INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING TECHNOLOGY AND APPLIED SCIENCES (IJIETAS)

REAL-TIME POSE ESTIMATION USING MEDIAPIPE FOR GESTURE AND SENTIMENT ANALYSIS

[1]Ishant Yadav, [2]Priyanshu Singh, [3]Manisha Kumari

[1]School of Engineering,Sharda University 2020501189, ishanrt@ug.sharda.ac.in
[2]School of Engineering,Sharda University, 2020551335, priyanshu@ug.sharda.ac.in
[3]School of Engineering,Sharda University 2021003255, manisha@ug.sharda.ac.in

ABSTRACT

Signs and Visual Learnings are considered as the easiest ways to learn and interact with your surroundings and people. These interpretations are performed by the changes in direction of hand landmarks, face landmarks, and body landmarks. Technological advancement in the field of computer vision, the possibility to predict these tasks will progress through the combination of image processing, deep learning, and machine learning techniques. In this research, we will learn how to leverage Mediapipe to estimate both facial and body landmarks. With the data we will then be able to build custom pose classification models that allow you to decode what a person might be saying with their body language with fine grain accuracy. While learning about the project, we can also customize the suits based on the needs. These estimations are low -dimensional based on skeleton poses. To predict the whole notion and description of the body and face in real time, we make a model using the real-time pipeline. We will then evaluate the routine of our project on a dataset preprocessed by us and show it achieve high accuracy in decoding body language.

Keywords: Body estimation, landmarks, Mediapipe, deep learning, computer vision, accuracy, pre-processing, facial expression.

DOI:

1 Introduction

Body language is an essential aspect of human communication and decoding it accurately can help understand a person’s emotional state and intentions. With the advent if Machine Learning and computer vision technologies , it is possible to automate the task of decoding body language using a combination of image processing, deep learning and machine learning techniques These face landmark recognition are already be available in market for facial detection whereas hand and pose estimation techniques can help in examining and estimation of various body movements and actions, such examples can of sign language and palm evaluation of different languages[1]. In some locations or areas there is a difference in languages, accents , dialects and many more , but some signs and normal head notations can help the person , further classification of implementing it in humanoid robot can help the artificial intelligences about the human characteristics and their behaviour, some other attentions can be using the conveying of sounds and noise of their surroundings which help the surrounding but basic difference is there can be disturbance in the background and change in amplitude and difficulty in conveying of their feelings in sound, thus using simple human visual behaviour can help them in the better understanding of how human sentiments actually behaves and works . Vision consists of many components, including coordination, memory retrieval, reasoning, estimation and many more. A machine with only one of this component cannot act as a systema as a whole thus should be able to at least summarises all the abilities. Human pose estimation can be determined as one of the most influential topics in the field of deep learning right now and around the tech world with connecting, it to the different mediums such as on the field of sports, work monitoring, home elderly care, home training entertainment, gesture control and many more. Human pose estimation can be divided into 2D and 3D based on the value of subjects, according to the variables recording the subjects and the views from the plane its capturing it[2].

Human Pose estimation can be described in single-stage methods and two-stage methods. The first method inputs the data captured into land points of different joints of the skeletal frame present in the camera which is of two classes: detection-based methods and regression-based MNE3 methods [3], the detection-based method signs the coordinates based on the heat map and predicts the likelihood of the joints whereas in regression based it predicts the landmarks of the coordinates using the angle between the joints and then the kinetic model of the structural frame, Media Pipe is presented in TensorFlow repository with the body tracking of 33 2d landmarks on the body with video frame of the dependencies based on the camera, in this project the RGB frame was selected, this model has the potential of selecting whole body frame but in this project, we would be using only the upper body frame which predicts the first 23 -25 landmarks. Media Pipe Face Mesh calculates about 476 3D junction landmarks, it fancies 3D surface geometry, needing only a solo frame point of objection thus not needing any depth module type [4][5].

For efficient results we would require an efficient subjects thus having a variety of results and better prediction it should be able to cover every aspect of the object, on large scale we would be using many frames but as it is small scale project we would be using a single frame of reference(laptop’s web cam) and having wide scale of variants of human pose for better dataset , and thus having large factors into many planer of x, y and z which further determines into many scaler planes. The video capture for holistic model was set to 0 but based on the cameras and the difference in various results it can be changed into 0.5 if possible and available [6].

There are various interfaces but we would be using OpenCV and then train the custom model using scikit learn with the help of various dependencies. An excel sheet stores all the data of the coordinate and helps in training the model.

The model which is presented in the project can further advance into many modulars such as face sentimental analysis, using into the humanoid robots for helping in elderly home and help in the field of sports for training and help in many models of basic physiotherapy help, yoga training and gym training

Background & Related Work

In this study, we compare various studies based on artificial pose estimation and then evaluate them on different scale thus helping us to better understand on why we use the media pipe holistic plane in our project.

There have been many works related to AI-based pose estimation in recent studies, such as:

1. **OpenPose:** Real time multi-person key point detection library for body, face, and hands estimation. This is used in variety of applications, where its further matured into various applications such as in sport analysis, fitness tracking and in the human-robot interaction.[7]
2. **Mask R-CNN:** It is a deep learning model that is used for instance segmentation and object detection tasks. It is an extension of the Faster R-CNN object detection model that adds a mask prediction branch to the existing Faster R-CNN framework.
3. **Mask R-CNN** considers its estate of art in a variety of instances and object detection benchmarks, such as COCO, PASCAL VOC, and cityscapes. These have already been used in range of applications, including robotics, autonomous vehicles, and augmented reality.[8]
4. **DeeperCut:** It is a multi-person pose estimation in images. It was introduced in research paper by MPI Informatik and University of Heidelberg in 2016, it is based on pervious state-of-the-art model, called “Poselets”, which uses a graphical model called a “poselet graph” to estimate the poses of individual people in an image. [9]
5. **PoseNet:** In collaboration with Google creative lab, it is released in the Tensorflow.js version of a Machine Learning Model which allows for a real-time human pose estimation in the browser.[10]

Mediapipe

It offers a wide range of computer vision solutions, including object detection, facial detection, hand tracking and many more. Some benefits of which why we use mediapipe were because of flexibility and its modular framework which supports a wide range of High accuracy: Mediapipe uses a combination of traditional computer vision and deep learning algorithms to achieve high accuracy in its predictions. It offers state-of-the-art models for various tasks and supports fine-tuning of these models for specific use cases [11]. Cross-platform compatibility: Mediapipe can run on a range of devices including CPUs, GPUs, and mobile devices, making it a versatile choice for developers who need to build applications that run on different hardware platforms. Multi-language support: Mediapipe supports multiple programming languages including C++, Python, and Java, which makes it easy for developers to integrate it into their existing software systems. Real-time performance: Mediapipe is optimized for real-time performance, which is particularly useful for applications that require fast and accurate computer vision processing, such as augmented reality, robotics, and self-driving cars [12].

In summary, Mediapipe offers a powerful and versatile framework for building various computer vision applications with high accuracy, cross-platform compatibility, multi-language support, and real-time performance. These benefits make it a valuable tool for developers who need to build computer vision applications
Pose estimation Pose detection for wide range of use cases [13].

Table 3.1 Difference between pose estimation and pose detection

Aspects	Pose estimation	Pose detection
Definition	Estimation of human poses in an image or video sequence	Detecting the presence of human body in an image or video sequence
Outputs	3D coordinates of key points, such as joints and landmarks	The output is a 2D bounding box around the pose
Complexity	It is more complex, which requires multi stages of pre-processing and machine learning technique	Less complex, often relying on simpler computer vision techniques
Input Requirements	High resolution images required in dataset	Low resolution images required in dataset
Examples:	Open Pose, MediaPipe pose, DensePose	OpenCV, Dlib, Haar Cascades

Methedology

1. **Dataset:** The dataset used in this study is loaded my marking the landmarks of the body coordinates and loading the dependencies for correct evaluation of the estimation of the dataset. The landmarks are first encoded with the clasification through RGB redlines and then marked according to the user . This is done to make the detections throgh OpenCV and load the coordinates to CSV so then we can learn about the model and then Use it when necessary . The dataset was ranging to [334 rows x 2005 columns]. The datahead for one emotion was enplaced to [855rows x 2005 columns].
2. **SciKit Learn:** Train the custom model using SciKit learn.It further evaluates down to:

(a). Read in Collected Data and Process:

To read in the collectd using scikit-lean , we needed to first import the neccesary module and read the CSV file using pandas. Once we have loaded the data into pandas DataFrame, we can then usi scikit-learn to preprocess and analyze the data .

To apply a learning model , first some preprocessing steps should be followed :

1. Spliting the data into training and test sets:

The Sklearn train_test_split function help us create our training data and test data. The training data and test data come from the same original dataset. To get the data to build a model , we start with a single dataset , and then we split into two dataset: train and test.[9]from sklearn.model_selection import train_test_splitX_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2).

3. Train Machine Learning Classification Model:

make_pipeline: In our dataset after splitting the train and test model , using the scikit-learn for classification tasks , we build a pipeline that data preprocessing steps anad the chosen model in a single object.It is a convenient function that allows us to chain multiple data preprocessing steps and a machine learning model intp a single pipeline.[14]

StandardScaler: is a preprocessing steps that sclaes the input the features to have a mean of 0 and variance of 1. This can be imported for many machine learning algorithims , such as linear regression and logostic regression , to ensure that all features are on a similar scale.

In our case we used :

a) **Logistic Regression**: It is a linear model for classification that uses a logistic function to predict the probability of each class. It is a widely approached algorithim in the learn based model , and is particulary used for binary classification models.[15]

b) **RidgeClassifier:** It is also a linear model for classification that uses a logistic function to predict the probability of each classs. It is a widely used algorithim in Machine Learning.[16]

c) **RandomForestClassifier:** An ensemble model of classification which is studied by constructing multiple tree and combing the prediction . Its maine advantages is that it can tackle the complex or branching relationships between features and target variables, and can be less prone to overfitting than individual decision tress. [17]

d) **Gradient Boosting Classifier:** On of the other ensemble learning model we used that evaluates multiple decision trees into a stronger decision trees. It trains the models in a sequential manner, where subsequent models tries to improve the mistakes of previous model. It is useful for handling imbalanced datset or dataset with complex interactions between the feartures[18]

e) **c)Evaluate and Serialize model:** We use pickle model from the numpy library where we use pickle.dump function which writes the model in binary format, the wb argu,emt specifies that the file is in write mode. We used pickle as it is convenientAfter algo fitting the model for accuracy metrics we found that accuracy_score of our models.

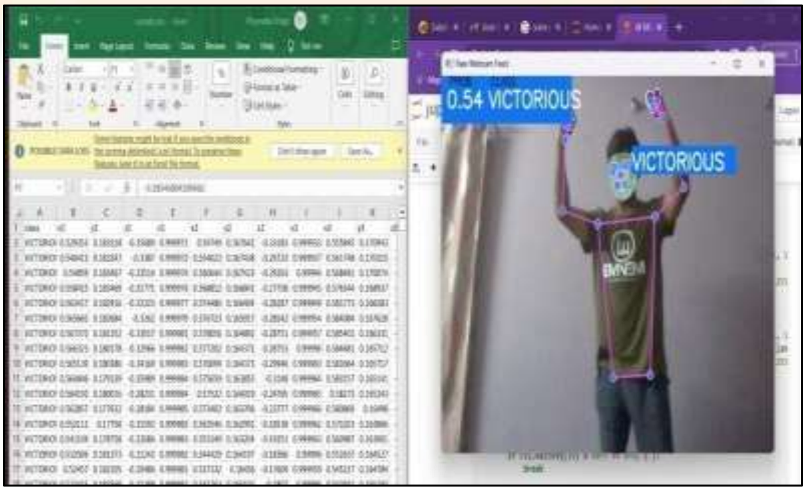
f) Logistic Regression = 1.0 Ridge Classifier = 1.0 Random Forest Classifier =1.0 Gradient Boost Classifier = 1.0

4. Make Detections with the model: The detection of the landmarks are done based on the setup coordinastes previously exporting the coordinates, the polarity of the different poses may differ based on the dataset and after evaluating the model and close redefing the variances of the model we will deploy it. The cv2After setting up the pipeline we will then deploy the holistic model after exporting the coordinates, the polarity of the different poses may differ based on the dataset and after evaluating the model and close redefing the variances of the model we will deploy it. Cv2.VideoCaputure would be different based on the camera module you are capturing .In our case, we would be setting it up to 0 , but in some cases the module would be different such as in linux device the video capture would be set up to 1. The showcase will shown by a display class with font of HERSHEY_SIMPLEX [19][20]. It would also be showing us the defined probability of the module of the different encdoing of the listed landmarks that were previously defined and were exported to the defined coords.csv file

.

Testing Of The Model

We tested variations of poses for radical poses to find the best results suitable for the environment present where the dataset was captured which counts as exposure of light to the model captured, plane of angle at which the camera was capturing the dataset and what was the variations of angles does the data was being captured for our model. We created an investigation model to check the probability of our application and what kind of variances are observed.



\

Fig 1. Detection results for victorious pose

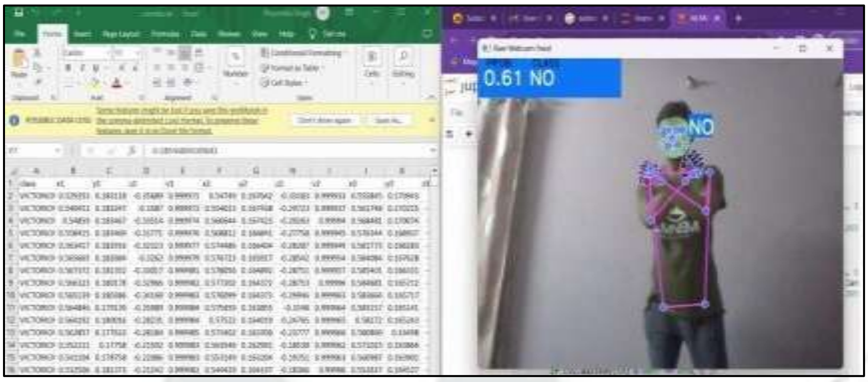


Fig 2. Detection results for No/Defense pose

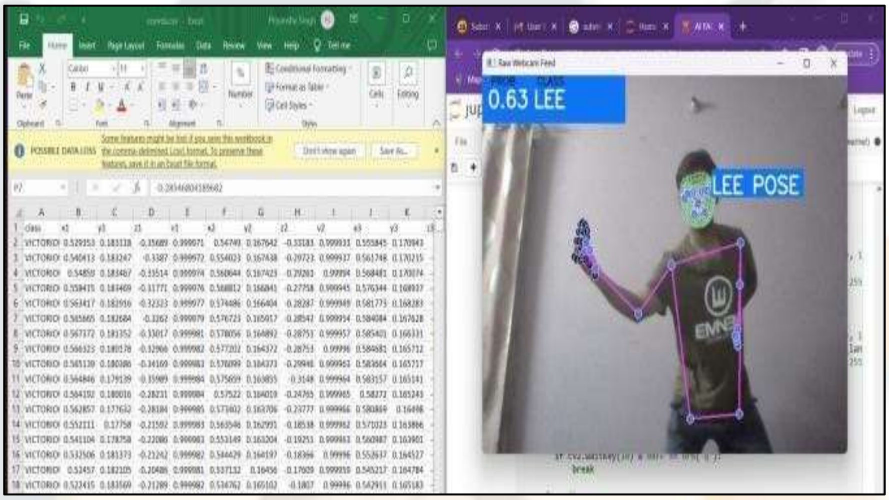


Fig 3. Detection results for LEE/Fighting pose

The above images show the difference in intentions between different poses and how the pipeline was able to detect the model and able to understand the after successfully applying it, the results were shown below:

Table 3.12 Evaluation result and predicting probability

S. No	Pose in Command	Predicted Pose	Resulted Pose	Evaluation Remarks
1.	Standing person with raise hands	Victorious	Victorius Prob:0.97	Good Result
.	Standing person with cross	NO	No Prob:0.98	Good Result
	Arms			

3.	Standing person with fighting stance	Fighti ng pose	Fighting Prob:0.87	Ok Result
4.	Face with opening mouth	Yawn	Yawn Prob:0.91	Good Result
5.	Standing pose with angry movement	Anger	Anger Prob 0.92	Good Result

Results And Discussion

The advancement of learning human behavior through the field of computer vison has been on large scope evidently for a huge period, with the major mode to understand a person emotion and how different elements of their gestures works in certain type of situations. In our model we were too able to analyze the certain movements of body by landmarking the various coordinates through the help of model and were able to give them identification of certain level. The probability of different postures was quite enraging with the certainty of positive results and can help in various fields of business. Various movements generations have already been captured before and deployed in the model such as sign language detection with varieties of signs that can be used for people with the problem of hearing ability , other language learning model consists of elements such as human gestures learning model while preparing for interviews as it can act as communication model with taking the client’s body gestures into consideration and giving them results and suggestions on how we can improve your posture and stance in different cases thus helping the candidate in various HR interviews. There are various models that have been released in the market such as PoseNet which is lightweight and can be deployed in the various android or other mobile devices. HRnet has shown better results in their field with better pyramidal results and can be followed up for research in their projected field. The application model was to be introduced in the AI humanoid model for better detection of its surroundings and how it can complete its tasks. Evaluation of emotions and object detection through the estimation of poses can help its recognition of better outcome of the task. Most preferably it is also used in advancement of learning of techniques in different fields of sports such as baseball player where a pitcher can able to correct its posture while throwing its various pitches such as there will be a different pose estimation and posture for a forkball than a slider from a fastball, same with estimation of yoga poses which on further learning can help in practice and can be modelled to be a yoga guru or ai fitness coach. There has also been research on hand-based biometric traits which were widely used in recognition systems and dynamic region of interest extraction was important for system enhancement. Which further evaluation of models the results are getting better such as the result in mobi-dev lab where they use the dataset of Human 3.6m and COCO model data to understand the dataset and give one of the best results. Also, with the introduction of new models such as Pose_reg where it uses the functationality of augmented application for better results although it has not been implemented yet but it can produce a gateway of vast opportunities to learn and understand the model for better application and business understanding

References

1. M. Murzamadieva, A. Ivashov, B. Omarov, B. Omarov, B. Kendzhayeva and R. Abdrakhmanov, "Development of a System for Ensuring Humidity in Sport Complexes," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 530-535.
2. N. Vasconcelos and A. Lippman, "Towards semantically meaningful feature spaces for the characterization of video content," Proceedings of International Conference on

Image Processing, Santa Barbara, CA, USA, 1997, pp. 25-28 vol.1

3. A. Datta, M. Shah and N. Da Vitoria Lobo, "Person-on-person violence detection in video data," 2002 International Conference on Pattern Recognition, Quebec City, QC, Canada,2002pp.433-438vol.1:10.1109/ICPR.2002.1044748
4. J. Chen, Y. Xu, C. Zhang, Z. Xu, X. Meng and J. Wang, "An Improved Two-stream 3D Convolutional Neural Network for Human Action Recognition," 2019 25th International Conference on Automation and Computing (ICAC), Lancaster, UK, 2019, pp.
5. Y. C. Shiu and S. Ahmad, "Calibration of wrist-mounted robotic sensors by solving homogeneous transform equations of the form AX=XB," in IEEE Transactions on Robotics and Automation, vol. 5, no. 1, pp. 16-29, Feb. 1989.
6. A. M. Zanchettin, N. M. Ceriani, P. Rocco, H. Ding and B.Matthias, "Safety in human-robot collaborative manufacturing environments: Metrics and control," in IEEE Transactions on Automation Science and Engineering, vol. 13, no. 2, pp. 882- 893, April 2016.
7. J. Ma, L. Ma, W. Ruan, H. Chen and J. Feng, "A Wushu Posture Recognition System Based on MediaPipe," 2022 2nd International Conference on Information Technology and Contemporary Sports (TCS), Guangzhou, China, 2022, pp. 10 -13, doi: 10.1109/TCS56119.2022.9918744
8. .[8] S. Adhikary, A. K. Talukdar and K. Kumar Sarma, "A Vision-based System for Recognition of Words used in Indian Sign Language Using MediaPipe," 2021 Sixth International Conference on Image Information Processing (ICIIP), Shimla, India, 2021, pp. 390-394, doi: 10.1109/ICIIP53038.2021.9702551.
9. K. Dixit and A. S. Jalal, "Automatic Indian Sign Language recognition system", 2013 3rd IEEE International Advance Computing Conference (IACC), pp. 883-887, 2013.
10. H. Muthu Mariappan and V. Gomathi, "Real-Time Recognition of Indian Sign Language", 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), pp. 1-6, 2019.
11. A. Chaikaew, K. Somkuan and T. Yuyen, "Thai Sign Language Recognition: an Application of Deep Neural Network", 2021 Joint International Conference on Digital Arts Media and Technology with ECTI Northern Section Conference on Electrical Electronics Computer and Telecommunication Engineering, pp. 128-131, 2021.
12. M. J. Cheok, Z. Omar and M. H. Jaward, "A review of hand gesture and sign language recognition techniques", International Journal of Machine Learning and Cybernetics, vol. 10, no. 1, 2019, [online] Available: https://doi.org/10.1007/s13042-017-0705-5.
13. B. Chakraborty, D. Sarma, M. Bhuyan and K. MacDorman, "A Review of Constraints on Vision-based Gesture Recognition for Human-Computer Interaction", IET Computer Vision, vol. 12, November 2017.
14. V. I. Pavlovic, R. Sharma and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: a review", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 677-695, July 1997.
15. J. Joy, K. Balakrishnan and M. Sreeraj, "SignQuiz: A Quiz Based Tool for Learning Fingerspelled Signs in Indian Sign Language Using ASLR", IEEE Access, vol. 7, pp. 28363-28371, 2019
16. [16] B. Unutmaz, A. C. Karaca and M. K. Güllü, "Turkish Sign Language Recognition Using Kinect Skeleton and Convolutional Neural Network", 2019 27th Signal Processing and Communications Applications Conference (SIU), pp. 1-4, 2019.
17. J. Browinee, "A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning", [online] Available: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/.
18. J. Jordan, "Evaluating a machine learning model", [online] Available: https://www.jeremyjordan.me/evaluating-a-machine-learning-model/.
19. H. V. Verma, E. Aggarwal and S. Chandra, "Gesture recognition using kinect for sign language translation", 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013), pp. 96-100, 2013.
20. X. Li, M. Zhang, J. Gu and Z. Zhang, "Fitness Action Counting Based on MediaPipe," 2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Beijing, China, 2022, pp. 1-7,

21. doi: 10.1109/CISP-BMEI56279.2022.9980337.
22. Cao Zhe, Hidalgo Martinez Gines, Simon Tomas, Wei Shih-En and Yaser A Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", IEEE transactions on pattern analysis and machine intelligence, pp. 2-6, 2019.
23. Cao Zhe, Hidalgo Martinez Gines, Simon Tomas, Wei Shih-En and Yaser A Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", IEEE transactions on pattern analysis and machine intelligence, pp. 2-6, 2019.
24. Valentin, Ivan Grishchenko, Karthik Raveendran, Tyler Lixuan Zhu, Fan Zhang and Matthias Grundmann, "BlazePose: On-device Real-time Body Pose tracking", ArXiv abs/2006.10204, 2020.
25. Yousef Alqasrawi, "Bridging the Gap between Local Semantic Concepts and Bag of Visual Words for Natural Scene Image Retrieval", International Journal of Sensors Wireless Communications and Control, pp. 1-2, 2016.