U-GO

AI4GA10 - AI Based Assessment Tool for Presentation Skills

Submitted By

Talha Tariq	*(18-CP-44)*
Muhammad Usman	*(18-CP-56)*
M. Salman Tahir	*(18-CP-58)*

Project Supervisor

Prof. Dr. M. Haroon Yousaf

DEPARTMENT OF COMPUTER ENGINEERING
UNIVERSITY OF ENGINEERING AND TECHNOLOGY TAXILA

July 2022

U-GO

AI4GA10 - AI Based Assessment Tool for Presentation Skills

Authors/Submitted By

Talha Tariq	(18-CP-44)
Muhammad Usman	(18-CP-56)
M. Salman Tahir	(18-CP-58)

Project Supervisor

____

Prof. Dr. M. Haroon Yousaf

DEPARTMENT OF COMPUTER ENGINEERING
UNIVERSITY OF ENGINEERING AND TECHNOLOGY TAXILA

July 2022

Abstract

U-GO is an Artificial Intelligence based tool for the assessment of presentation skills. Presentations are an effective way to showcase a person's ideas and knowledge. Good presentation skills not only help people in their careers, but also help build effective Verbal and Non-Verbal communication skills which play a key role in their personality development. In this era, learning any new technical skill is just a click away. Many online resources are available which are easily accessible and free to use. This means that there is a high competition in the market and people who can communicate better will get the better opportunities. Keeping their importance in mind, we see that there is still no such tool or platform available through which people can practice or evaluate these soft skills. Lack of such a tool affects students, teachers, educational institutes, corporate sector, and departments like marketing, HR, and sales. This effect is seen in a greater magnitude in countries like Pakistan where resources are already scarce and not much effort is made on promoting the development of its students or professionals. Such regions require that a low maintenance system be provided to them so that they can deploy it in the local infrastructure with minimum effort.

As a solution, we propose an Automated AI based Tool for Assessment of Presentation and Communication Skills. The tool is simple to use, a recorded video of a presentation, interview or test should be uploaded on the platform. The video will then be analyzed on both visual and audio aspects like body language, body movements, facial emotions, and speech features like rate of speech and variation of pitch of voice. Next, all these features would be scored and displayed on a dashboard to the user with an overall score (out of 10) as well.

Undertaking

We certify that research work titled "*U-GO AI-Based Tool for Assessment of Presentation Skills*" is our own work. The work has not, in whole or in part, been presented elsewhere for assessment. Where material has been used from other sources it has been properly acknowledged/ referred.

______________________	______________________	______________________
Talha Tariq	Muhammad Usman	M. Salman Tahir
18-CP-44	18-CP-56	18-CP-58

Acknowledgement

First and foremost we are extremely grateful to our supervisors, Prof. Dr. Muhammad Haroon Yousaf for his invaluable advice, continuous support, and patience during our final year project. His immense knowledge and plentiful experience have encouraged us in all the time of our academic research and daily life. We would also like to thank Mr. Saad Saeed for their technical support on our study. We would like to thank all the members in the SWARM Robotics Lab UET Taxila. It is their kind help and support that have made our study and life in the lab a wonderful time. Finally, we would like to express our gratitude to our parents. Without their tremendous understanding and encouragement in the past few years, it would be impossible for us to complete our study.

Table of Contents

Abstract 3
Undertaking 4
Acknowledgement 5
Table of Contents 6
1.0 Introduction 8
1.1 Problem Statement 8
1.1.1 Presentation Skills and Students 8
1.1.2 Presentation Skills and Teachers 8
1.1.3 Presentation Skills and Industry Professionals 9
1.2 Background of Study 9
1.2.1 Factors of a Good Presentation 10
1.2.2 Types of Presentations 11
1.3 Aims and Objectives 12
2.0 Literature Review 14
2.1 Face Detection 14
2.1.1 Haar Cascade 14
2.1.2 Dlib using CNN 16
2.1.3 MTCNN 17
2.2 Posture Estimation 18
2.2.1 OpenPose 18
2.2.2 BlazePose 19
2.3 Speech Analysis 21
2.3.1 My-Voice Analysis Library 21
2.3.2 Speech Recognition Lib 21
2.4 Sequential Data Handling 22
2.4.1 Recurrent Neural Networks 22
2.4.2 Long Short-Term Memory Model 24
3.0 Implementation 26
3.1 Finalized Approach 26
3.2 Working of Algorithm 26
3.3 AI Pipeline of Algorithm 27
3.3.1 Visual Stream 27
3.3.2 Audio Stream 28
3.4 Scoring Mechanism 28

3.4 User Interface and Dashboard 28
4.0 Conclusions 29
References 30

1.0 Introduction

Presentation skills can be defined as the abilities that people use to deliver information to different kinds of audiences in an effective and engaging manner. Presentation skills involve organizing your time, using body language, choosing the proper presentation material, answering event attendees' questions, and providing audience participation.

Presenting information clearly and effectively is a key skill in getting your message across. Today, presentation skills are required in almost every field, and most of us are required to give presentations on occasions. While some people take this in their stride, others find it much more challenging. Many people feel terrified when asked to talk in public, especially to bigger groups. However, these fears can be reduced by good preparation, which will also lay the groundwork for making an effective presentation. It is, however, possible to improve your presentation skills with a bit of work. This research is done in such a way to provide solution for the above.

1.1 Problem Statement

Fear of presenting yourself in public is the number one phobia and 77% of people around the globe suffer from it. Presentation skills are critical to one's success at work. Lack of these skills is affecting our academic as well as our corporate sector. Many students these days are technically sound, yet they are held back by their inability to present their knowledge in front of people. Our educational institutions do not prioritize presentation skills and so, we are neglecting the talent and skills of our youth.

1.1.1 Presentation Skills and Students

A presentation is an integral part of modern classrooms. Having presentations helps the students practice their language systems, improve public-speaking skills, learn how to find, and organize information, create slides, and share information with the audience. Therefore, developing presentation skills is very important in the developmental stage of students. Making a presentation is a great way for students to become more confident in themselves.

During the process of presentation preparation, they learn plenty of useful technical skills, and sometimes they even learn scientific presentation techniques that can be useful in their career development. Developing presentation skills will help the students learn how to express their ideas clearly and become more creative. All these things pave the way to the future world of work [1].

1.1.2 Presentation Skills and Teachers

Another pressing issue is the high pupil-teacher ratio in Pakistan (44.08 in 2018 – reported by World Bank) [2]. According to the OBE system of education (based on Washington Accord), the teacher must evaluate the interpersonal skills of each student based on the rubrics set in the Affective Domain [3]. With such a high pupil-teacher ratio, it is difficult for the teachers to individually target every student while also keeping in mind all the rubrics and providing them with valuable feedback as well.

In our region, due to limited resources, the quality and quantity of skilled professionals is low. There is a need of a system that can operate with minimum effort and can promote the development of these people who are an asset to the nation. The lack of communication skills affects people from multiple domains, in case of students, they rely heavily on their institutes to equip them with these skills. Unfortunately, not all educational institutions are up to the mark in our region and as a result students are left on their own. Lack of such skills affects their careers negatively, and they lose a lot of good opportunities.

1.1.3 Presentation Skills and Industry Professionals

Communication is the soul of every firm. Everything you do in the organization results from communication. So good reading, speaking, listening, presenting and writing skills are necessary if targets are to be accomplished and to achieve the goal. As you advance your course you will find numerous reasons why successful communication skills are important to you [4].

Most of the people in developing nation or even in developed ones face the problem of effective communication. It may be due to lack of knowledge, illiteracy, low confidence, negative impression in the society and so on. According to various research developments around the globe, we tend to have wrong assumption about communication. Communication doesn't mean that every individual needs to speak grammatically correct; Communication is how you explain your thought through your words and gestures to make the other person understand your views about anything.

There is a significant demand of professionals that have excellent Verbal and Non-Verbal communication skills in fields like marketing, sales, and HR. Here we observe that there is lack of such a tool in the market that can provide training to the existing personnel in these fields or act as a screening test for hiring new ones.

Similarly, imagine you have a business meeting to finalize a funding opportunity with some investors. The idea of the company and its work is great, but you still lost the opportunity, just because you were not able to deliver the idea to the investors. That is a big loss for the company. Recognizing the issue, u-GO uses state-of-the-art Artificial Intelligence algorithms and a Progress Tracking System to aid the communication skill growth of its users. u-GO will provide a platform for self-evaluation of individuals to achieve confidence and excel in professional life. Through the detailed analysis and guidance of u-GO, any individual, be it a business professional or a student, can improve the delivery of their content. u-GO also aims to be a benchmark to aid HR departments in evaluating the communication skills of any candidate to judge their competency.

1.2 Background of Study

Communication and presentation skills are essential for everyone either it's a students or a professional to develop a persuasive personality. These interpersonal skills are polished during their academic and professional career paths. There are several activities practiced in academia that help to develop such interpersonal skills. Having strong communication skills aids in all aspects of life – from professional life to personal life and everything that falls in between. From a business standpoint, all transactions result from communication. Good communication skills are essential to allow others and yourself to understand information more accurately and quickly. In contrast, poor communication skills lead to frequent misunderstanding and frustration. In a 2016 LinkedIn survey conducted in the United States, communication topped the list of the most sought-after soft skills among employers [5].

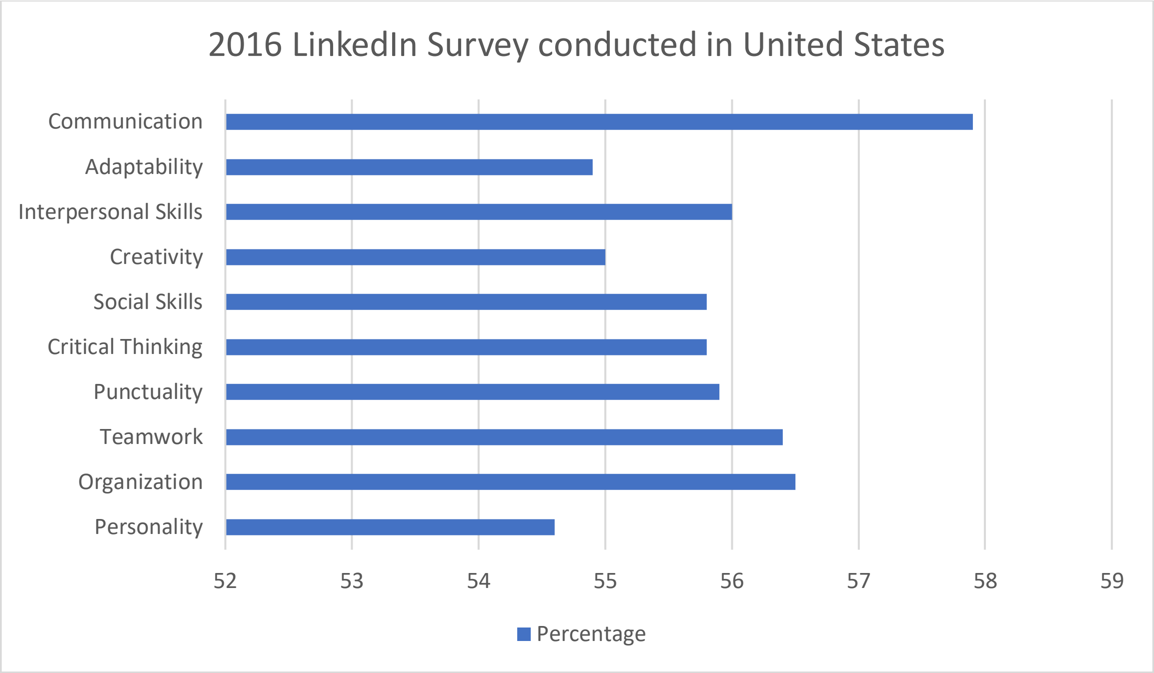


Figure 1.0 – LinkedIn Survey on most sought-after skills in employees

Among them, 'Presentation' is the most used practice. Unfortunately, our institutions do not target the development of these skills in students due to which they are at risk of falling behind their peers. U-GO solved this problem by providing an AI-based tool for assessment of presentation skills.

1.2.1 Factors of a good presentation

A good presentation contains many factors. Some of the factors that we are targeting are stated below,

Visual Communication

Visual information is the first thing that forms the audience's impression. Good presentations include effective and influential slides. Your presentation should not contain more than 100-word text slides.

Interpersonal Communication

Giving a good presentation is unimaginable without building a rapport with the audience. Effective interpersonal communication means convincing each member of the audience that you're speaking directly to them.

Verbal Communication

Speaking in a clear and confident way is key to delivering your message to the audience. Verbal communication is the most obvious part of our communication and plays a major role in presentation as well.

Audience Interaction

To inspire your audience, you need to attract them in the first place. A great balance between verbal and non-verbal communication, as well as engaging visual materials, lead presenters to attract the audience. Your presentation should be centered on the audience, not on yourself. The main purpose of the presentation is to engage the audience. So, good presenters pay attention to the needs and preferences of the people listening to them.

Body Language

Body language accounts for as much as 55% of our communication. Effective presenters try to control their bodies, gestures, and pose to make an impact on the audience.

Emotion Management

Sometimes people don't like the way you present things or are not interested in the topic. However, rejection shouldn't affect your self-image. Effective presenters know how to deal with the stress of public speaking and manage their emotions.

1.2.2 Types of presentations

Even though the majority of presentations aim to explain something to the audience and inspire them, presentations have plenty of different forms and purposes. Understanding the most common types of presentations across different industries is useful for organizing your information better and determining how to deliver your message to the audience.

Nowadays, there are six major types of presentations that differ from each other in structure and purpose.

Providing Information

Informative presentations are the most common type. It's an educational, to-the-point presentation that aims to provide new information or updates to the audience. For instance, it could be public consultation meetings about an upcoming project in the company.

The main goal here is to share information instead of entertaining or inspiring the audience. Informative presentations often include a few educational slides on a slide sorter with short and on-point information. However, you can add presentation notes in order to help your audience better understand the topic and generate questions.

Teaching a Skill

Teaching new skills during public meetings is especially frequent in companies. In this case, as well, a presenter is sharing information with the team. However, the purpose isn't only to share information but to instruct the audience on a particular topic.

The audience is attending the presentation in order to learn new skills and have a better understanding of the topic. Teaching new skills can take the form of training, workshops, or webinars.

On each occasion, the presenter offers specific instructions to help the audience use the new information in practice. Using techniques such as mind maps could be useful in the process of teaching new skills.

Reporting Progress

Reporting progress has a bit different structure than standard presentation structures. Each team leader or company leader wants to know how their team is working. Therefore, the practice of scheduling presentations about the team's progress is very common in the business industry.

Of course, the leader won't assess the individual strengths and weaknesses of team members, but they might point out important achievements or overall drawbacks. So, if you want to report the progress of a new campaign or project, don't forget to keep the outcome in mind and use as much informative visual information as possible.

Selling a Product or Service

Selling a product or service via presentation is an integral part of marketing campaigns. Product sales presentations usually start with the introduction of the company and point out its main ideas and purposes. Then in the mid-presentation, a product or service is introduced to the audience. This type of presentation is a useful way to increase awareness about your product and attract potential customers.

Decision Making

Decision-making presentations are used when the company has a specific problem or problems that need to be discussed among the team members. In this case, the presentation points out the problem, includes existing reasons for the problem, and provides possible options for the solution.

Solving a Problem

This type is very similar to the previous one, but the difference is that during a solution-oriented presentation, the presenter and the audience work towards solving the problem during the presentation. Here as well, the problem is initially identified, and possible solutions and recommendations are discussed.

1.3 Aims and Objectives

The purpose of presentation skills is to help people deliver their message to the audience. Great presentation skills are one of the most worthwhile abilities in modern workplace culture. The reason is that presentations are an integral part of today's work ethic. Consequently, they have a great impact on job performance, employee satisfaction, and work engagement. Aside from organizations, presentations are also very common for high school, college, or university students. Often, students are afraid of speaking publicly and think that they don't have the proper abilities to present the information. Regardless of the reason for public speaking-related anxiety, presentation skills can be improved. Our aim is to provide a platform to the society where they can exercise these skills. This will have a direct impact on the social life of an individual as well as the society. This will result in better career opportunities for students and professionals.

The six presentation goals are:

- To inform
- To educate
- To persuade or convince
- To activate

- To inspire or motivate
- To entertain

We are aiming to enable the students, professionals, or any individual to achieve the above six goals of a good presentation and our tool will help them achieve these goals. We aim to develop he abve in individual, be it a student or an industry professional so that they can communicate clearly, effectively, and confidently with a range of audiences in a range of different contexts. We aim to enhance their verbal and non-verbal skills and increase their confidence level.

2.0 Literature Review

In this modern era, every aspect of human living is being automated. Human lives are getting more and more complex. The task of human recognition and detection has also been actively researched to get more better and fast results. In this paper, we provide an in-dept view of different aspects that help in the detection and recognition of a human body. Some of these aspects are face detection, emotion classification, human pose estimation etc.

2.1 Face Detection

In computer vision, face detection has been an active part of research in the past decade. Face is the most important part of a human body. Face detection has numerous applications from biometrics, facial recognition, verification, security surveillance and access control to human interaction and communication. Apart from the fact that each human being has a different face, our face plays a main role in our daily life. Below is a detailed analysis on all the research done in the previous years and discussion about different face detection algorithms or models.

2.1.1 Haar Cascades

Haar Cascades, also known as Viola-Jones Face Detection Technique, is an object detection algorithm used to detect object in images or videos. It was proposed by Paul Viola and Michael Jones in 2001 in one of their papers, "Rapid Object Detection using a Boosted Cascade of Simple Features" [6]. It was the first ever object detection framework for face detection in real time footage. Their focus was face detection but still the algorithm can detect other objects like cars, fruits, buildings etc.

First, we train the models by images with and without faces as positive and negative images respectively. Then we extract the features using five different feature extractors listed below. Each feature obtained is in form of a single value that we get by subtracting sum of pixels under white rectangle from sum of pixels under black rectangle.

The **Haar Features** travel across the frame from top left to bottom right to search and extract features. It travels pixel by pixel over the image. Depending on the features that we want to extract, we use one of the five feature extractors. It can be seen in Figure 1.0. These feature extractors are classified into three groups; Edge Features for finding out vertical and horizontal edges, Line features for finding out that if there is lighter region surrounded by dark region and vice versa, and Four-rectangle features for finding out pixel intensity change across diagonals.

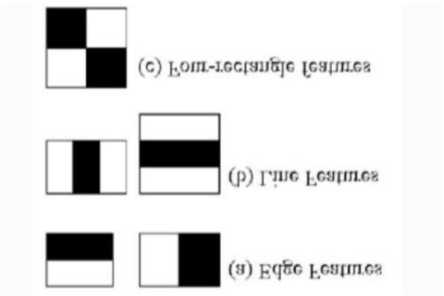


Figure 2.0 – Types of Haar Features

That is basically all the working and functionality of haar feature extractors. After a complete traversal, these features are very hard to determine. To solve this problem, we use integral images to minimize the number of calculations performed in extracting the features.

Integral images help us extract Haar features more efficiently. Instead of computing at every pixel, it instead creates sub-rectangles and creates array references for each of those sub-rectangles. These are then used to compute the Haar features.

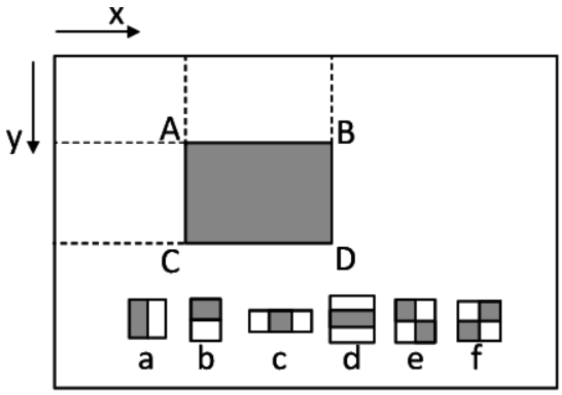
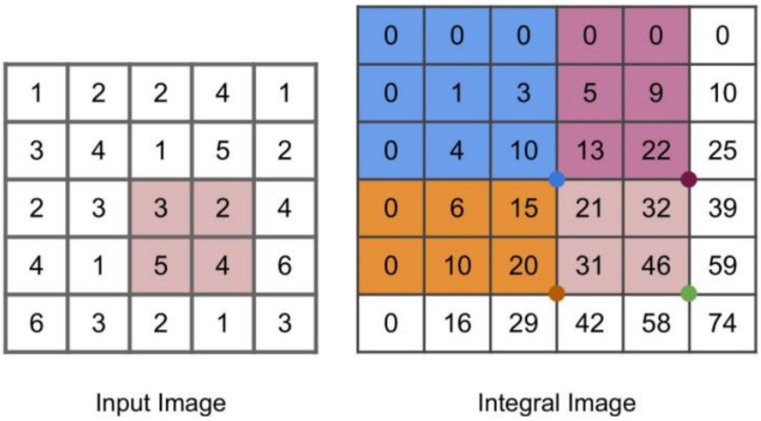


Figure 2.1- Illustration for how an integral image works

An Integral Image is calculated from the original image in such a way that each pixel in this is the sum of all the pixels lying in its left and above in the original image. This can also be seen in Figure 1.2. Each pixel in integral image is the result of sum of all pixels on its left side and all pixels above it including the value of pixel of input image itself.



1	2	2	4	1
3	4	1	5	2
2	3	3	2	4
4	1	5	4	6
6	3	2	1	3

Input Image

0	0	0	0	0	0
0	1	3	5	9	10
0	4	10	13	22	25
0	6	15	21	32	39
0	10	20	31	46	59
0	16	29	42	58	74

Integral Image

Figure 2.2- Calculating integral image from input image

When extracting features in an image, most of the features are irrelevant as we are only concerned about the features of the object that we want to detect. Now the question is that how do we select best features out of hundreds of thousands of features that we get from an image? This is achieved using Adaboost.

Adaboost is a boosting Feature Selection technique, to select a subset of features from the huge set which would not only select features performing better than the others, but also will eliminate the irrelevant ones. Boosting refers to any method that can combine several weak learners into a strong learner. In this technique, all features are sapplied to the training images. For each feature, it finds the best threshold which will classify the faces to positive and negative. We select the features with minimum error rate, which means they are the features that best classifies the face and non-face images. Final classifier is a weighted sum of these weak classifiers. It is called weak because it alone can't classify the image, but together with others forms a strong classifier. After this technique the final number of features gets reduced.

In an image, most of the region is non-face region. So it is a better idea to have a simple method to check if a window is not a face region. If it is not, discard it in a single shot. Don't process it again. Instead focus on region where there can be a face. This way, we can find more time to check a possible face region.

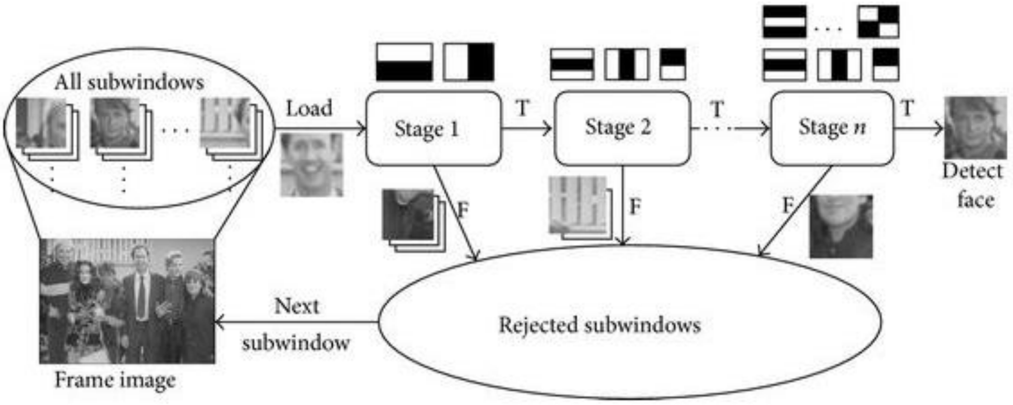


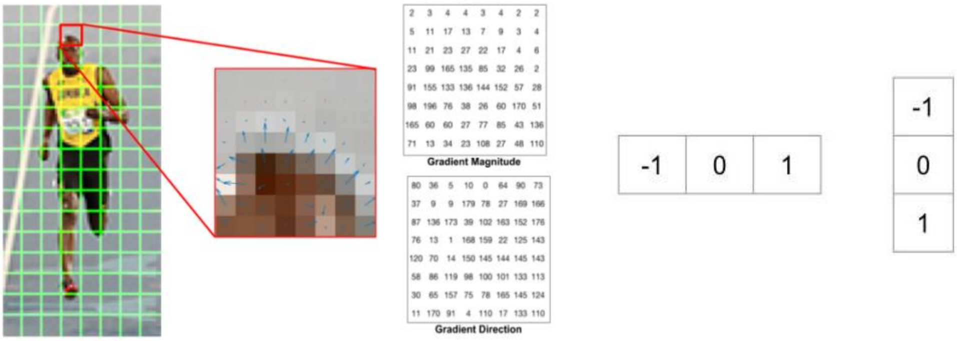
Figure 2.3- Illustration of n-stages in cascade of classifiers

For this they introduced the concept of **Cascade of Classifiers**. Instead of applying all the remaining features on the image, features are grouped into different stages and apply one by one. If a window fails at any stage, it means that the region does not contain face. Hence, it discards it. After that we don't consider remaining features on it. Then we move to the next window where we might find facial features. Only one window which contains a face, runs both the stage and detects the face.

2.1.2 Dlib using CNN

Dlib is a toolkit for making real world machine learning and data analysis applications in C++ [7]. While the library is originally written in C++, it has good, easy to use Python bindings. The dlib library is arguably one of the most utilized packages for face recognition. A Python package appropriately named face_recognition wraps dlib's face recognition functions into a simple, easy to use API. The frontal face detector in dlib works really well. It is simple and just works out of the box. This detector is based on histogram of oriented gradients (HOG) and linear SVM. While the HOG+SVM based face detector has been around for a while and has gathered a good amount of users. HOG based face detector in dlib will not detect faces at odd angles. It is meant to be a good "frontal" face detector and it is, indeed. It detects faces even when they are not perfectly frontal to a good extend. Which is really good for a frontal face detector. But you can only expect so much from it.

The CNN based detector is capable of detecting faces almost in all angles. Unfortunately it is not suitable for real time video. It is meant to be **executed** on a GPU. To get the same speed as the HOG based detector you might need to run on a powerful Nvidia GPU. The Histogram of Oriented Gradients (HoG) + Linear Support Vector Machine (SVM) algorithm in Dlib offers very fast recognition of front-on faces, but has limited capabilities in terms of recognizing face poses at acute angles (such as CCTV footage, or casual surveillance environments where the subject is not actively participating in the ID process).



2	3	4	4	3	4	2	2
5	11	17	13	7	9	3	4
11	21	23	27	22	17	4	6
23	99	165	135	85	32	26	2
91	155	133	136	144	152	57	28
98	196	76	38	26	60	170	51
165	60	60	27	77	85	43	136
71	13	34	23	108	27	48	110

Gradient Magnitude

80	36	5	10	0	64	90	73
37	9	9	179	78	27	169	166
87	136	173	39	102	163	152	176
76	13	1	168	159	22	125	143
120	70	14	150	145	144	145	143
58	86	119	98	100	101	133	113
30	65	157	75	78	165	145	124
11	170	91	4	110	17	133	110

Gradient Direction

-1	0	1

-1
0
1

Figure 2.4- Illustration of gradient magnitude and gradient direction

The first step is to compute **the horizontal and vertical gradients** of the image, by applying the following kernels. The gradient of an image typically removes non-essential information. The image is then divided into 8x8 cells to offer a compact representation and make our HOG more robust to noise. Then, we **compute a HOG** for each of those cells. To estimate the direction of a gradient inside a region, we simply build a histogram among the 64 values of the gradient directions (8x8) and their magnitude (another 64 values) inside each region. The categories of the histogram correspond to angles of the gradient, from 0 to 180°. There are 9 categories overall : 0°, 20°, 40°… 160°.

Finally, for **Block Normalization**, a 16x16 block can be applied in order to normalize the image and make it invariant to lighting for example. This is simply achieved by dividing each value of the HOG of size 8x8 by the L2-norm of the HOG of the 16x16 block that contains it, which is in fact a simple vector of length 9*4 = 36. Finally, all the 36x1 vectors are concatenated into a large vector. Now, we have our feature vector, on which we can train a soft **SVM classifier**.

2.1.3 MTCNN Face Detector

MultiTask Cascaded Convolutional Neural Network is a modern tool for face detection, leveraging a 3-stage neural network detector [8]. The MTCNN algorithm works in three steps and use one neural network for each. The first part is a proposal network. It will predict potential face positions and their bounding boxes like an attention network in Faster R-CNN.

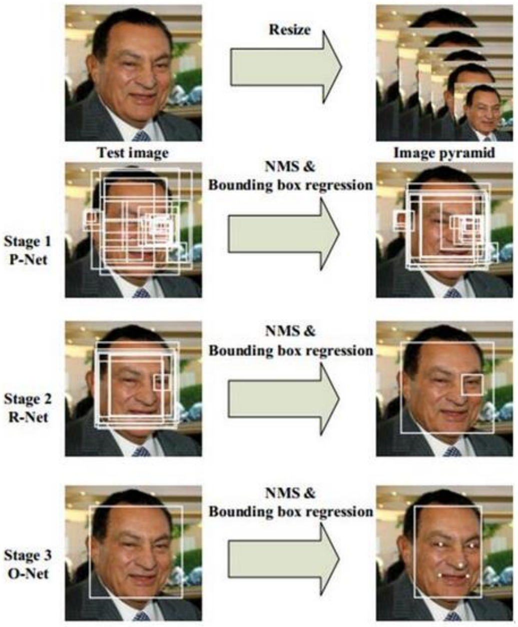


Figure 2.5- Working of MTCNN face detector

The result of this step is a large number of face detections and lots of false detections. The second part uses images and outputs of the first prediction. It makes a refinement of the result to eliminate most of false detections and aggregate bounding boxes. The last part refines even more the predictions and adds facial landmarks predictions

First, the image is resized multiple times to detect faces of different sizes. Then the P-network (Proposal) scans images, performing first detection. It has a low threshold for detection and therefore detects many false positives, even after NMS (Non-Maximum Suppression), but works like this on purpose.The proposed regions (containing many false positives) are input for the second network, the R-network (Refine), which, as the name suggests, filters detections (also with NMS) to obtain quite precise bounding boxes. The final stage, the O-network (Output) performs the final refinement of the bounding boxes. This way not only faces are detected, but bounding boxes are very right and precise.

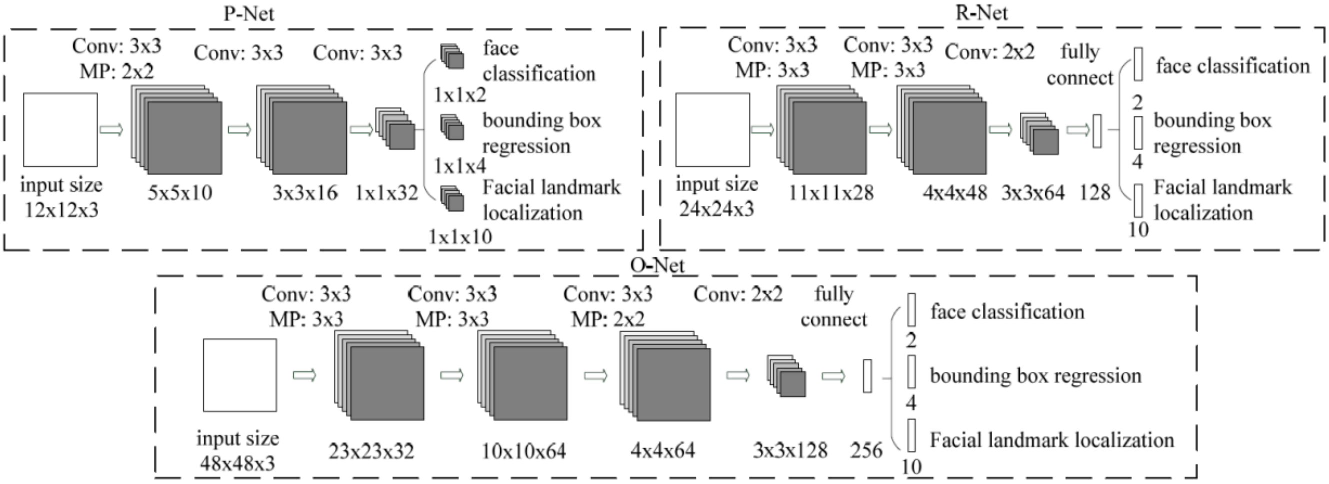


Figure 2.6- Illustration of P-Net, R-Net and O-Net

2.2 Posture Estimation

Pose estimation is a computer vision technique that predicts and tracks the location of a person or object. This is done by looking at a combination of the pose and the orientation of a given person/object. We can also think of pose estimation as the problem of determining the position and orientation of a camera relative to a given person or object. This is typically done by identifying, locating, and tracking a number of **keypoints** on a given object or person. For objects, this could be corners or other significant features. And for humans, these keypoints represent major joints like an elbow or knee. The goal of our machine learning models are to track these keypoints in images and videos.

2.2.1 OpenPose

OpenPose is a real-time multi-person human pose detection library that has for the first time shown the capability to jointly detect the human body, foot, hand, and facial keypoints on single images [9]. OpenPose is capable of detecting a total of 135 keypoints. It was proposed by researchers at Carnegie Mellon University. They have released in the form of Python code, C++ implementation and Unity Plugin.

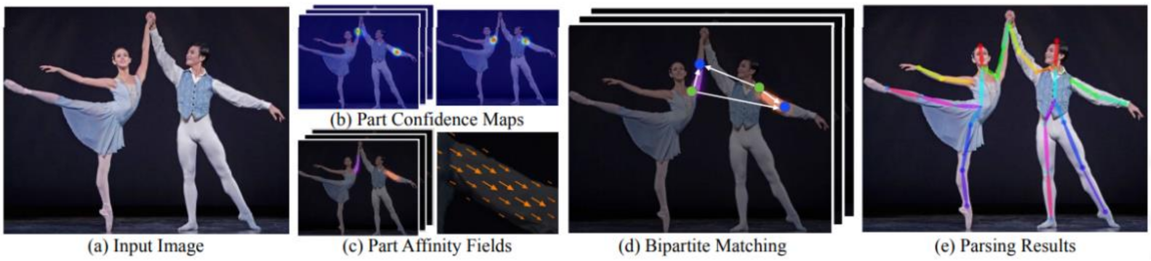


Figure 2.7- Working of OpenPose Estimator

- In first step the image is passed through baseline CNN network to extract the feature maps of the input In the paper. In this paper the authors used first 10 layers of VGG-19 network.
- The feature map is then process in a multi-stage CNN pipeline to generate the Part Confidence Maps and Part Affinity Field
    - Part Confidence Maps
    - Part Affinity Field
- In the last step, the Confidence Maps and Part Affinity Fields that are generated above are processed by a greedy bipartite matching algorithm to obtain the poses for each person in the image.

A **Confidence Map** is a 2D representation of the belief that a particular body part can be located in any given pixel. **Part Affinity** is a set of 2D vector fields that encodes location and orientation of limbs of different people in the image. It encodes the data in the form of pairwise connections between body parts. The above multi-CNN architecture has three major steps:

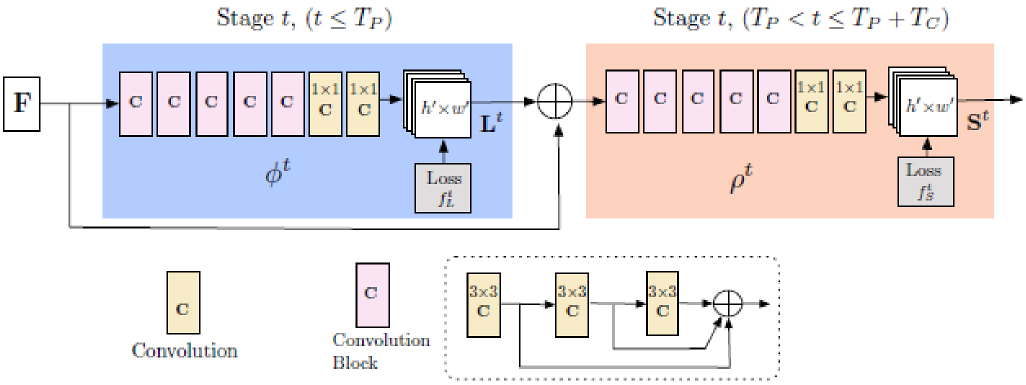


Figure 2.8- Stages of OpenPose Estimator

The first set of stages predicted the Part Affinity Fields refines Lt from the feature maps of base network F. The second set of stages takes use the output Part Affinity Fields from the previous layers to refine the prediction of confidence maps detection. The final S (confidence maps) and L (Part Affinity Field) are then passed into the greedy algorithm for further process.

OpenPose have problems estimating pose when the ground truth example has non typical poses and upside down examples. In highly crowded images where people are overlapping, the approach tends to merge annotations from different people, while missing others, due to the overlapping PAFs that make the greedy multi-person parsing fail.

2.2.2 BlazePose

BlazePose (Full Body) is a pose detection model developed by Google that can compute (x,y,z) coordinates of 33 skeleton keypoints [10]. It can be used for example in fitness applications.

BlazePose consists of two machine learning models: a Detector and an Estimator. The Detector cuts out the human region from the input image, while the Estimator takes a 256x256 resolution image of the detected person as input and outputs the keypoints.

The **Detector** is an Single-Shot Detector(SSD) based architecture. Given an input image (1,224,224,3), it outputs a bounding box (1,2254,12) and a confidence score (1,2254,1). The 12 elements of the bounding box are of the form (x,y,w,h,kp1x,kp1y,…,kp4x,kp4y), where kp1x to kp4y are additional keypoints. Each one of the 2254 elements has its own anchor, anchor scale and offset need to be applied.

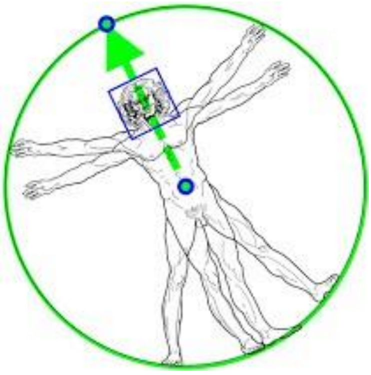


Figure 2.9- Illustration of alignment via two key points and face detection

The **Estimator** uses heatmap for training, but computes keypoints directly without using heatmap for faster inference. The first output of the Estimator is (1,195) landmarks , the second output is (1,1) flags. The landmarks are made of 165 elements for the (x,y,z,visibility,presence) for every 33 keypoints. The z-values are based on the person's hips, with keypoints being between the hips and the camera when the value is negative, and behind the hips when the value is positive.

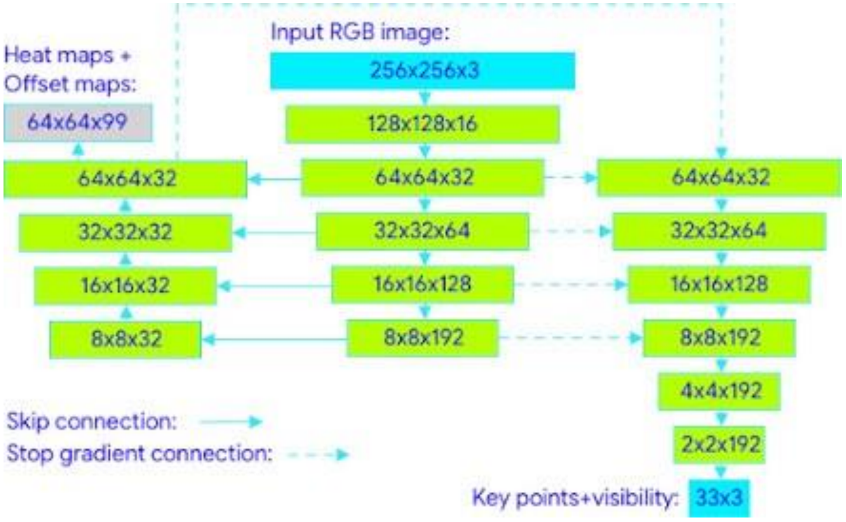


Figure 2.10- Tracking network architecture: regression with heatmap supervision

The visibility and presence are stored in the range of [min_float,max_float] and are converted to probability by applying a sigmoid function. The visibility returns the probablity of keypoints that exist in the frame and are not occluded by other objects. presence returns the probablity of keypoints that exist in the frame.

2.3 Speech Analysis

Speech analysis is the process of transcribing a recorded conversation and analyzing it to derive valuable insights. The conversation can be between any number of people. It doesn't matter, it can be a single person or multiple persons. First, the data is processed, and transcription is done. Transcription can be done both manually and by using an automated algorithms or programs. The data collected after transcription is then analysed according to pre-determined criteria. Our interest can be search for specific keywords spoken by one or both speakers, sentiments displayed by one or both speakers, and many other categories of analysis. A detailed report on the analysis is then delivered, providing us with valuable insights about the conversation and the speaker(s).

2.3.1 My-voice Analysis Library

My-Voice Analysis is a Python library [11] for the analysis of voice (simultaneous speech, high entropy) without the need of a transcription. It breaks utterances and detects syllable boundaries, fundamental frequency contours, and formants. Its built-in functions recognise and measures

- gender recognition,
- speech mood (semantic analysis)
- pronunciation posterior score
- articulation-rate
- speech rate
- filler words

The library was developed based upon the idea introduced by Nivja DeJong and Ton Wempe [12], Paul Boersma and David Weenink [13], Carlo Gussenhoven [14], S.M Witt and S.J. Young [15] and Yannick Jadoul [16]. Peaks in intensity (dB) that are preceded and followed by dips in intensity are considered as potential syllable cores. My-Voice Analysis is unique in its aim to provide a complete quantitative and analytical way to study acoustic features of a speech. Moreover, those features could be analysed further by employing Python's functionality to provide more fascinating insights into speech patterns. This library is for Linguists, scientists, developers, speech and language therapy clinics and researchers. Please note that My-Voice Analysis is currently in initial state though in active development. While the amount of functionality that is currently present is not huge, more will be added over the next few months.

2.3.2 SpeechRecognition Library

The SpeechRecognition interface of the Web Speech API [17] is the controller interface for the recognition service; this also handles the SpeechRecognitionEvent sent from the recognition service. It is a Library for performing speech recognition, with support for several engines and APIs, online and offline.

Speech recognition engine/API support:

- CMU Sphinx (works offline)
- Google Speech Recognition
- Google Cloud Speech API
- Wit.ai
- Microsoft Bing Voice Recognition
- Houndify API

- IBM Speech to Text
- Snowboy Hotword Detection (works offline)

From the above listed API's, the Google speech recognition API was used. It enables developers to convert audio to text in over 125 languages and variants, by applying powerful neural network models in an easy-to-use API.

Figure 2.11- Conversion of speech to text

It enables easy integration of Google speech recognition technologies into developer applications. You can send audio data to the Speech-to-Text API, which then returns a text transcription of that audio file. Speech-to-Text has three main methods to perform speech recognition. These are Synchronous Recognition, Asynchronous Recognition and Streaming Recognition. Speech-to-Text API synchronous recognition request is the simplest method for performing recognition on speech audio data. Speech-to-Text can process up to 1 minute of speech audio data sent in a synchronous request. After Speech-to-Text processes and recognizes all the audio, it returns a response. A synchronous request is blocking, meaning that Speech-to-Text must return a response before processing the next request. Speech-to-Text typically processes audio faster than real time, processing 30 seconds of audio in 15 seconds on average. In cases of poor audio quality, your recognition request can take significantly longer.

2.4 Sequential Data Handling

Sequence prediction problems have been around for a long time. They are considered as one of the hardest problems to solve in the data science industry. These include a wide range of problems; from predicting sales to finding patterns in stock markets' data, from understanding movie plots to recognizing your way of speech, from language translations to predicting your next word on your iPhone's keyboard.

2.4.1 Recurrent Neural Network

Recurrent Neural Network also known as RNN are a type of Neural Network where the output from previous step is fed as input to the current step [18]. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words. Thus, RNN came into existence, which solved this issue with the help of a Hidden Layer. The main and most important feature of RNN is Hidden state, which remembers some information about a sequence.

RNN have a “memory” which remembers all information about what has been calculated. It uses the same parameters for each input as it performs the same task on all the inputs or hidden layers to produce the output. This reduces the complexity of parameters, unlike other neural networks. RNN works on the principle of saving the output of a particular layer and feeding this back to the input in order to predict the output of the layer. Below is how you can convert a Feed-Forward Neural Network into a Recurrent Neural Network:

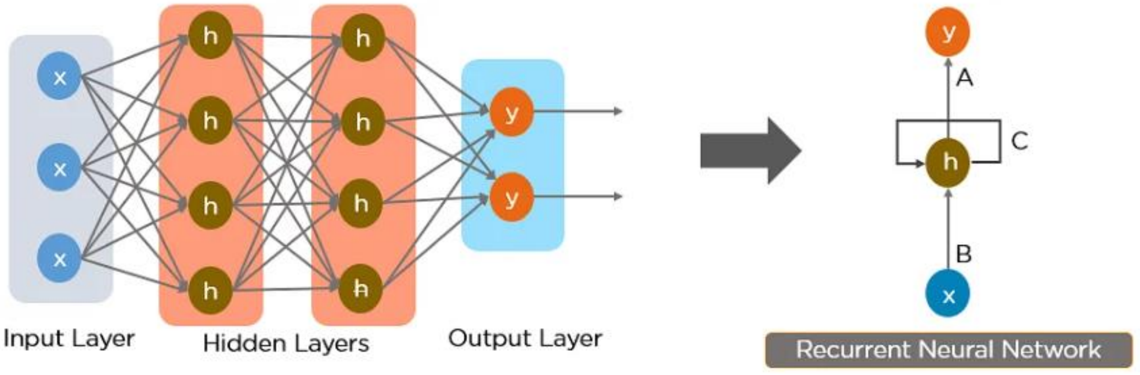


Figure 2.12- Simple Recurrent Neural Network

The nodes in different layers of the neural network are compressed to form a single layer of recurrent neural networks. A, B, and C are the parameters of the network. Here, “x” is the input layer, “h” is the hidden layer, and “y” is the output layer. A, B, and C are the network parameters used to improve the output of the model. At any given time t, the current input is a combination of input at x(t) and x(t-1). The output at any given time is fetched back to the network to improve on the output.

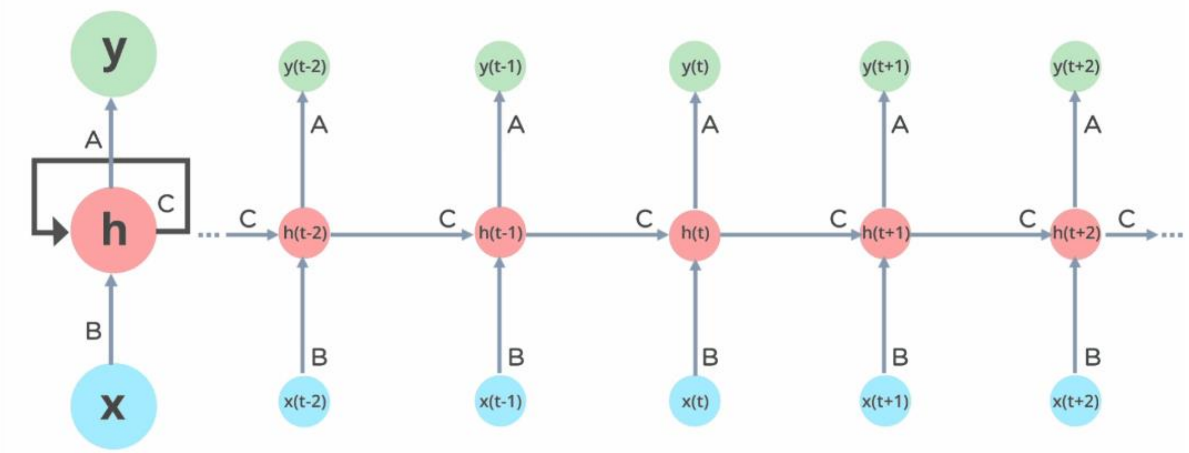


Figure 2.13- Architecture of Recurrent Neural Network

The input layer x receives and processes the neural network’s input before passing it on to the middle layer.

Multiple hidden layers can be found in the middle layer h, each with its own activation functions, weights, and biases. You can utilize a recurrent neural network if the various parameters of different hidden layers are not impacted by the preceding layer, i.e. There is no memory in the neural network.

The different activation functions, weights, and biases will be standardized by the Recurrent Neural Network, ensuring that each hidden layer has the same characteristics. Rather than constructing numerous hidden layers, it will create only one and loop over it as many times as necessary [C].

2.4.2 Long Short-Term Memory Models

With the recent breakthroughs that have been happening in data science, it is found that for almost all these sequence prediction problems, long short-Term Memory networks, also known as, LSTMs, have been observed as the most effective solution. LSTMs have an edge over conventional feed-forward neural networks and RNN in many ways. This is because of their property of selectively remembering patterns for long durations of time[19].

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. This is a behaviour required in complex problem domains like machine translation, speech recognition, and more.

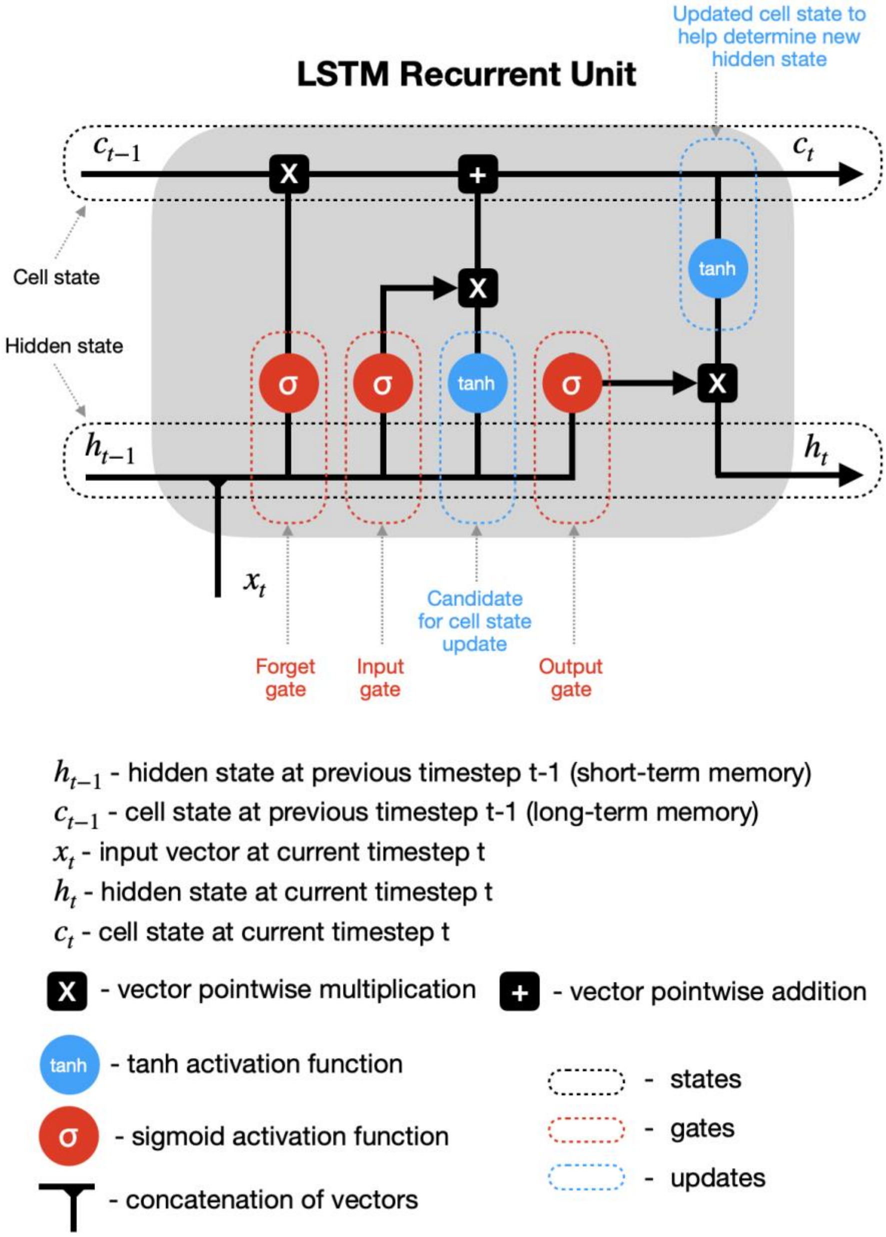

LSTM Recurrent Unit
Updated cell state to help determine new hidden state
c_{t-1}
c_t
Cell state
Hidden state
h_{t-1}
h_t
x_t
σ
tanh
X
+
Forget gate
Input gate
Candidate for cell state update
Output gate
h_{t-1} - hidden state at previous timestep t-1 (short-term memory)
c_{t-1} - cell state at previous timestep t-1 (long-term memory)
x_t - input vector at current timestep t
h_t - hidden state at current timestep t
c_t - cell state at current timestep t
X - vector pointwise multiplication
+ - vector pointwise addition
tanh - tanh activation function
σ - sigmoid activation function
- concatenation of vectors
- states
- gates
- updates

Figure 2.14- Architecture of LSTM's

A typical LSTM network is comprised of different memory blocks called cells. There are two states that are being transferred to the next cell, the cell state, and the hidden state. The memory blocks are responsible for remembering things and manipulations to this memory is done through three major mechanisms, called gates. These gates optionally let the information flow in and out of the cell. It contains a pointwise multiplication operation and a sigmoid neural net layer that assist the mechanism. The sigmoid layer gives out numbers between zero and one, where zero means ‘nothing should be let through,’ and one means ‘everything should be let through.’

Hidden state & new inputs — hidden state from a previous timestep (h_t-1) and the input at a current timestep (x_t) are combined before passing copies of it through various gates.

Forget gate — this gate controls what information should be forgotten. Since the sigmoid function ranges between 0 and 1, it sets which values in the cell state should be discarded (multiplied by 0), remembered (multiplied by 1), or partially remembered (multiplied by some value between 0 and 1).

Input gate helps to identify important elements that need to be added to the cell state. Note that the results of the input gate get multiplied by the cell state candidate, with only the information deemed important by the input gate being added to the cell state.

Update cell state —first, the previous cell state (c_t-1) gets multiplied by the results of the forget gate. Then we add new information from [input gate × cell state candidate] to get the latest cell state (c_t).

Update hidden state — the last part is to update the hidden state. The latest cell state (c_t) is passed through the tanh activation function and multiplied by the results of the output gate.

Finally, the latest cell state (c_t) and the hidden state (h_t) go back into the recurrent unit, and the process repeats at timestep t+1. The loop continues until we reach the end of the sequence.

3.0 Implementation

We needed to design a tool that could analyse a presentation depending upon the facial expressions, body posture and verbal delivery of the speaker. For that purpose, we did research on most efficient models. After a through literature view of the all three aspects we implemented our learning on U-GO. It is an AI-based automated tool that will provide detailed analysis of your presentation in form of a dashboard. where you can see for yourself what you did good and what you need to improve. It will also provide an overall confidence score which can help you track your performance throughout your learning journey. U-GO will be your complete self-assessment tool on your mobile phone as an application and on your computer as a web app. All you need to do is upload a recording of your presentation and the results are just a click away.

3.1 Finalized Approach

Purpose	Approach
Sequential Data Analysis	LSTM
Face Detection	MTCNN
Posture Analysis	BlazePose
Speech Analysis	SpeechRecognition Library

Table 1- Finalized approach for implementation

3.2 Working of Algorithm

The U-GO algorithm takes the input video, face detector detects the face of user and performs emotion recognition. The posture estimation model extracts the landmarks and analyse the body language based on the position of those landmarks. For the face detection, we used the MTCNN model and for posture estimation, we used BlazePose by MediaPipe. The role of BlazePose here is the extraction of landmarks. We have trained an LSTM model. The LSTM model was trained by custom dataset that was created by us. LSTM was trained on the landmarks extracted from the train set. The video recorded for test purpose is fed into the BlazePose. As a result, we get the landmarks extracted from the video. These landmarks are then fed into the custom trained LSTM. The model analyses the video and predicts the posture of subject in the video. At the same time the audio in the video is extracted and speech analysis is performed on the extracted audio. As a result, useful information like the rate of speech, filler words, unique words and repeated words used is generated.

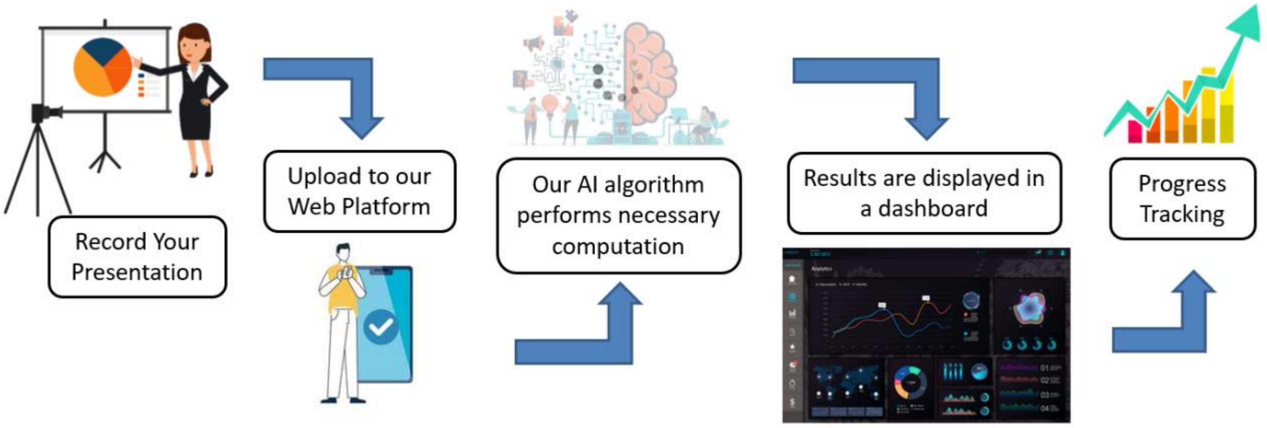

Record Your Presentation
Upload to our Web Platform
Our AI algorithm performs necessary computation
Results are displayed in a dashboard
Progress Tracking

Figure 3.0- Working of U-GO

3.3 AI Pipeline of Algorithm

The AI pipeline can be seen below. The pipeline is further broken into Visual Stream and Audio Stream. The recorded video of a presentation passes through the pipeline and analytics are displayed on the dashboard.

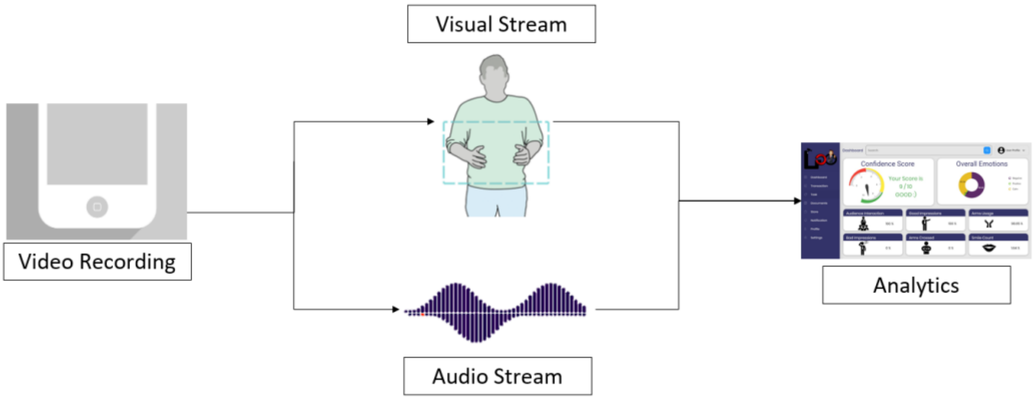


Figure 3.1- AI Pipeline of U-GO

3.3.1 Visual Stream

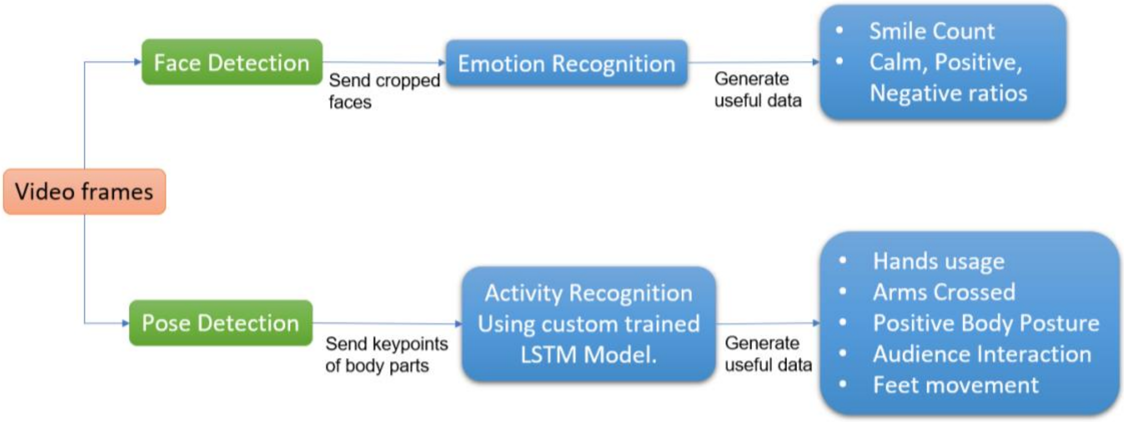


Figure 3.2- Visual Stream of Pipeline

3.3.2 Audio Stream

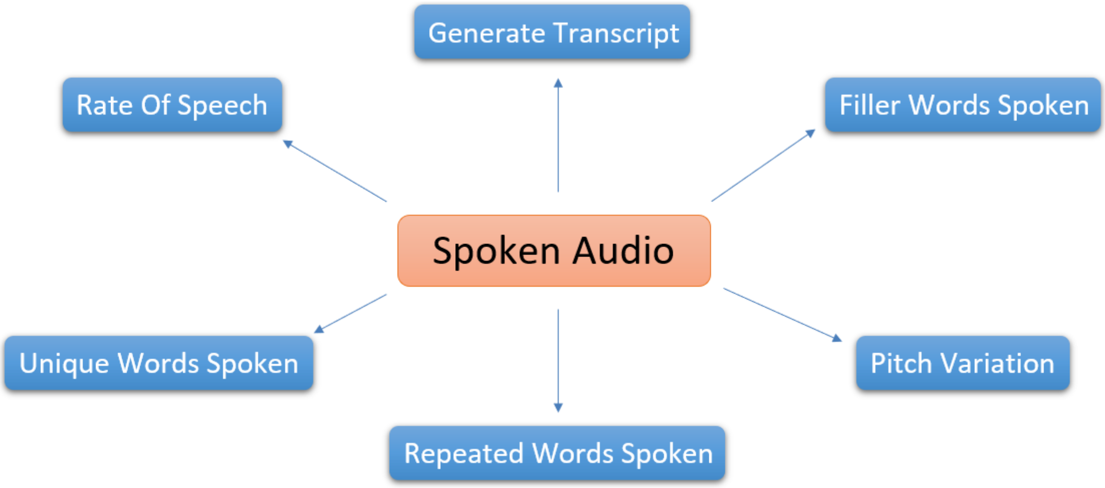


Figure 3.3- Audio Stream of Pipeline

3.2 Scoring Mechanism

Professor Albert Mehrabian has pioneered the understanding of communications since the 1960s. He received his PhD from Clark University and in l964 commenced an extended career of teaching and research at the University of California, Los Angeles [20].

He currently devotes his time to research, writing, and consulting as Professor Emeritus of Psychology, UCLA. Mehrabian's work featured strongly (the mid-late 1900s) in establishing an early understanding of body language and non-verbal communications. Aside from his many and various other fascinating works, Mehrabian's research provided the basis for the widely quoted and often much over-simplified statistic for the effectiveness of spoken communications. Here is a more precise (and necessarily detailed) representation of Mehrabian's findings than is typically cited or applied:

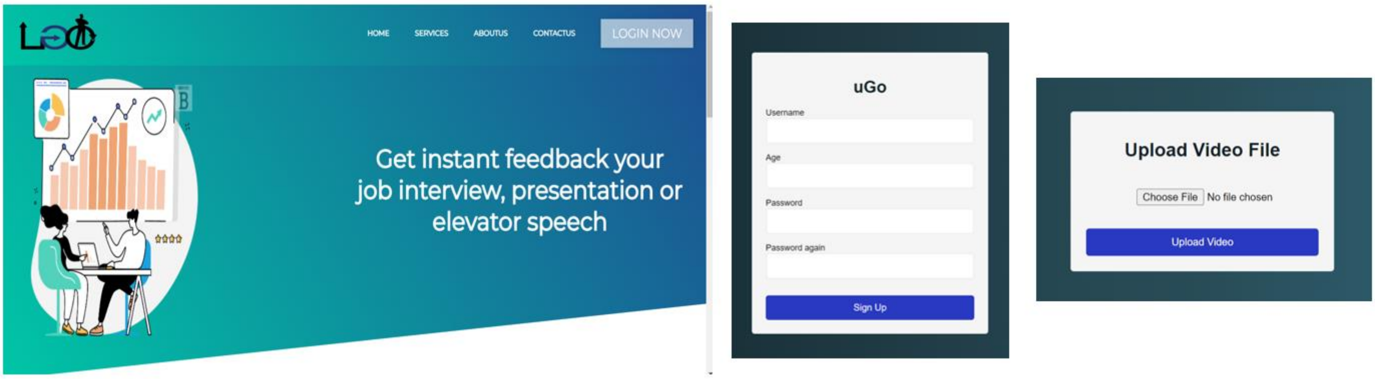
- 7 percent of meaning is communicated through spoken word,
- 38 percent through tone of voice, and
- 55 percent through body language.

Speaker's Impact During Presentation		
Text	**Visual**	**Vocal**
7%	55%	38%

Table 2- Speakers Impact during Presentation

3.3 User Interface and Dashboard

In the end we get a detailed analysis in form of graphs on a dashboard. Our algorithm also assigns an overall confidence score which can later be used to track the performance. The dashboard can be seen in the figure below,



Get instant feedback your job interview, presentation or elevator speech

Figure 3.4- User Interface of U-GO

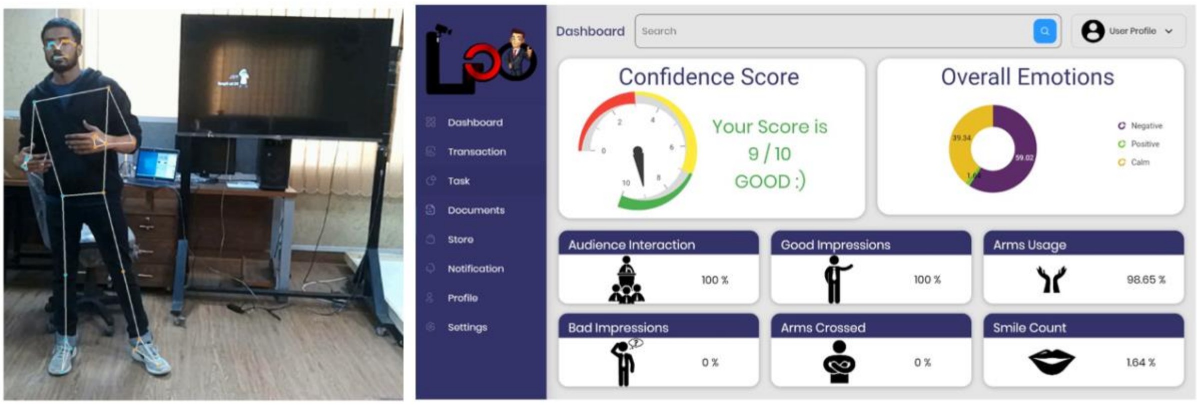


Figure 3.4- Dashboard of U-GO

4.0 Conclusion

Presentation skills are critical for our success. Yet these skills are ignored by our academia. This directly affects the confidence level and inter-personal skills of students. Presentations are major part of our corporate firms. Communication is the soul of every firm. Everything you do in the organization results from communication. If you have effective presentation skills, this means you are good at communicating. By speaking clearly, and getting your ideas and message across to people well, there will be less miscommunication in your life. This means less stress and a healthy professional life.

For this problem, we used Artificial Intelligence (Computer Vision) and machine learning to devise a solution. Based on features like emotions, body posture and verbal delivery, we create an AI-based tool that analyses human behaviour through face emotions, body posture, body movements and verbal delivery. These aspects are then used to classify the presentation as good or bad presentation. At the end, a detailed analysis is generated which highlights the areas in which the person excels or lacks. An overall confidence score is generated and saved for the user so that he can track his progress. Our academic and corporate sector can use this tool to overcome these issues in our society.

References

1. Presentation skills for students https://high5test.com/presentation-skills/
2. Pakistan - Pupil-teacher Ratio, Primary2022 Data 2023 Forecast 1971-2018 Historical https://tradingeconomics.com/pakistan/pupil-teacher-ratio-primary-wb-data.html
3. Kamran, Muhammad & Nisa, Badar & Fazal, Muhammad & Abid, Muhammad & Abid, Irsa. (2020). IMPLEMENTATION OF THE OUTCOME-BASED EDUCATION SYSTEM IN ENGINEERING PROGRAMS FOR PAKISTAN ENGINEERING COUNCIL ACCREDITATION UNDER WASHINGTON ACCORD SIGNATORY. 32. 197-206.
4. Communication skills in Professional world https://madhavuniversity.edu.in/communication-skills.html
5. Guy Berger, Ph.D. Principal Economist at LinkedIn August 30, 2016, Data Reveals the Most In-Demand Soft Skills Among Candidates. https://www.linkedin.com/business/talent/blog/talent-strategy/most-indemand-soft-skills
6. Rapid Object Detection using a Boosted Cascade of Simple Features, Paul Viola, Mitsubishi Electric Research Labs, 201 Broadway, 8th FL, Cambridge, MA 02139 and Michael Jones, Compaq CRL One Cambridge Centre Cambridge, MA 02142
7. Davisking/Dlib https://github.com/davisking/dlib
8. Yang, XianBen & Zhang, Wei. (2022). Heterogeneous face detection based on multi-task cascaded convolutional neural network. IET Image Processing. 16. 10.1049/ipr2.12344.
9. Cao, Zhe & Martinez, Gines & Simon, Tomas & Wei, Shih-En & Sheikh, Yaser. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence. PP. 1-1. 10.1109/TPAMI.2019.2929257.
10. Mroz, Sarah & Baddour, Natalie & McGuirk, Connor & Juneau, Pascale & Tu, Albert & Cheung, Kevin & Lemaire, Edward. (2021). Comparing the Quality of Human Pose Estimation with BlazePose or OpenPose. 1-4. 10.1109/BioSMART54244.2021.9677850.
11. My-voice Analysis Library, https://github.com/Shahabks/my-voice-analysis
12. DeJong N.H, and Ton Wempe [2009]; “Praat script to detect syllable nuclei and measure speech rate automatically”; Behaviour Research Methods, 41(2).385-390.
13. Paul Boersma and David Weenink; http://www.fon.hum.uva.nl/praat/
14. Gussenhoven C. [2002]; “Intonation and Interpretation: Phonetics and Phonology”; Centre for Language Studies, Univerity of Nijmegen, The Netherlands.
15. Witt S.M and Young S.J [2000]; “Phone-level pronunciation scoring and assessment or interactive language learning”; Speech Communication, 30 (2000) 95-108.
16. Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. Journal of Phonetics, 71, 1-15. https://doi.org/10.1016/j.wocn.2018.07.001 (https://parselmouth.readthedocs.io/en/latest/)
17. Zhang, A. (2017). Speech Recognition (Version 3.8) [Software]. Available from https://github.com/Uberi/speech_recognition#readme.
18. A Brief Overview of Recurrent Neural Networks (RNN) https://www.analyticsvidhya.com/blog/2022/03/a-brief-overview-of-recurrent-neural-networks-rnn/
19. LSTM Recurrent Neural Networks — How to Teach a Network to Remember the Past https://towardsdatascience.com/lstm-recurrent-neural-networks-how-to-teach-a-network-to-remember-the-past-55e54c2ff22e
20. Mehrabian, A. (1981) Silent messages: Implicit communication of emotions and attitudes. Belmont, CA: Wadsworth (currently distributed by Albert Mehrabian, email: am@kaaj.com)