrases to create completeutterances. It is used in applications where the variety of texts the system will output is limitedto a particular domain, like transit schedule announcements or weather reports. The technology is very simple to implement, and has been in commercial use for a long time, in devices like talking clocks and calculators. The level of naturalness of these systems can be very high because the variety of sentence types is limited, and they closely match the prosody and intonation of the original recordings. Because these systems are limited by the words and phrases in their databases, they are not general-purpose and can only synthesize the combinations of words and phrases with which they have been preprogrammed. The blending ofwords within naturally spoken language however can still cause problems unless the many variations are taken into account. For example, in non-rhotic dialects of English the "r" in wordslike "clear" is usually only pronounced when the following word has a vowel as its first letter (e.g. "clear out" is realized as /). Likewise in French, many final consonants become no longer silent if followed by a word that begins with a vowel, an effect called liaison. This alternation cannot be reproduced by a simple word-concatenation system, which would require additional complexity to be context-sensitive.

Formant synthesis

Formant synthesis does not use human speech samples at runtime. Instead, the synthesized speech output is created using additive synthesis and an acoustic model (physical modelling synthesis). Parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech. This method is sometimes called rules-based synthesis; however, many concatenative systems also have rules-based components. Manysystems based on formant synthesis technology generate artificial, robotic-sounding speech that would never be mistaken for human speech. However, maximum naturalness is not always the goal of a speech synthesis system, and formant synthesis systems have advantages over concatenative systems. Formant-synthesized speech can be reliably intelligible, even at very high speeds, avoiding the acoustic glitches that commonly plague concatenative systems. High- speed synthesized speech is used by the visually impaired to quickly navigate computers using ascreen reader. Formant synthesizers are usually smaller programs than concatenative systems because they do not have a database of speech samples. They can therefore be used in embedded systems, where memory and microprocessor power are especially limited. Because formant-based systems have complete control of all aspects of the output speech, a wide variety of prosodies and intonations can be output, conveying not just questions and statements, but a variety of emotions and tones of voice. Examples of non-real-time but highly accurate intonation control in formant synthesis include the work done in the late 1970s for the Texas Instruments toy Speak & Spell, and in the early 1980s Sega arcade machines and in many Atari,Inc. arcade games using the TMS5220 LPC Chips. Creating proper intonation for these projects was painstaking, and the results have yet to be matched by real-time text-to-speech interfaces.

Articulatory synthesis

Articulatory synthesis refers to computational techniques for synthesizing speech based on models of the human vocal tract and the articulation processes occurring there. The firstarticulatory synthesizer regularly used for laboratory experiments was developed at Haskins

Laboratories in the mid-1970s by Philip Rubin, Tom Baer, and Paul Mermelstein. This synthesizer, known as ASY, was based on vocal tract models developed at Bell Laboratories in the 1960s and 1970s by Paul Mermelstein, Cecil Coker, and colleagues.

Until recently, articulatory synthesis models have not been incorporated into commercial speech synthesis systems. A notable exception is the NeXT-based system originally developed and marketed by Trillium Sound Research, a spin-off company of the University of Calgary, where much of the original research was conducted. Following the demise of the various incarnations of NeXT (started by Steve Jobs in the late 1980s and merged with Apple Computer in 1997), the Trillium software was published under the GNU General Public License, with work continuing as gnu speech. The system, first marketed in 1994, provides full articulatory-based text-to-speech conversion using a waveguide or transmission-line analog of the human oral and nasal tracts controlled by "distinctive region model".

More recent synthesizers, developed by Jorge C. Lucero and colleagues, incorporate models of vocal fold biomechanics, glottal aerodynamics and acoustic wave propagation in the bronchi, trachea, nasal and oral cavities, and thus constitute full systems of physics-based speech simulation.

HMM-based synthesis

HMM-based synthesis is a synthesis method based on hidden Markov models, also called Statistical Parametric Synthesis. In this system, the frequency spectrum (vocal tract), fundamental frequency (voice source), and duration (prosody) of speech are modeled simultaneously by HMMs. Speech waveforms are generated from HMMs themselves based on the maximum likelihood criterion.

Sinewave synthesis

Sinewave synthesis is a technique for synthesizing speech by replacing the formants (main bands of energy) with pure tone whistles.

Deep learning-based synthesis

Deep learning speech synthesis uses deep neural networks (DNN) to produce artificial speech from text (text-to-speech) or spectrum (vocoder). The deep neural networks are trained using a large amount of recorded speech and, in the case of a text-to-speech system, the associatedlabels and/or input text. The DNN-based speech synthesizers are approaching the naturalness of the human voice.Examples of disadvantages of the method are low robustness when the data are not sufficient, lack of controllability and low performance in auto-regressive models.

For tonal languages, such as Chinese or Taiwanese language, there are different levels of tone sandhi required and sometimes the output of speech synthesizer may result in the mistakes of tone sandhi.

Audio deepfakes

This section is an excerpt from Audio deepfake.

The audio deepfake (also known as voice cloning) is a type of artificial intelligence used to create convincing speech sentences that sound like specific people saying things they did not say. This technology was initially developed for various applications to improve human life. Forexample, it can be used to produce audiobooks, and also to help people who have lost their voices (due to throat disease or other medical problems) to get them back. Commercially, it has opened the door to several opportunities. This technology can also create more personalized digital assistants and natural-sounding text-to-speech as well as speech translation services.

Audio deepfakes, recently called audio manipulations, are becoming widely accessible using simple mobile devices or personal PCs. These tools have also been used to spread misinformation using audio. This has led to cybersecurity concerns among the global public about the side effects of using audio deepfakes. People can use them as a logical access voice spoofing technique, where they can be used to manipulate public opinion for propaganda, defamation, or terrorism. Vast amounts of voice recordings are daily transmitted over the Internet, and spoofing detection is challenging. Audio deepfake attackers have targeted individuals and organizations, including politicians and governments. In early 2020, some scammers used artificial intelligence-based software to impersonate the voice of a CEO to authorize a money transfer of about $35 million through a phone call. Therefore, it is necessaryto authenticate any audio recording distributed to avoid spreading misinformation.

Text-to-speech systems

Text-to-speech (TTS) refers to the ability of computers to read text aloud. A TTS engine converts written text to a phonemic representation, then converts the phonemic representation towaveforms that can be output as sound. TTS engines with different languages, dialects and specialized vocabularies are available through third-party publishers.

Android

Version 1.6 of Android added support for speech synthesis (TTS).

Internet

Currently, there are a number of applications, plugins and gadgets that can read messages directly from an e-mail client and web pages from a web browser or Google Toolbar. Some specialized software can narrate RSS-feeds. On one hand, online RSS-narrators simplify information delivery by allowing users to listen to their favourite news sources and to convert them to podcasts. On the other hand, on-line RSS-readers are available on almost any personal computer connected to the Internet. Users can download generated audio files to portable devices, e.g. with a help of podcast receiver, and listen to them while walking, jogging or commuting to work. A growing field in Internet based TTS is web-based assistive technology,

e.g. 'Browsealoud' from a UK company and Readspeaker. It can deliver TTS functionality to anyone (for reasons of accessibility, convenience, entertainment or information) with access to aweb browser. The non-profit project Pediaphon was created in 2006 to provide a similar web- based TTS interface to the Wikipedia. Other work is being done in the context of the W3C through the W3C Audio Incubator Group with the involvement of The BBC and Google Inc.

Open source

Some open-source software systems are available, such as:

RHVoice with support for multiple languages. Festival Speech Synthesis System which uses diphone-based synthesis, as well as more modern and better-sounding techniques. Speak which supports a broad range of languages. GNU speech which uses articulatory synthesis from the FreeSoftware Foundation. MaryTTS, web based and open source.

Others

Following the commercial failure of the hardware-based Intellivoice, gaming developers sparingly used software synthesis in later games[citation needed]. Earlier systems from Atari, such as the Atari 5200 (Baseball) and the Atari 2600 (Quadrun and Open Sesame), also had games utilizing software synthesis. Some e-book readers, such as the Amazon Kindle, Samsung E6, PocketBookeReader Pro, enTourageeDGe, and the Bebook Neo. The BBC Micro incorporated the Texas Instruments TMS5220 speech synthesis chip, Some models of Texas Instruments home computers produced in 1979 and 1981 (Texas Instruments TI-99/4 and TI- 99/4A) were capable of text-to-phoneme synthesis or reciting complete words and phrases (text-to-dictionary), using a very popular Speech Synthesizer peripheral. TI used a proprietary codec to embed complete spoken phrases into applications, primarily video games. IBM's OS/2 Warp

4 included VoiceType, a precursor to IBM ViaVoice. GPS Navigation units produced byGarmin, Magellan, TomTom and others use speech synthesis for automobile navigation. Yamaha produced a music synthesizer in 1999, the Yamaha FS1R which included a Formant synthesis capability. Sequences of up to 512 individual vowel and consonant formants could be stored and replayed, allowing short vocal phrases to be synthesized.

Digital sound-alikes

At the 2018 Conference on Neural Information Processing Systems (NeurIPS) researchers from Google presented the work 'Transfer Learning from Speaker Verification to Multispeaker Text- To-Speech Synthesis', which transfers learning from speaker verification to achieve text-to- speech synthesis, that can be made to sound almost like anybody from a speech sample of only5 seconds. Also researchers from Baidu Research presented a voice cloning system with similar aims at the 2018 NeurIPS conference, though the result is rather unconvincing. By 2019 the digital sound-alikes found their way to the hands of criminals as Symantec researchers know of 3 cases where digital sound-alikes technology has been used for crime. This increases the stress on the disinformation situation coupled with the facts that Human image synthesis since the early 2000s has improved beyond the point of human's inability to tell a real human imaged with a real camera from a simulation of a human imaged with a simulation of a camera. 2D video forgery techniques were presented in 2016 that allow near real-time counterfeiting of facial expressions in existing 2D video. In SIGGRAPH 2017 an audio driven digital look-alike of upper torso of Barack Obama was presented by researchers from University of Washington.It was driven only by a voice track as source data for the animation after the training phase to acquire lip sync and wider facial information from training material consisting of 2D videoswith audio had been completed. In March 2020, a freeware web application called 15.ai that generates high-quality voices from an assortment of fictional characters from a variety of media sources was released. Initial characters included GLaDOS from Portal, Twilight Sparkle and Fluttershy from the show My Little Pony: Friendship Is Magic, and the Tenth Doctor from Doctor Who.

Speech synthesis markup languages

A number of markup languages have been established for the rendition of text as speech in an XML-compliant format. The most recent is Speech Synthesis Markup Language (SSML),which became a W3C recommendation in 2004. Older speech synthesis markup languages include Java Speech Markup Language (JSML) and SABLE. Although each of these was proposed as a standard, none of them have been widely adopted. Speech synthesis markup languages are distinguished from dialogue markup languages. VoiceXML, for example, includes tags related to speech recognition, dialogue management and touchtone dialing, in addition to text-to-speech markup.

Applications

Speech synthesis has long been a vital assistive technology tool and its application in this area is significant and widespread. It allows environmental barriers to be removed for people with a wide range of disabilities. The longest application has been in the use of screen readers for people with visual impairment, but text-to-speech systems are now commonly used by people with dyslexia and other reading disabilities as well as by pre-literate children. They are also frequently employed to aid those with severe speech impairment usually through a dedicated voice output communication aid. Work to personalize a synthetic voice to better match aperson's personality or historical voice is becoming available. A noted application, of speech synthesis, was the Kurzweil Reading Machine for the Blind which incorporated text-to- phonetics software based on work from Haskins Laboratories and a black-box synthesizer built by Votrax. Speech synthesis techniques are also used in entertainment productions such as games and animations. In 2007, Animo Limited announced the development of a software application package based on its speech synthesis software FineSpeech, explicitly geared towards customers in the entertainment industries, able to generate narration and lines ofdialogue according to user specifications. The application reached maturity in 2008, when NEC Biglobe announced a web service that allows users to create phrases from the voices of characters from the Japanese anime series Code Geass: Lelouch of the Rebellion R2.

In recent years, text-to-speech for disability and impaired communication aids have become widely available. Text-to-speech is also finding new applications; for example, speech synthesiscombined with speech recognition allows for interaction with mobile devices via natural language processing interfaces. Text-to-speech is also used in second language acquisition. Voki, for instance, is an educational tool created by Oddcast that allows users to create theirown talking avatar, using different accents. They can be emailed, embedded on websites or shared on social media. Another area of application is AI video creation with talking heads. Tools, like Elai.io are allowing users to create video content with AI avatars who speak using text-to-speech technology.

In addition, speech synthesis is a valuable computational aid for the analysis and assessment of speech disorders. A voice quality synthesizer, developed by Jorge C. Lucero et al. at the University of Brasília, simulates the physics of phonation and includes models of vocal frequency jitter and tremor, airflow noise and laryngeal asymmetries. The synthesizer has been used to mimic the timbre of dysphonic speakers with controlled levels of roughness, breathinessand strain.

3.3.Google Trans

Google Translate is a multilingual neural machine translation service developed by Google to translate text, documents and websites from one language into another. It offers a website interface, a mobile app for Android and iOS, and an API that helps developers build browser extensions and software applications. As of April 2023, Google Translate supports 133 languages at various levels, and as of April 2016, claimed over 500 million total users, with more than 100 billion words translated daily, after the company stated in May 2013 that it served over 200 million people daily.

Launched in April 2006 as a statistical machine translation service, it used United Nations and European Parliament documents and transcripts to gather linguistic data. Rather than translating languages directly, it first translates text to English and then pivots to the target language in most of the language combinations it posits in its grid, with a few exceptions including Catalan- Spanish. During a translation, it looks for patterns in millions of documents to help decide which words to choose and how to arrange them in the target language. Its accuracy, which has been criticized on several occasions, has been measured to vary greatly across languages. In November 2016, Google announced that Google Translate would switch to a neural machine translation engine – Google Neural Machine Translation (GNMT) – which translates "whole sentences at a time, rather than just piece by piece. It uses this broader context to help it figure out the most relevant translation, which it then rearranges and adjusts to be more like a human speaking with proper grammar"

- Functions : Google Translate can translate multiple forms of text and media, which includes text, speech, and text within still or moving images. Specifically, its functions include:
- Written Words Translation: a function that translates written words or text to a foreign language.
- Website Translation: a function that translates a whole webpage to selected languages.
- Document Translation: a function that translates a document uploaded by the users to selected languages. The documents should be in the form of: .doc, .docx, .odf, .pdf, .ppt, .pptx, .ps, .rtf, .txt, .xls, .xlsx.
- Speech Translation: a function that instantly translates spoken language into the selected foreign language.

- Mobile App Translation: in 2018, Google introduced its new Google Translate feature called "Tap to Translate", which made instant translation accessible inside any app without exiting or switching it.
- Image Translation: a function that identifies text in a picture taken by the users and translates text on the screen instantly by images.
- Handwritten Translation: a function that translates language that are handwritten on the phone screen or drawn on a virtual keyboard without the support of a keyboard.
- Bilingual Conversation Translation: a function that translates conversations in multiple languages.
- Transcription: a function that transcribes speech in different languages.
- For most of its features, Google Translate provides the pronunciation, dictionary, and listening to translation. Additionally, Google Translate has introduced its own Translate app, so translation is available with a mobile phone in offline mode.

3.3.Grey Scale Images

If the original color image has no defined colorspace, or if the grayscale image is not intendedto have the same human-perceived achromatic intensity as the color image, then there is no unique mapping from such a color image to a grayscale image.

Converting color to grayscale:

Conversion of an arbitrary color image to grayscale is not unique in general; different weightingof the color channels effectively represent the effect of shooting black-and-white film with different-colored photographic filters on the cameras.

Colorimetric (perceptual luminance-preserving) conversion to grayscale:

A common strategy is to use the principles of photometry or, more broadly, colorimetry to calculate the grayscale values (in the target grayscale colorspace) so as to have the same luminance (technically relative luminance) as the original color image (according to its colorspace). In addition to the same (relative) luminance, this method also ensures that both images will have the same absolute luminance when displayed, as can be measured by instruments in its SI units of candelas per square meter, in any given area of the image, given equal white points. Luminance itself is defined using a standard model of human vision, so preserving the luminance in the grayscale image also preserves other perceptual lightnessmeasures, such as L* (as in the 1976 CIE Lab color space) which is determined by the linear luminance Y itself (as in the CIE 1931 XYZ color space) which we will refer to here as Ylinear to avoid any ambiguity.

To convert a color from a colorspace based on a typical gamma-compressed (nonlinear) RGB color model to a grayscale representation of its luminance, the gamma compression function must first be removed via gamma expansion (linearization) to transform the image to a linear RGB colorspace, so that the appropriate weighted sum can be applied to the linear colorcomponents {\displaystyle R_{\mathrm {linear} },G_{\mathrm {linear} },B_{\mathrm {linear}

}}) to calculate the linear luminance Ylinear, which can then be gamma-compressed back again if the grayscale result is also to be encoded and stored in a typical nonlinear colorspace.

Csrgb represents any of the three gamma-compressed sRGB primaries (Rsrgb, Gsrgb, and Bsrgb, each in range [0,1]) and Clinear is the corresponding linear-intensity value (Rlinear, Glinear, and Blinear, also in range [0,1]). Then, linear luminance is calculated as a weightedsum of the three linear-intensity values. The sRGB color space is defined in terms of the CIE 1931 linear luminance Ylinear{\displaystyle Y_{\mathrm {linear} }=0.2126R_{\mathrm

{linear} }+0.7152G_{\mathrm {linear} }+0.0722B_{\mathrm {linear} }}.

These three particular coefficients represent the intensity (luminance) perception of typical trichromat humans to light of the precise Rec. 709 additive primary colors (chromaticities) that are used in the definition of sRGB. Human vision is most sensitive to green, so this has the greatest coefficient value (0.7152), and least sensitive to blue, so this has the smallest coefficient (0.0722) {\displaystyle Y_{\mathrm {linear} },Y_{\mathrm {linear} },Y_{\mathrm

{linear} }} to get this linear grayscale), which then typically needs to be gamma compressed to get back to a conventional non-linear representation {\displaystyle Y_ {\mathrm{srgb}

}={\begin{cases}12.92\ Y_{\mathrm {linear} },&{\text{if }}Y_{\mathrm {linear} }\leq 0.0031308\\1.055\ Y_{\mathrm {linear} }^{1/2.4}-0.055,&{\text{otherwise}}\end{cases}}}

Because the three sRGB components are then equal, indicating that it is actually a gray image (not color), it is only necessary to store these values once, and we call this the resulting grayscale image. This is how it will normally be stored in sRGB-compatible image formats that support a single-channel grayscale representation, such as JPEG or PNG. Web browsers and other software that recognizes sRGB images should produce the same rendering for such a grayscale image as it would for a "color" sRGB image having the same values in all three color channels.

Luma coding in video systems

For images in color spaces such as Y'UV and its relatives, which are used in standard color TV and video systems such as PAL, SECAM, and NTSC, a nonlinear luma component (Y′) is calculated directly from gamma-compressed primary intensities as a weighted sum, which, although not a perfect representation of the colorimetric luminance, can be calculated more quickly without the gamma expansion and compression used in photometric/colorimetric calculations. In the Y'UV and Y'IQ models used by PAL and NTSC,

But if the luma component Y' itself is instead used directly as a grayscale representation of the color image, luminance is not preserved: two colors can have the same luma Y′ but different CIE linear luminance Y (and thus different nonlinear Ysrgb as defined above) and therefore appear darker or lighter to a typical human than the original color. Similarly, two colors having the same luminance Y (and thus the same Ysrgb) will in general have different luma by eitherof the Y′ luma definitions above.

3.3.Neural Network

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. Neuralnetworks can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria. The concept of neural networks, which has its roots in artificial intelligence, is swiftly gaining popularity in the development of trading systems.

Neural networks, in the world of finance, assist in the development of such processes as time- series forecasting, algorithmic trading, securities classification, credit risk modeling, and constructing proprietary indicators and price derivatives. A neural network works similarly to the human brain's neural network as shown in fig. 4. A "neuron" in a neural network is a mathematical function that collects and classifies information according to a specific architecture. The network bears a strong resemblance to statistical methods such as curve fitting and regression analysis. A neural network contains layers of interconnected nodes. Each node isa known as perceptron and is similar to a multiple linear regression. The perceptron feeds the signal produced by a multiple linear regression into an activation function that may be nonlinear.

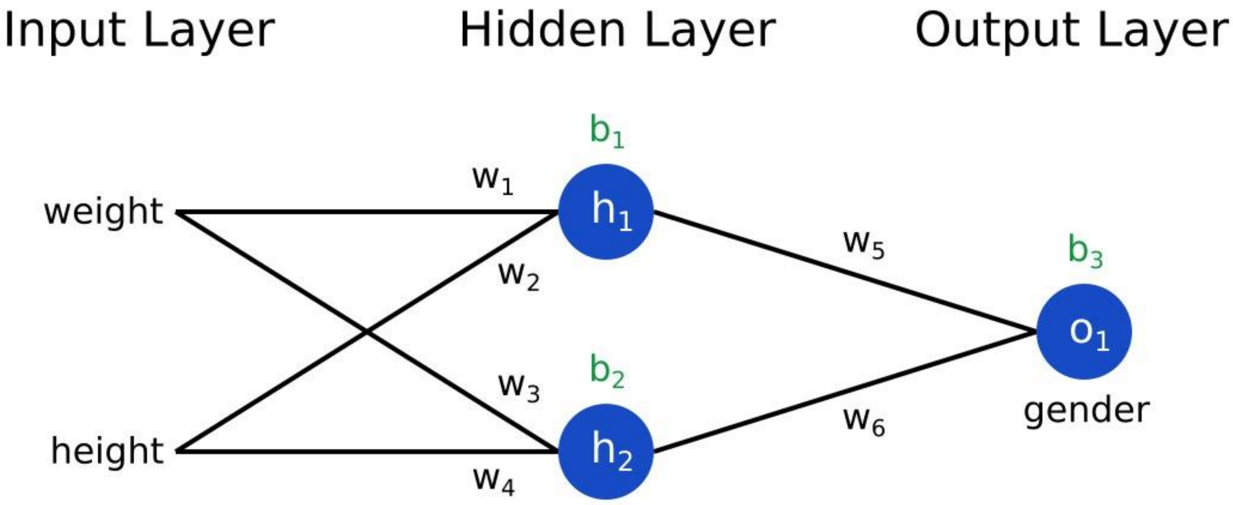

Fig. 4. A Simple Neural Network

Types of Neural Networks

- Feed-Forward Neural Networks

  Feed-forward neural networks are one of the more simple types of neural networks. It conveys information in one direction through input nodes; this information continues to be processed in this single direction until it reaches the output mode. Feed-forwardneural networks may have hidden layers for functionality, and this type of most often used for facial recognition technologies.

- Recurrent Neural Networks

  A more complex type of neural network, recurrent neural networks take the output of a processing node and transmit the information back into the network. This results in theoretical "learning" and improvement of the network. Each node stores historicalprocesses, and these historical processes are reused in the future during processing. This becomes especially critical for networks in which the prediction is incorrect; the system will attempt to learn why the correct outcome occurred and adjust accordingly. This typeof neural network is often used in text-to-speech applications.

- Convolutional Neural Networks

  Convolutional neural networks, also called ConvNets or CNNs, have several layers in which data is sorted into categories. These networks have an input layer, an output layer,and a hidden multitude of convolutional layers in between. The layers create feature maps that record areas of an image that are broken down further until they generate valuable outputs. These layers can be pooled or entirely connected, and these networks are especially beneficial for image recognition applications.

- Deconvolutional Neural Networks

  Deconvolutional neural networks simply work in reverse of convolutional neural networks. The application of the network is to detect items that might have been recognized as important under a convolutional neural network. These items would likely have been discarded during the convolutional neural network execution process. This type of neural network is also widely used for image analysis or processing.

- Modular Neural Networks

  Modular neural networks contain several networks that work independently from one another. These networks do not interact with each other during an analysis process. Instead, these processes are done to allow complex, elaborate computing processes to be done more efficiently. Similar to other modular industries such as modular real estate,the goal of the network independence is to have each module responsible for a particularpart of an overall bigger picture.

Application of Neural Networks

Neural networks are broadly used, with applications for financial operations, enterprise planning, trading, business analytics, and product maintenance. Neural networks have also gained widespread adoption in business applications such as forecasting and marketing researchsolutions, fraud detection, and risk assessment.

A neural network evaluates price data and unearths opportunities for making trade decisions based on the data analysis. The networks can distinguish subtle nonlinear interdependencies andpatterns other methods of technical analysis cannot. According to research, the accuracy of neural networks in making price predictions for stocks differs. Some models predict the correct stock prices 50 to 60% of the time, while others are accurate in 70% of all instances. Some haveposited that a 10% improvement in efficiency is all an investor can ask for from a neural network.

Specific to finance, neural networks can process hundreds of thousands of bits of transaction data. This can translate to a better understanding of trading volume, trading range, correlation between assets, or setting volatility expectations for certain investments. As a human may not be able to efficiently pour through years of data (sometimes collected down second intervals), neural networks can be designed to spot trends, analyze outcomes, and predict future asset class value movements.

3.3.Digital Image Processing

Digital image processing is the use of a digital computer to process digital images through an algorithm. As a subcategory or field of digital signal processing, digital image processing has many advantages over analog image processing. It allows a much wider range of algorithms to be applied to the input data and can avoid problems such as the build-up of noise and distortion during processing. Since images are defined over two dimensions (perhaps more) digital image processing may be modeled in the form of multidimensional systems. The generation and development of digital image processing are mainly affected by three factors: first, the development of computers; second, the development of mathematics (especially the creationand improvement of discrete mathematics theory); third, the demand for a wide range of applications in environment, agriculture, military, industry and medical science has increased.

Image sensors

The basis for modern image sensors is metal–oxide–semiconductor (MOS) technology, which originates from the invention of the MOSFET (MOS field-effect transistor) by Mohamed M. Atalla and DawonKahng at Bell Labs in 1959. This led to the development of digital semiconductor image sensors, including the charge-coupled device (CCD) and later the CMOS sensor. The charge-coupled device was invented by Willard S. Boyle and George E. Smith at Bell Labs in 1969.While researching MOS technology, they realized that an electric charge was the analogy of the magnetic bubble and that it could be stored on a tiny MOS capacitor. As it was fairly straightforward to fabricate a series of MOS capacitors in a row, they connected a suitable voltage to them so that the charge could be stepped along from one to the next. The CCD is a semiconductor circuit that was later used in the first digital video cameras fortelevision broadcasting. The NMOS active-pixel sensor (APS) was invented by Olympus in Japan during the mid-1980s. This was enabled by advances in MOS semiconductor device fabrication, with MOSFET scaling reaching smaller micron and then sub-micron levels. The NMOS APS was fabricated by Tsutomu Nakamura's team at Olympus in 1985. The CMOS active-pixel sensor (CMOS sensor) was later developed by Eric Fossum's team at the NASA Jet Propulsion Laboratory in 1993. By 2007, sales of CMOS sensors had surpassed CCD sensors.

Image compression

An important development in digital image compression technology was the discrete cosine transform (DCT), a lossy compression technique first proposed by Nasir Ahmed in 1972.DCT compression became the basis for JPEG, which was introduced by the Joint Photographic Experts Group in 1992. JPEG compresses images down to much smaller file sizes, and has become the most widely used image file format on the Internet. Its highly efficient DCT compression algorithm was largely responsible for the wide proliferation of digital images and digital photos, with several billion JPEG images produced every day as of 2015.

Digital signal processor (DSP)

Electronic signal processing was revolutionized by the wide adoption of MOS technology in the1970s. MOS integrated circuit technology was the basis for the first single-chip microprocessors and microcontrollers in the early 1970s, and then the first single-chip digital signal processor (DSP) chips in the late 1970s. DSP chips have since been widely used in digital image processing. The discrete cosine transform (DCT) image compression algorithm has been widely implemented in DSP chips, with many companies developing DSP chips based on DCT technology. DCTs are widely used for encoding, decoding, video coding, audio coding, multiplexing, control signals, signaling, analog-to-digital conversion, formatting luminance and color differences, and color formats such as YUV444 and YUV411. DCTs are also used for encoding operations such as motion estimation, motion compensation, inter-frame prediction, quantization, perceptual weighting, entropy encoding, variable encoding, and motion vectors, and decoding operations such as the inverse operation between different color formats (YIQ, YUV and RGB) for display purposes. DCTs are also commonly used for high-definition television (HDTV) encoder/decoder chips.

Medical imaging

In 1972, the engineer from British company EMI Housfield invented the X-ray computed tomography device for head diagnosis, which is what is usually called CT (computer tomography). The CT nucleus method is based on the projection of the human head section and is processed by computer to reconstruct the cross-sectional image, which is called imagereconstruction. In 1975, EMI successfully developed a CT device for the whole body, which obtained a clear tomographic image of various parts of the human body. In 1979, this diagnostic technique won the Nobel Prize. Digital image processing technology for medical applications was inducted into the Space Foundation Space Technology Hall of Fame in 1994.

Digital image processing allows the use of much more complex algorithms, and hence, can offerboth more sophisticated performance at simple tasks, and the implementation of methods which would be impossible by analogue means.

In particular, digital image processing is a concrete application of, and a practical technology based on:

- Classification
- Feature extraction
- Multi-scale signal analysis
- Pattern recognition
- Projection
- Some techniques which are used in digital image processing include:
- Anisotropic diffusion
- Hidden Markov models
- Image editing
- Image restoration
- Independent component analysis
- Linear filtering
- Neural networks
- Partial differential equations
- Pixelation
- Point feature matching
- Principal components analysis
- Self-organizing maps
- Wavelets
- Digital image transformations
- Filtering

Digital filters are used to blur and sharpen digital images. Filtering can be performed by: convolution with specifically designed kernels (filter array) in the spatial domain masking specific frequency regions in the frequency (Fourier) domain.

3.3.OCR

Python OCR is a technology that recognizes and pulls out text in images like scanned documents and photos using Python. It can be completed using the open-source OCR engine Tesseract. We can do this in Python using a few lines of code. One of the most common OCR tools that are used is the Tesseract. Tesseract is an optical character recognition engine for various operating systems.

A process called Optical Character Recognition (OCR) converts printed texts into digital image files. It is a digital copier that uses automation to convert scanned documents into editable, shareable PDFs that are machine-readable. OCR may be seen in action when you use your computer to scan a receipt. The scan is then saved as a picture on your computer. The words in the image cannot be searched, edited, or counted, but you may use OCR to convert the image to a text document with the content stored as text. OCR software can extract data from scanned documents, camera photos, and image-only PDFs. It makes static material editable and does away with the necessity for human data entry.

Widely used as a form of data entry from printed paper data records – whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation – it is a common method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as cognitive computing, machine translation, (extracted) text- to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. Early versions needed to be trained with images of each character, and worked on one font at a time. Advanced systems capable of producing a high degree of recognition accuracy for most fonts are now common, and with support for a varietyof digital image file format inputs.[2] Some systems are capable of reproducing formatted output that closely approximates the original page including images, columns, and other non- textual components. OCR engines have been developed into many kinds of domain-specific OCR applications, such as receipt OCR, invoice OCR, check OCR, legal billing document OCR.

They can be used for:

- Data entry for business documents, e.g. Cheque, passport, invoice, bank statement and receipt
- Automatic number plate recognition
- In airports, for passport recognition and information extraction Automatic insurance documents key information extraction [citation needed] Traffic-sign recognition, Extracting business card information into a contact list
- More quickly make textual versions of printed documents, e.g. book scanning for Project Gutenberg Make electronic images of printed documents searchable, e.g. Google Books Converting handwriting in real-time to control a computer (pen computing)
- Defeating CAPTCHA anti-bot systems, though these are specifically designed to prevent OCR.The purpose can also be to test the robustness of CAPTCHA anti-bot systems.
- Assistive technology for blind and visually impaired users Writing the instructions for vehicles by identifying CAD images in a database that are appropriate to the vehicle design as it changes in real time. Making scanned documents searchable by converting them to searchable PDFs

Types:

- Optical character recognition (OCR) – targets typewritten text, one glyph or character ata time.
- Optical word recognition – targets typewritten text, one word at a time (for languagesthat use a space as a word divider).(Usually just called "OCR".)
- Intelligent character recognition (ICR) – also targets handwritten print script or cursivetext one glyph or character at a time, usually involving machine learning.

- Intelligent word recognition (IWR) – also targets handwritten print script or cursive text, one word at a time. This is especially useful for languages where glyphs are not separated in cursive script.
- OCR is generally an "offline" process, which analyses a static document. There are cloud based services which provide an online OCR API service. Handwriting movementanalysis can be used as input to handwriting recognition. Instead of merely using the shapes of glyphs and words, this technique is able to capture motions, such as the order in which segments are drawn, the direction, and the pattern of putting the pen down and lifting it. This additional information can make the end-to-end process more accurate. This technology is also known as "on-line character recognition", "dynamic character recognition", "real-time character recognition", and "intelligent character recognition".

Techniques:

- Pre-processing - OCR software often "pre-processes" images to improve the chances of successful recognition. Techniques include:
- De-skew – If the document was not aligned properly when scanned, it may need to be tilted a few degrees clockwise or counterclockwise in order to make lines of textperfectly horizontal or vertical.
- Despeckle – remove positive and negative spots, smoothing edges
- Binarisation – Convert an image from color or greyscale to black-and-white (called a "binary image" because there are two colors). The task of binarisation is performed as a simple way of separating the text (or any other desired image component) from the background. The task of binarisation itself is necessary since most commercial recognition algorithms work only on binary images since it proves to be simpler to do so. In addition, the effectiveness of the binarisation step influences to a significant extentthe quality of the character recognition stage and the careful decisions are made in the choice of the binarisation employed for a given input image type; since the quality of thebinarisation method employed to obtain the binary result depends on the type of the input image (scanned document, scene text image, historical degraded document etc.).
- Line removal – Cleans up non-glyph boxes and lines

- Layout analysis or "zoning" – Identifies columns, paragraphs, captions, etc. as distinct blocks. Especially important in multi-column layouts and tables.
- Line and word detection – Establishes baseline for word and character shapes, separates words if necessary.
- Script recognition – In multilingual documents, the script may change at the level of the words and hence, identification of the script is necessary, before the right OCR can be invoked to handle the specific script.
- Character isolation or "segmentation" – For per-character OCR, multiple characters that are connected due to image artifacts must be separated; single characters that are broken into multiple pieces due to artifacts must be connected.
- Normalize aspect ratio and scale
- Segmentation of fixed-pitch fonts is accomplished relatively simply by aligning the image to a uniform grid based on where vertical grid lines will least often intersect blackareas. For proportional fonts, more sophisticated techniques are needed because whitespace between letters can sometimes be greater than that between words, and vertical lines can intersect more than one character.

There are two basic types of core OCR algorithm, which may produce a ranked list of candidate characters.

Matrix matching involves comparing an image to a stored glyph on a pixel-by-pixel basis; it is also known as "pattern matching", "pattern recognition", or "image correlation". This relies on the input glyph being correctly isolated from the rest of the image, and on the stored glyph being in a similar font and at the same scale. This technique works best with typewritten textand does not work well when new fonts are encountered. This is the technique the early physical photocell-based OCR implemented, rather directly.

Feature extraction decomposes glyphs into "features" like lines, closed loops, line direction, andline intersections. The extraction features reduces the dimensionality of the representation and makes the recognition process computationally efficient. These features are compared with an abstract vector-like representation of a character, which might reduce to one or more glyph prototypes. General techniques of feature detection in computer vision are applicable to this type of OCR, which is commonly seen in "intelligent" handwriting recognition and indeed most modern OCR software. Nearest neighbour classifiers such as the k-nearest neighbors algorithm are used to compare image features with stored glyph features and choose the nearest match.

Software such as Cuneiform and Tesseract use a two-pass approach to character recognition. The second pass is known as "adaptive recognition" and uses the letter shapes recognized with high confidence on the first pass to recognize better the remaining letters on the second pass. This is advantageous for unusual fonts or low-quality scans where the font is distorted (e.g. blurred or faded).

Modern OCR software include Google Docs OCR, ABBYY FineReader and Transym. Others like OCRopus and Tesseract uses neural networks which are trained to recognize whole lines of text instead of focusing on single characters.

A new technique known as iterative OCR automatically crops a document into sections based on page layout. OCR is performed on the sections individually using variable character confidence level thresholds to maximize page-level OCR accuracy. A patent from the United States Patent Office has been issued for this method

The OCR result can be stored in the standardized ALTO format, a dedicated XML schema maintained by the United States Library of Congress. Other common formats include hOCR andPAGE XML.

For a list of optical character recognition software see Comparison of optical characterrecognition software.

Post-processing

OCR accuracy can be increased if the output is constrained by a lexicon – a list of words that are allowed to occur in a document. This might be, for example, all the words in the English language, or a more technical lexicon for a specific field. This technique can be problematic if the document contains words not in the lexicon, like proper nouns. Tesseract uses its dictionary to influence the character segmentation step, for improved accuracy.

The output stream may be a plain text stream or file of characters, but more sophisticated OCR systems can preserve the original layout of the page and produce, for example, an annotated PDF that includes both the original image of the page and a searchable textual representation.

"Near-neighbor analysis" can make use of co-occurrence frequencies to correct errors, by notingthat certain words are often seen together. For example, "Washington, D.C." is generally far more common in English than "Washington DOC".

Knowledge of the grammar of the language being scanned can also help determine if a word is likely to be a verb or a noun, for example, allowing greater accuracy.

The Levenshtein Distance algorithm has also been used in OCR post-processing to further optimize results from an OCR API.

Application-specific optimizations

In recent years,[when?] the major OCR technology providers began to tweak OCR systems to deal more efficiently with specific types of input. Beyond an application-specific lexicon, better performance may be had by taking into account business rules, standard expression,[clarification needed] or rich information contained in color images. This strategy is called "Application-Oriented OCR" or "Customized OCR", and has been applied to OCR of license plates, invoices, screenshots, ID cards, driver licenses, and automobile manufacturing.

The New York Times has adapted the OCR technology into a proprietary tool they entitle, Document Helper, that enables their interactive news team to accelerate the processing of documents that need to be reviewed. They note that it enables them to process what amounts to as many as 5,400 pages per hour in preparation for reporters to review the contents.

Tesseract was in the top three OCR engines in terms of character accuracy in 1995. It is available for Linux, Windows and Mac OS X. However, due to limited resources it is only rigorously tested by developers under Windows and Ubuntu.

Tesseract up to and including version 2 could only accept TIFF images of simple one-column text as inputs. These early versions did not include layout analysis, and so inputting multi- columned text, images, or equations produced garbled output. Since version 3.00 Tesseract has supported output text formatting, OCR positional information and page-layout analysis. Supportfor a number of new image formats was added using the Leptonica library. Tesseract can detect whether text is monospaced or proportionally spaced.

The initial versions of Tesseract could only recognize English-language text. Tesseract v2 added six additional Western languages (French, Italian, German, Spanish, Brazilian Portuguese, Dutch). Version 3 extended language support significantly to include ideographic (Chinese & Japanese) and right-to-left (e.g. Arabic, Hebrew) languages, as well as many more scripts. New languages included Arabic, Bulgarian, Catalan, Chinese (Simplified andTraditional), Croatian, Czech, Danish, German (Fraktur script), Greek, Finnish, Hebrew, Hindi, Hungarian, Indonesian, Japanese, Korean, Latvian, Lithuanian, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak (standard and Fraktur script), Slovenian, Swedish, Tagalog, Tamil, Thai, Turkish, Ukrainian and Vietnamese. V3.04, released in July 2015, added an additional 39 language/script combinations, bringing the total count of support languages to over 100. New language codes included: amh (Amharic), asm (Assamese), aze_cyrl(Azerbaijana in Cyrillic script), bod (Tibetan), bos (Bosnian), ceb (Cebuano), cym (Welsh), dzo (Dzongkha), fas (Persian), gle (Irish), guj (Gujarati), hat (Haitian and Haitian Creole), iku (Inuktitut), jav (Javanese), kat (Georgian), kat_old (Old Georgian), kaz (Kazakh), khm (Central Khmer), kir (Kyrgyz), kur (Kurdish), lao (Lao), lat (Latin), mar (Marathi), mya (Burmese), nep (Nepali), ori (Oriya), pan (Punjabi), pus (Pashto), san (Sanskrit), sin (Sinhala), srp_latn (Serbian in Latin script), syr (Syriac), tgk (Tajik), tir (Tigrinya), uig (Uyghur), urd (Urdu), uzb (Uzbek), uzb_cyrl (Uzbek in Cyrillic script), yid (Yiddish).

CHAPTER 4:

RESULT AND DISCUSSIONS

4.1 .Introduction

This paper is to review various approaches used for providing the assistive reading framework for the visually challenged persons. Visually challenged persons are the persons who are either blind or having any kind of difficulty in reading any printed material to acquire the domain knowledge. A lot of research is being done for visually challenged to make them independent intheir life. We want to study the existing assistive technology for visually challenged and then propose a robust reading framework for visually challenged. The discussed reading framework will help the visually challenged person to read normal printed books, typed documents, journals, magazines, newspapers and computer displays of emails, Web pages, etc., like normal persons. It is a system based on image processing and pattern recognition using which visually challenged person carries or wears a portable camera as a digitizing device and uses computeras a processing device. The camera captures the image of the text to be read along with the relevant image and data. An optical character recognition (OCR) system segregates the image into text and non-text boxes, and then, the OCR converts the text from the text boxes to ASCII or text file. The text file is converted to voice by a text-to-speech (TTS) converter. Thus, blind person would 'hear' the text information that has been captured. So this technology can help thevisually challenged to read the captured textual information independently which will be communicated as voice signal output of a text-to-speech converter.

4.1.Methodology

TTS software in general is considered an assistive technology tool that can be used in many ways. Another early application of this technology was to help people who have troublereading. The amendment of the Individuals with Disabilities Education Act (IDEA) in 2004 compelled educational institutions to seek out technology to assist in fulfilling this mandate.The IDEA is a federal law ensuring educational services to children with disabilities throughout the United States. 4 TTS allows users to see text and hear it read aloud simultaneously. There are many apps available, but typically as text appears on the screen, it’s spoken. Some software uses a computer-generated voice and others use a recorded human voice. Very often the user has a choice of gender and accent as well. People with learning disabilities who have difficulty reading large amounts of text due to dyslexia or other problems really benefit from TTS, offering them an easier option for experiencing website content. People who have literacy issuesand those trying to learn another language often get frustrated trying to browse the internet because so much text is confusing. Many people have difficulty reading fluently in a second language even though they may be able to read content with a basic understanding. TTS technology allows them to understand information in a way that makes content easier to retain.

4.1.Result and Discussion

Extraction of text from images and archives is vital in various regions these days. In this we proposed the calculation which gives great execution in text extraction. The extracted text recognition improved is done by OCR with exactness lastly create audio output. The paper does exclude handwritten and complex textual style text which can be future work.

CHAPTER 5:

CONCLUSIONS AND FUTURE SCOPE

5.1.Conclusion

Text-to-Speech device can change the text image input into sound with a performance that is high enough and a readability tolerance of less than 2%, with the average time processing less than three minutes for A4 paper size. This portable device, does not require internet connection, and can be used independently by people. Through this method, we can make editing process ofbooks or web pages easier.

5.1.Future Scope

It would be a thing of the past to type out your message or paper, because we could just use our voice. It does make sense in some regards, because we can probably speak much faster than typing in most cases. But there are certain drawbacks that could hinder the expansion of this idea.

The only thing that we do right now is wait and see how the text to speech world is going to change. Not only are the converters getting better with their methods, but the entire industry has the ability to make a lasting impact very soon. Keep your eyes peeled to see which industries it will affect and how many lives it will impact.

REFERENCES

[1] P. Shivakumara, S. Bhowmick, B. Su, C. L. Tan, and U. Pal, ‘‘A new gradient based character segmentation method for video text recognition,’’ in Proc. Int. Conf. Document Anal. Recognit., Sep. 2011, pp. 126–130.

[2] M.M. Luqman, pp. Gomez–Krämer, and JM. Ogier, ‘‘Mobile phonecamera-based video scanning of paper documents,’’ in Camera-Based Document Analysis and Recognition (LectureNotes in Computer Science), vol. 8357. Cham, Switzerland: Springer, 2013.

[3] T. Saba and A. Rehman, ‘‘Effects of artificially intelligent tools on pattern recognition,’’ Int. J. Mach. Learn. Cyber., vol. 4, no. 2, pp. 155–162, Apr. 2013.

[4] K. Bulatov, V. V. Arlazarov, T. Chernov, O. Slavin, and D. Nikolaev,‘‘SmartIDReader: Document recognition in video stream,’’ in Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR), Nov. 2017, pp. 39–44

[5] F. Jia, C. Shi, Y. Wang, C. Wang, and B. Xiao, ‘‘Grayscale-projectionbased optimal character segmentation for camera-captured faint text recognition,’’ in Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR), Nov. 2017, pp. 1301–1306

[6] K. Bulatov, ‘‘Selecting optimal strategy for combining per-frame characterrecognition results in video stream,’’ J. Inf. Technol. Comput. Syst., no. 3, pp. 45–55, 2017

[7] R. Hussain, H. Gao, and R. A. Shaikh, ‘‘Segmentation of connected characters in text- based CAPTCHAs for intelligent character recognition,’’Multimedia Tools Appl., vol. 76, no. 24, pp. 25547–25561, Dec. 2017.

[8] G. Renton, C. Chatelain, S. Adam, C. Kermorvant, and T. Paquet, ‘‘Handwritten text line segmentation using fully convolutional network,’’ in Proc.14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR), vol. 1. Kyoto,Japan, Nov. 2017, pp. 5–9.

[9] P. Sahare and S. B. Dhok, ‘‘Multilingual character segmentation andrecognition schemes for Indian document images,’’ IEEE Access, vol. 6, pp. 10603–10617, 2018.

[10] P. Sahare and S. B. Dhok, ‘‘Multilingual character segmentation andrecognition schemes for Indian document images,’’ IEEE Access, vol. 6, pp. 10603–10617, 2018.

[11] G. Nanfack, A. Elhassouny, and R. O. H. Thami, ‘‘Squeeze-SegNet: A newfast deep convolutional neural network for semantic- segmentation,’’ Proc. SPIE, vol. 10696, Apr. 2018, Art. no. 106962O

[12] Rampurkar, V. V., Shah, S. K., Chhajed, G. J., &Biswash, S. K. (2018,January). An approach towards text detection from complex images using morphological techniques. In 2018 2nd International Conference on Inventive Systems and Control (ICISC) (pp. 969-973). IEEE.

[13] Ling, O. Y., Theng, L. B., Chai, A., & McCarthy, C. (2018,November). A Model for Automatic Recognition of Vertical Texts in Natural Scene Images. In 2018 8th IEEE International Conference on Control System, Computing and Engineering (ICCSCE) (pp. 170- 175). IEEE.

[14] Ling, O. Y., Theng, L. B., Chai, A., & McCarthy, C. (2018,November). A Model for Automatic Recognition of Vertical Texts in Natural Scene Images. In 2018 8th IEEE International Conference on Control System, Computing and Engineering (ICCSCE) (pp. 170- 175). IEEE.

[15] Gao, X., Han, S., & Luo, C. (2019). A Detection and VerificationModel Based on SSD and Encoder-Decoder Network for Scene Text Detection. IEEE Access, 7, 71299-71310.

[16] Liu, F., Chen, C., Gu, D., & Zheng, J. (2019). FTPN: scene textdetection with feature pyramid based text proposal network. IEEE Access, 7, 44219-44228.

[17] Su, Y. M., Peng, H. W., Huang, K. W., & Yang, C. S. (2019,November). Image processing technology for text recognition. In 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI) (pp. 1-5). IEEE.

[18] Chernyshova, Y. S., Sheshkus, A. V., &Arlazarov, V. V. (2020). Twostepcnn framework for text line recognition in camera-capturedimages. IEEE Access, 8, 32587-32600

[19] Liang, Q., Xiang, S., Wang, Y., Sun, W., & Zhang, D. (2020). RNTRNet: A robust natural text recognition network. IEEE Access, 8, 7719-7730.

[20] KiranRakshana R, chittaC(2019) “A Smart Navguide System for visually Impaired”, IJITE, ISSN: 2278-3075, Issue 6S3, Vol. *, No.),pp. 0.

[21] Vaibhav V. Govekar, Meenakshi A(2018) “A Smart Reader for Blind People”, IJSTE, Issn:n2349-8958

[22] ShraddhaHingankar, PrachiTardekar, SantoshiPote(2020) “A Smart Reader for Visually Impaired Individuals”, International Research Journal of Engineering and Technology, P-ISSN: 2395-0072, E-ISSN: 2395-0056, Issue 07, Vol. 07, No. 0, pp. 0

[23] Aravind S, Roshna E(2013) “A Text Reding System for the Visually Disabled”, International Journal of Research in Computer Application & Management, ISSN: 2231-1009, Issue 12, Vol. 3, No. 0, pp. 0.

[24] AkhileshPanchal, ShrugalVarde, M.S Panse(2016) “Automatic Scene Text Detection and Recognition system for visually Impaired People”, International Journal for Research in Emerging Science and Technology, E-ISSN: 2349-761, Issue 6, Vol. 3, No. 0, pp. 0.

[25] N. S. Lokhande, P.B. Pawar, S.J. Shelke, A.R. Wagh(2019) “Book Reader for Visually Impaired Using Raspberry Pi”, International Research Journal of Engineering and Technology, P-ISSN: 2395-0072, E-ISSN: 2395-0056, Issue 02, Vol. 06, No. 0, pp 0.

R. ShanthaSlevaKumari, R. Sangeetha “Conversion of English Text To Speech (TTS) Using Indian Speech Signal”, Mathematical and Computational Methods in Electrical Engineering, Vol. 0, No.