I SEE YOU : AN ANALYSIS OF THE BODY LANGUAGE OF A SPEAKER USING COMPUTER VISION

Abdul Hakeem Ahmed
Department of Computer Engineering
University of Sri Jayewardenepura
Ratmalana, Sri Lanka
hakeemahmed166@gmail.com

***Abstract*— Imagine you are standing in front of people, ready to give a speech. You are nervous, you have butterflies in your stomach, and you think to yourself, "I'm so not ready to do this". But what if you could improve your speech beforehand, without even having to speak in front of people? This project aims on improving a major aspect of that speech, the body language. It uses a system that analyzes the body language of a speaker from the video of his/her speech using computer vision techniques. It gives the speaker an analysis of his/her facial expressions, how hand gestures were used and how the stage was utilized during the speech. The speaker can identify seven different types of facial expressions used in the speech and their frequency, two types of hand gestures used in the speech and their frequency and the percentage of each part of the stage utilized during the speech. The speaker can also recognize the stage transitions made during the speech. This system is also applied to analyze the body language of the speeches of several winners of the Toastmasters World Championship of Public Speaking. The results of how these body language components are used by these speakers gives the user the ability to learn from those at the pinnacle of their public speaking journey. By providing such an analysis, this system can greatly improve the body language skills of a speaker without the need of human interaction.**

Keywords—computer vision, body language, public speaking, self-improvement

I. INTRODUCTION

Public Speaking is an extremely important form of communication to transmit information, persuade or entertain people. When presenting a speech, a vast amount of information is visually conveyed by appearance, manner, and physical behavior. In public speaking, the body can be an effective tool for adding emphasis and clarity to words. It is also the most powerful instrument for convincing an audience of the speaker's sincerity, earnestness, and enthusiasm. [1].

Often speakers need an evaluation of their speeches to improve. This is mostly done through peer evaluation. Allowing the speaker, the opportunity to get an evaluation on an individual basis using video analysis would be a great advantage for one's journey as a Public Speaker.

Since body language is a key component in the arsenal of a Public Speaker, this project focuses on analyzing the speech and providing feedback to the speaker of his/her body language. Here we focus on the Toastmasters International Speech format of 5–7-minute short public speeches.

.

II. RELATED WORK

Computer Vision allows computers to see and understand images and videos. It is concerned with extraction, analysis and understanding of useful information from an image or image sequence. This encompasses many sub domains such as image classification, object detection and pose estimation.

YOLO (You Only Look Once) is a particularly popular algorithm that can be used for real-time object detection [2]. Facial detection and facial emotion recognition have also become popular applications of computer vision in recent times. Another significant development in this field is Pose Estimation. Pose estimation is a computer vision technique to detect human figures in images and videos. It determines the positions of the key body joints. It can be done in either 2d pose estimation or 3d pose estimation. Open pose, PoseNet and deepPose are examples of such Pose estimation algorithms [3].

Furthermore, these technologies have been used in various applications. Emotional Body Gesture Recognition System [4] which is an automatic emotion recognition system incorporating Body Gestures can be cited as an example. "Everybody can dance now", a simple method for dancing "do as I do" motion transfer using Poses is another application of such technology [5].

Similarly, using the mentioned domains and technologies in the context of public speaking, this project aims to provide feedback to the speaker regarding his or her Body Language.

III. METHODOLOGY

A. Posenet

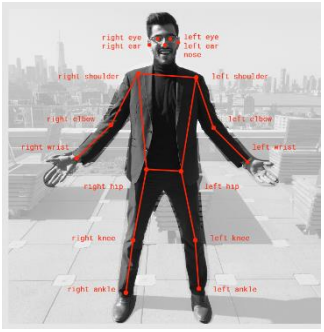

Fig. 1. Seventeen pose key points detected by posenet[6]

For certain parts of the system, pose estimation was used to detect the body joints of the speaker. Pose Estimation is a technique used to estimate the pose of a person from an image or video by estimating the locations of key body joints (key points).

The pose estimation algorithm chosen for this project was a python implementation of PoseNet. This model takes the processed camera image as the input and outputs information about key points. These key points represent 17 body joints of the person such as nose, right wrist, left wrist, right eye, left eye, and so on.

B. Facial Expressions

To detect facial expressions, a model was created using Deep Learning. A dataset from Kaggle of grayscale images of size 48x48 were obtained for this purpose. This consisted of images of faces which was separated into 7 classes representing 7 different emotions - anger, disgust, fear, happy, sad, neutral and surprise. 80% of the dataset was used as training data and the remaining 20% was used as test data. A Convolutional Neural Network was created using 4 convolutional blocks and then trained and evaluated. Through this, a model was generated for facial expression detection. And a frontal face detection model in OpenCV is first used to detect the faces in the frame and to each of these the above detected faces, the developed model was applied, and the predictions were obtained.

This system is applied to the video for frames at every 1s interval and the number of facial expressions of each emotion is recorded and displayed in the frame. Once the analysis is done to the entire video, the final frame with the final values is saved and can be obtained by the speaker.

Also, throughout the speech, a per-second timeline is created indicating emotion detected during each second of the speech which would give the speaker an overall idea of his/her facial expression usage. To make it easier to visualize the predominant emotion for 10 instances of these per-second predictions is identified and saved as well, which would give a clearer overview of the expressions used by the speaker at different intervals of the speech. These two reports generated are saved as text files.

C. Hand Gestures

This system identifies two universal hand gestures. The generally encouraged ‘Open Hands’ gesture with hands spread out and the clasping of one's hands which is a gesture that is better avoided in Public Speaking.

Pose Estimation is used to detect the wrist positions of the speaker and the wrist distance can then, in turn, be used to identify these gestures. One challenge was that the distance between the wrists of the speaker is obtained as pixel distances on the image and may significantly vary depending on the closeness of the speaker to the camera. To counter this, the distance between the shoulders of the speaker is also obtained and the detection is based on the relative distance between the shoulders and the wrists. Another problem in this approach was that when the speaker turned to a certain direction during the speech, the shoulder distance significantly drops causing erroneous identification of certain hand gestures. To avoid this, an average shoulder- distance measure was maintained, and the algorithm was applied only when the detected distance was above a factor of the average distance.

This was applied to the speech video for frames at every 2s interval, and each frame in which a gesture was detected was saved which could be reviewed later, and the final frame consisting

of the total number of these gestures is also saved.

This system was mainly applied on a video recorded by a single, stationary camera. A further enhanced version was also created to detect the hand gestures of a speaker on a stage recorded by multiple camera angles. This was applied to World Champion Speeches and the results could be viewed in the coming sections.

D. Analysing Stage Usage

To track the speaker the nose coordinate of the speaker is used. This is because it is a key point located at the central axis of the body of the speaker. The input video is analyzed at each 1-second interval and the location of this key point is tracked and recorded. The stage (frame) is divided into 5 sections, these sections representing five zones of the stage. The count of the frequency that the speaker is in each zone is calculated in real time and using this the percentage usage of each zone of the stage by the speaker is displayed.

Once the entire video analysis is done, the final frame with the stage usage for the entire speech is saved. Also, throughout the speech, a per-second timeline is created indicating in which zone the speaker was during each second of the speech which would give the speaker an overall idea of his/her stage movement. The report generated is saved as a text file.

E. Transitions

With regards to stage usage, there are mainly two primary mistakes made by a speaker. The first is standing stationary at one place during the entire speech. The other is moving across the stage continuously while speaking which would distract the audience from the message of the speaker. The key to effective stage usage is to use the stage with a purpose. One of the main ways to do this is to use transitions. For example, if the speech consisted of several stories, after one story the speaker will ‘transition’ from one place on the stage to another making it clear to the audience the shift from one part of the speech to another. This cannot be judged using the system as the system is unaware of the speech and can only analyze the body language of the speaker.

Therefore, what is done is the transitions of the speaker on the stage are identified and a short clip of this transition is cut, and all these transitions are saved and played back to the user, so that the user can identify whether a certain transition was appropriate. Transitions are identified by the movement of the key point coordinates in consecutive frames.

IV. EXPERIMENTS AND RESULTS.

A speech was prerecorded and analyzed using the system developed. Some of the results are shown below.

A. *Facial Expressions*

Fig. 2. Facial Expression analysis indicating the 7 emotions

```
1s : Neutral
2s : Sad
2s : Sad
4s : Neutral
4s : Happy
5s : Neutral
6s : Neutral
7s : Sad
8s : Neutral
9s : Neutral
10s : Neutral
```

Fig. 3. A snapshot of the timeline report of Facial Expression Analysis

```
Avg 1 : Neutral
Avg 2 : Neutral
Avg 3 : Neutral
Avg 4 : Neutral
Avg 5 : Neutral
Avg 6 : Neutral
Avg 7 : Neutral
Avg 8 : Neutral
Avg 9 : Neutral
```

Fig. 4. A snapshot of the average timeline for the entire speech

B. *Hand Gestures*

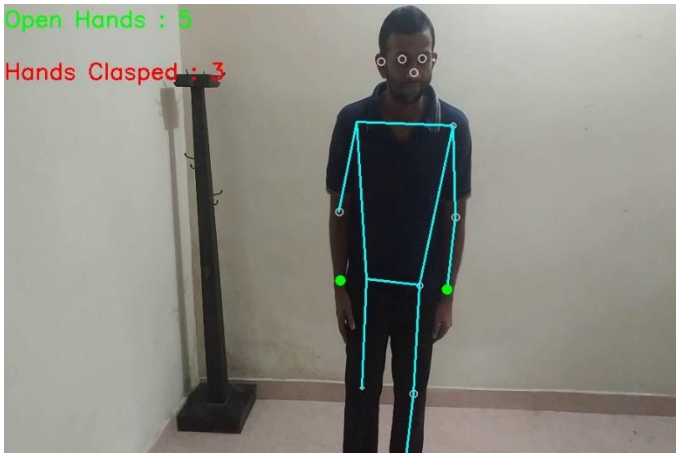

Fig. 5. Final Result indicating the frequency of hand gestures

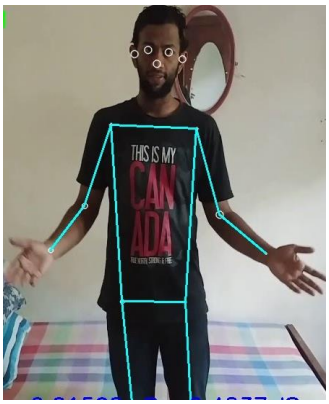

Fig. 6. An instance of an open hand gesture

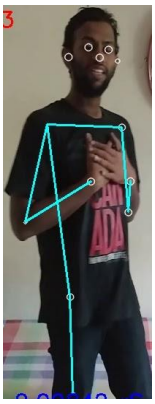

Fig. 7. Stage usage analysis with percentage stage usage. (Green markers indicate the movement of the speaker during the speech)

C. *Analysing Stage Usage*

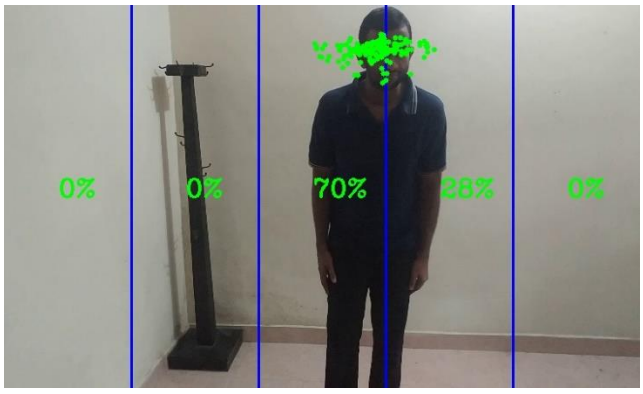

Fig. 8. A snapshot of the timeline report for stage analysis

```
0s : Zone 3
1s : Zone 3
2s : Zone 3
3s : Zone 4
4s : Zone 4
5s : Zone 3
6s : Zone 3
7s : Zone 4
8s : Zone 3
9s : Zone 3
10s : Zone 3
11s : Zone 4
```

Fig. 9. A snapshot of the timeline report for stage analysis

D. Transitions

A 14-second clip (7 before and 7 after) of the transition in the video is cut from the original video and all these transitions are concatenated into one video and saved to be viewed by the user.

V. Comparison

Every good speech is honed through multiple deliveries incorporating the feedback from previous speeches. As Darren Lacroix, the 2001 World Champion says “A great speech isn’t written. It’s re-written”. And Aaron Beverly, the 2019 World Champion describes the speech process as “Write, Read, Record, Listen. Re-Write, Re-Read, Re-Record, Re-Listen, Repeat.”

Thus, the outputs from different deliveries of the same speech can be compared by the user to improve.

Fig. 10. Hand gestures comparison between two deliveries of the same speech

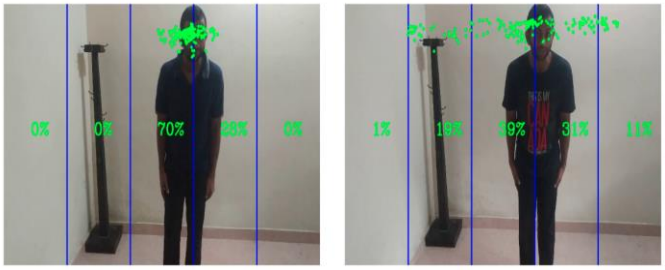

Fig. 11. Stage analysis comparison between two deliveries of the same speech

VI. Learning from World Champions

One of the best ways to develop any skill is to learn from someone who has already accomplished what you want. By applying these same algorithms to speeches of world class speakers, users can see how they use these body language components and can learn from the World Champions and use this to improve their body language skills. Below images show the application to these speeches.

A. Facial Expressions

Fig. 12. Facial Emotion Recognition [7]

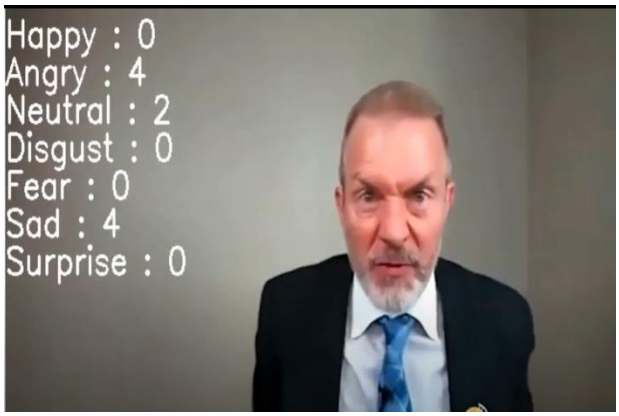

Fig. 13. Analysing the Facial Expressions of the World Champion of Public Speaking 2020 [7]

B. Hand Gestures

Fig. 14. Hand gestures analysis snapshot, 1st Runner up, 2020 [8]

C. Analysing Stage Usage

Fig. 15. Stage Usage analysis of the World Championship speech 2020 (Green markers indicate the movement of the speaker during the speech) [7]

D. Transitions

A 14-second clip (7 before and 7 after) of the transition in the video is cut from the original video and all these transitions are concatenated into one video and saved to be viewed by the user.

E. Hand Gestures – Stage Version

The above algorithms are applied to speeches recorded by a single, stationary camera. A further enhanced version of the Hand Gesture detection was created to detect hand gestures of a speaker on stage recorded using multiple camera angles.

Fig. 16. Hand gesture analysis snapshot, World Champion 2019 [9]

Fig. 17. Hand gestures analysis snapshot, World Champion 2014 [10]

Fig. 18. Hand gestures analysis snapshot, World Champion 2018 [11]

VII. Feedback from Toastmasters and Gaveliers

Since the target users of this system are anyone who wants to improve public speaking, such as Toastmasters, feedback for the system was obtained from a Toastmaster and a Gavelier. (Akkheel Mohammed, Competent Communicator, Level 9, Gavel Club of University of Sri Jayewardenepura and Yashmi Jayaweera, Toastmaster - SJMS Toastmasters Club, Former President - Gavel Club of University of Sri Jayewardenepura.)

One of the highlighted benefits of the system was that it was not subjective, as was the case with human evaluators. Another benefit highlighted was that the improvement between two speeches will not be marked to a human evaluator as the speech may be given after several weeks, but this system is able to handle this. Also, certain future improvements were suggested by them.

VIII. Limitations

- The video for the analysis of the stage analysis and transitions needs to be recorded from a single, stationary camera.
- A single person needs to be within the frame that is analysed.
- The accuracy of pose estimation is less when the entire body of the speaker is not within the frame

IX. Future Improvements

This is a suggestion based on the feedback from a Toastmaster. Each public speech consists of various stories. And these stories may convey different emotions. For example, it may be a humorous story or an emotional story. We could first get the emotion that the speaker is expecting the story to convey and by analysing the emotions of that speech we could output the feedback to the speaker confirming as to whether indeed the expected emotion was conveyed by the speech or story.

Also, we could provide quarterly reports on the analysis instead of at the end so that the speaker has an idea about which part of the stage was utilized more during the first part of the speech, the predominant emotions produced during the first part of the speech etc.

REFERENCES

[7] T. International, Effective Body Language Level 3 Project. 2016, p. 2.

[8] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", 2016.

[9] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1-1, 2019. Available: 10.1109/tpami.2019.2929257.

[10] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera and G. Anbarjafari, "Survey on Emotional Body Gesture Recognition", *IEEE Transactions on Affective Computing*, pp. 1-1, 2019. Available: 10.1109/taffc.2018.2874986.

[11] C. Chan, S. Ginosar, T. Zhou and A. A. Efros, "Everybody Dance Now", 2019.

[12] "Real-time Human Pose Estimation in the Browser with TensorFlow.js", *Blog.tensorflow.org*, 2021. [Online]. Available: https://blog.tensorflow.org/2018/05/real-time-human-pose-estimation-in.html. [Accessed: 21- May- 2021].

[13] "2020 Toastmasters World Champion of Public Speaking: Mike Carr", *Youtube.com*, 2020. [Online]. Available: https://www.youtube.com/watch?v=7Tev43VNRIc. [Accessed: 01- Sep- 2020].

[14] "Linda-Marie Miller: 2nd place winner, 2020 World Championship of Public Speaking", *Youtube.com*, 2020. [Online]. Available: https://www.youtube.com/watch?v=c1YNZ42JxN4. [Accessed: 01- Sep- 2020].

[15] "2019 Toastmasters World Champion of Public Speaking, Aaron Beverly", *Youtube.com*, 2019. [Online]. Available: https://www.youtube.com/watch?v=xmj1LBJu_Ss. [Accessed: 26- Aug- 2020].

[16] "Dananjaya Hettiarachchi - World Champion of Public Speaking 2014 - Full Speech", *Youtube.com*, 2014. [Online]. Available: https://www.youtube.com/watch?v=bbz2boNSeL0. [Accessed: 26- Dec- 2019].

[17] "2018 Toastmasters World Champion of Public Speaking, Ramona J. Smith", *Youtube.com*, 2018. [Online]. Available: https://www.youtube.com/watch?v=7Tev43VNRIc. [Accessed: 26- Sep- 2018].