MACHINE LEARNING SPEECH PRODUCTION SYSTEM BASED ON FACIAL RECOGNITION

Nuthanakanti Bhaskar
Associate Professor
Department of CSE
CMR TECHNICAL CAMPUS, HYDERABAD

D. Sri Durga Sai Alekhya
B. Tech student
Department of IT
CMR TECHNICAL CAMPUS, HYDERABAD

P. Sanjana
B. Tech student
Department of IT
CMR TECHNICAL CAMPUS, HYDERABAD

J. Vikas
B. Tech student
Department of IT
CMR TECHNICAL CAMPUS, HYDERABAD

ABSTRACT: Face detection is a computer vision technology that helps to locate/visualize human faces in digital images. This technique is a specific use case of object detection technology that deals with detecting instances of semantic objects of a certain class (such as humans, buildings or cars) in digital images and videos. With the advent of technology, face detection has gained a lot of importance especially in fields like photography, security, and marketing, hands-on knowledge. When coming to speech production, it is a useful technology that converts any text into a speech signal and it has a lot of importance in car navigation ,email reading etc. Speech synthesis can be described as artificial production of human speech. In this project we provide voice output or speech production additionally when the face is recognized. After recognizing the faces the text gets converted into speech so that we get voice as output. Generally speech is one of the oldest and most natural means of information exchange between human. Now, the aim of this project is to exchange the information between human and machine and has to detect and recognize faces in a scene and to provide the user with the person's name as well as additional speech in the form of voice. We use machine learning algorithm to validate unspecified images by comparing them to previously saved pictures in the training set and to provide information about when the individual recognized by the system.

I. INTRODUCTION

Face recognition is the problem of identifying and verifying people in a photograph by their face or it can identify in real time. It is a task that is trivially performed by humans, even under varying light and when faces are changed by age or obstructed with accessories and facial hair. Nevertheless, it is remained a challenging computer vision problem for decades until recently, Machine learning methods are able to leverage very large datasets of faces and learn rich and compact representations of faces, allowing modern models to first perform as-well and later to outperform the face recognition capabilities of humans[5]. Face recognition is a broad problem of identifying or verifying people in photographs and videos. Face recognition is a process comprised of detection, alignment, feature extraction, and a recognition task Machine learning models first approached then exceeded human performance for face recognition tasks[1]. Now, after face recognition we need to know about the speech production there are modules which converts text to speech when the face is detected. Text-to-speech (TTS) technology reads aloud digital text. It can take words on computers, smartphones, tablets and convert them into audio[6]. Also, all kinds of text files can be read aloud, including Word, pages document, online web pages can be read aloud. TTS can help kids who struggle with reading. Many tools and apps are available to convert text into speech. Different API's are available in Python in order to convert text to speech. One of Such API's is the Google Text to Speech commonly known as the gTTS API[7]. It is very easy to use the library which converts the text entered, into an audio file which can be saved as a mp3 file. It supports several languages and the speech can be delivered in any one of the two available audio speeds, fast or slow. TTS is an easy tool to convert text to voice, but it requires an internet connection to operate, because it depends entirely on Google to get the audio data[8]. A detailed study of the process must be made by various techniques like Image processing, feature recognition etc. The data collected by these sources

must be scrutinized to arrive to a conclusion[4]. The proposals. The proposal is then weighed with the existing system analytically and the best one is selected. The proposal is presented to the user for an endorsement by the user. The proposal is reviewed on user request and suitable changes are made. This is loop that ends as soon as the user is satisfied with proposal.

The Scope of the project is that it will uses machine-learning methods to identify or recognize the individual from detected faces and produces speech as output. We used machine learning algorithm to validate unspecified images by comparing them to previously saved pictures in the dataset and to provide information or speech about the individual recognized by the system[2]. The main features of it has the new way of designing the speech production that is to physically impaired and the vocally disturbed individuals using English language. It can enable the blind and elderly people enjoy a User-Friendly computer interface[1]. There are few features which are important to identify the face that is role of eyebrows and if it is not detected we may not get the speech. Here, we are having a feature in speech production that converts the text to speech synthesizer can be described as artificial production of human speech[8].

II. LITERATURE SURVEY

Face processing, amongst many basic visual skills, is thought to be invariant across all humans. From as early as 1965, studies of eye movements have consistently revealed a systematic triangular sequence of fixations over the eyes and the mouth, suggesting that faces elicit a universal, biologically-determined information extraction pattern[1]. It is a widely held belief that many basic visual processes are common to all humans, independent of culture. Here we monitored the eye movements of Western Caucasian and East Asian observers while they learned, recognized, and categorized by race Western Caucasian and East Asian faces. Western Caucasian observers reproduced a scattered triangular pattern of fixations for faces of both races and across tasks[2]. Contrary to intuition, East Asian observers focused more on the central region of the face. Face recognition is considered to be one such process, as this basic biological skill is necessary for effective social interactions[3]. Any approach aiming to understand face perception must recognize, however, that only a small part of the visual information available on faces is actually used. These results demonstrate that face processing can no longer be considered as arising from a universal series of perceptual events. The strategy employed to extract visual information from faces differs across cultures[4].

This paper presents an extension and refinement of the author's theory for human visual information processing, which is then applied to the problem of human facial recognition[5]. Several fundamental processes are implicated: encoding of visual images into neural patterns, detection of simple facial features, size standardization, reduction of the neural patterns in dimensionality, and finally correlation of the resulting sequence of patterns with all visual patterns already stored in memory. In the theory presented here, this entire process is automatically "driven" by the storage system in what amounts to an hypothesis verification paradigm[1].

Elderly persons exceed young adults in false recognitions of new faces. One account claims there are age-related deficits in memory for context of encounter with faces. Because of these deficits, elderly persons frequently recognize faces on the basis of perceived familiarity (i.e., resemblance to face representations in memory), which is high for some new faces[5]. To test this context-recollection hypothesis, we had young adult and elderly subjects judge whether faces: (1) had been seen previously in a test (though no face was repeated), and (2) were subjectively familiar (though no face was famous). So different authors was visualized the theory of face recognition based on their research. So, now the paper explained half of the project. Now, lets continue to know the text to speech production technology and how it developed and so on.

Purpose Accessing auditory and written material simultaneously benefits people with aphasia; however, the extent of benefit as well as people's preferences and experiences may vary given different auditory presentation rates. This study's purpose was to determine how 3 text-to-speech rates affect comprehension when adults with aphasia access newspaper articles through combined modalities. Secondary aims included exploring time spent reviewing written texts after speech output cessation, rate preference, preference consistency, and participant rationales for preferences[7]. Method Twenty-five adults with aphasia read and listened to passages presented at slow (113 words per minute [wpm]), medium (154 wpm), and fast (200 wpm) rates. So that’s how the author explained the problem occurs when people facing problem during newspaper.

As a cross-disciplinary, speech recognition is based on the voice as the research object. Speech recognition allows the machine to turn the speech signal into text or commands through the process of

identification and understanding, and also makes the function of natural voice communication. Speech recognition involves many fields of physiology, psychology, linguistics, computer science and signal processing, and is even related to the person's body language, and its ultimate goal is to achieve natural language communication between man and machine[6]. The speech recognition technology is gradually becoming the key technology of the IT man-machine interface [1] .The paper describes the development of speech recognition technology and its basic principles, methods, reviewed the classification of speech recognition systems and voice recognition technology, analyzed the problems faced by the speech recognition[8].

A Text-to-speech synthesizer is developed that converts text into spoken word, by analysing and processing it using Natural Language Processing (NLP) and then using Digital Signal Processing (DSP) technology to convert this processed text into synthesized speech representation of the text. A useful text-to-speech synthesizer in the form of a simple application that converts inputted text into synthesized speech and reads out to the user which can then be saved as an mp3 file[6].

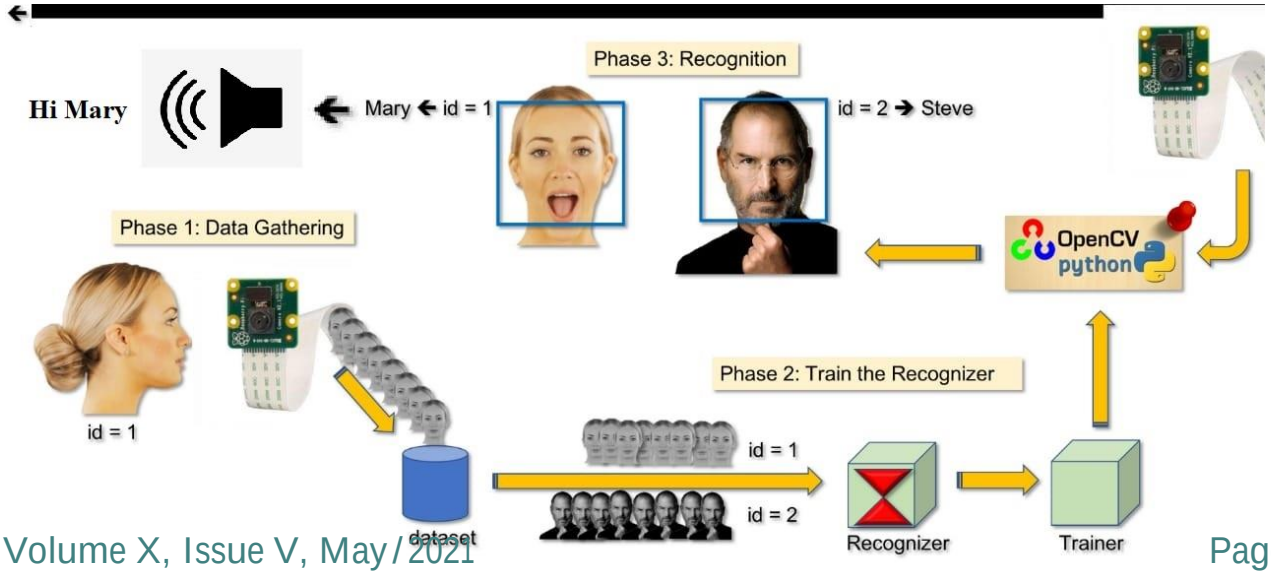
III. EXISTING SYSTEM

We already have face recognition techniques like smart security cameras which is used in digital medical applications and also used to detect a rare disease called DiGeorge Syndrome. We have five different algorithms that have been preferred based on the most widely used criteria. The algorithms are Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), **skin** colour, wavelet and Artificial Neural Network (ANN)[1].We also have voice output technologies like Digital Assistant- Amazon Echo, Apples Siri and Google Assistant use Voice Recognition to Interact with digital Assistance[6].

IV. PROPOSED SYSTEM ARCHITECTURE

In the proposed system, we mainly focus on face recognition technique. When the person comes infront of the camera then it captures the face of the person and the system will detect the individual face and if it is in the dataset then it wishes the individual. To implement this, first the individual should click on take images then the system will take around 80 images and then save the individual profile. Then, it asks you to enter the password and you can check the recognition then if your face is registered in dataset then it produces the speech as output.

Below figure shows the architecture of our project, which explain about the functionality of each step. Here, if system wants to detect the face, first we need to train the face to the machine with some identity with different shades.so camera captures the face of the person it detects and it will be stored in training module with some identity and name. Now, if the user stands infront of the camera the system will recognize face and it will retrieve the name of the user and produce the speech according to that. Here, we are using haarcascade algorithm to detect the faces. While to produce the speech we are using gtts module that will convert text to speech. These are all modules provided by opencv library to detect the real time computer vision.



V. IMPLEMENTATION

The objective of this project is to detect and recognize the faces from the scene and give the person name as voice output to the user. It was implemented by modules.

It has Four modules:

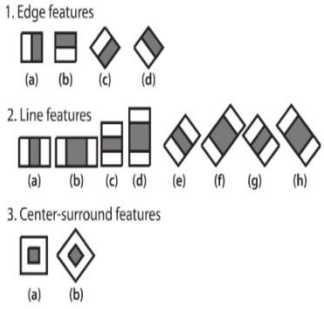
1. Creator
2. Trainer
3. Recognizer
4. Speech Production

1. Creator Module

This module is used to create datasets i.e. we save faces and their names. It first takes the id and name of the person and stores in a file (datatext file). Then it will detect the faces from scene using a pre-executed xml file called haarcascade_frontalface_default (included in OpenCV file). It takes photos in range 50-100 (this range is based on my personal experience but you may take more if you want) and saves all photos in grayscale for better feature extraction.

A. Haarcascade Algorithm

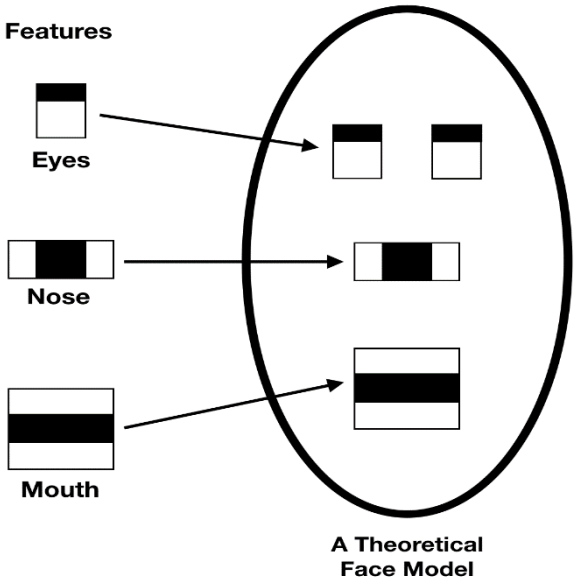
A Haar Cascade is based on "Haar Wavelets". A sequence of rescaled "square-shaped" functions which together form a wavelet family or basis. It is based on the Haar Wavelet technique to analyze pixels in the image into squares by function. This uses machine learning techniques to get a high degree of accuracy from what is called "training data". This uses "integral image" concepts to compute the "features" detected. Haar Cascades use the **Adaboost** learning algorithm which selects a small number of important features from a large set to give an efficient result of classifiers.



Face Detection determines the locations and sizes of human faces in arbitrary (digital) images.

In **Face Recognition**, the use of Face Detection comes first to determine and isolate a face before it can be recognized.

As I mentioned earlier, Haar Cascades use machine learning techniques in which a function is trained from a lot of positive and negative images. This process in the algorithm is feature extraction.



The training data used in this project is an XML file called: haarcascade_frontalface_default.xml. We will apply three Haar cascades to a real-time video stream. These Haar cascades reside in the cascades directory and include:

- **haarcascade_frontalface_default.xml:** Detects faces.
- **haarcascade_eye.xml:** Detects the left and right eyes on the face.
- **haarcascade_smile.xml:** While the filename suggests that this model is a "smile detector," it actually detects the presence of the "mouth" on a face.

2. Trainer Module

Datasets are used for training the model. It does by first loading the LBPHFaceRecognizer which is an algorithm called LBPH (Local Binary Pattern Histograms).

Which is used for finding patterns in the image and remembering them. After loading, it then separates the name and associated image and gives it to train() for training the model to remember the image for the given name. The trained data is saved in yml format.

A. LBPH Face Recognizer

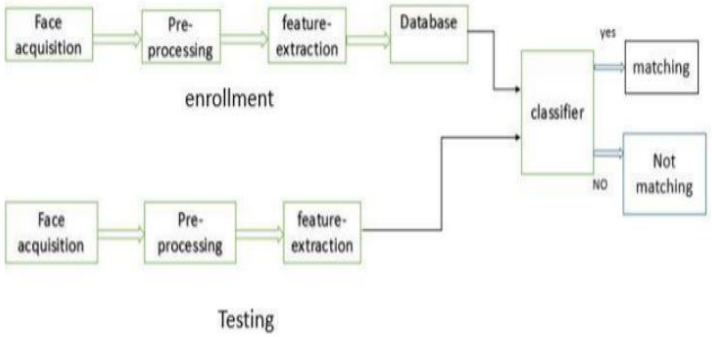
Local Binary Patterns Histogram algorithm was proposed in 2006. It is based on local binary operator. It is widely used in facial recognition due to its computational simplicity and discriminative power.

The steps involved to achieve this are:

- creating dataset
- face acquisition
- feature extraction
- classification

The LBPH algorithm is a part of opencv.

Steps



- Suppose we have an image having dimentions N x M.
- We divide it into regions of same height and width resulting in m x m dimension for every region.
- Local binary operator is used for every region. The LBP operator is defined in window of 3x3.

$$LBP(x_c, y_c) = \sum_{p=0}^{P-1} 2^p s(i_p - i_c)$$

- here '(Xc,Yc)' is central pixel with intensity 'Ic'. And 'In' being the intensity of the the neighbor pixel.
- Using median pixel value as threshold, it compares a pixel to its 8 closest pixels using this function.

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

3. Recognizer

This module is responsible for recognizing and giving voice output. It first detects the faces from the scene and gives to the predict() function which returns id and confidence with which it thinks the detected face(s) belong to that id(s). The received id(s) are examined to find out to which name(s) it belongs and gives to say() which is a function of pyttsx for voice output to the user. It continues until user quits.

4. Speech Production

This is the final module where is helps to convert the text to speech using gtts module when the face is detected. Now let us know about gtts module.

A. gtts Module

gTTS (Google Text-to-Speech)is a Python library and CLI tool to interface with Google Translate text-to-speech API. The **text** variable is a string used to store the user's input. The **tts** variable is used to perform the Google text-to-speech translation on the user's input. The output of the converted text is stored in the form of speech in the tts variable. The tts.save function allows us to save the converted speech in a format that allows us to play sounds.

The gTTS module can be used extensively on other languages such as French, German, Hindi, etc., as well. This is extremely useful when there is a communication barrier and the user is unable to convey his messages to people. Text-to-speech is a great help to the visually impaired people or people with other disabilities as it can help them by assisting in the text to speech translation. There are also many ideas possible with the gTTS module and it can be used for other languages as well.

IV. CONCLUSION AND FUTURE WORK

The project titled as “Machine learning speech production system based on facial recognition” is a console based application. This software detects and recognizes the faces and produces voice as output. So, we have used machine learning algorithm for detecting and recognizing the faces and we have trained our faces in different shades so that system can detect and also giving the voice output as per that. Our algorithm detects the faces more or less with having 99% accuracy. Once faces are detected it can produce a voice output as per that. This becomes one of the best experience for the individual who comes over this.

This software is developed with scalability in mind. The software is developed with modular approach. All modules in the system have been tested with valid data and invalid data and everything work successfully. Thus the system has fulfilled all the objectives identified and is able to replace the existing system.

The constraints are met and overcome successfully. The system is designed as like it was decided in the design phase. The project gives good idea on developing a full-fledged application satisfying the user requirements.

The system is very flexible and versatile. Validation checks induced have greatly reduced errors. Provisions have been made to upgrade the software. The application has been tested with live data and has provided a successful result. Hence the software has proved to work efficiently.

In future we can use other technologies by downloading the modules directly into the project files and implement them. One can use this software for attendance system in future by adding some modules and even they can implement other . One can even use this software to identify mask, check students I’d card by adding some other modules. One can also use this software for warning the people who are not following the traffic signals. The software can be developed further to include lot of modules because the proposed system is developed on the view of future. We can connect to other datasets by including them .

REFERENCES

[1] Blais C., Jack R. E., Scheepers C., Fiset D., Caldara R. (2008). Culture shapes how we look at faces. PLoS ONE 3:e3022. 10.1371/journal.pone.0003022 - DOI - PMC - PubMed

[2] R. Baron, “Mechanisms of human facial recognition,” *International Journal of Man-Machine Studies*, pp. 137–178, 1981.Google Scholar

[3] Boduroglu A., Shah P., Nisbett R. E. (2009). Cultural differences in allocation of attention in visual information processing. J. Cross Cult. Psychol. 40, 349–360. 10.1177/0022022108331005 - DOI - PMC - PubMed

[4] Caldara R. (2017). Culture reveals a flexible system for face processing. Curr. Dir. Psychol. Sci. 26, 249–255. 10.1177/0963721417710036 - DOI.

[5] J. C. Bartlett and A. Fulton, “Familiarity and recognition of faces in old age,” *Memory and Cognition*, Vol. 19, No. 3, pp. 229–238, 1991.Google Scholar.

[6] Effect of Digital Highlighting on Reading Comprehension Given Text-to-Speech Technology for People with Aphasia. Brown JA, Knollman-Porter K, Hux K, Wallace SE, Deville C.

[7] Yu Tiecheng. The current development of speech recognition [J]. Communication World, 2005.Google Scholar

[8] C. S. T. Thu and T. Zin, "Implementation of Text to Speech Conversion", *International Journal of Engineering Research & Technology*, vol. 3, no. 3, pp. 911-915, 2014. Google Scholar