Human Body Pose Estimation and Applications

Amrutha K
Dept. of Computer Science
CHRIST (Deemed to be University)
Bengaluru, India
amrutha.k@res.christuniversity.in

Prabu P
Dept. of Computer Science
CHRIST (Deemed to be University)
Bengaluru, India
prabu.p@christuniversity.in

Joy Paulose
Dept. of Computer Science
CHRIST (Deemed to be University)
Bengaluru, India
joy.paulose@christuniversity.in

***Abstract*—Human Pose Estimation is one of the challenging yet broadly researched areas. Pose estimation is required in applications that include human activity detection, fall detection, motion capture in AR/VR, etc. Nevertheless, images and videos are required for every application that captures images using a standard RGB camera, without any external devices. This paper presents a real-time approach for sign language detection and recognition in videos using the Holistic pose estimation method of MediaPipe. This Holistic framework detects the movements of multiple modalities-facial expression, hand gesture and body pose, which is the best for the sign language recognition model. The experiment conducted includes five different signers, signing ten distinct words in a natural background. Two signs, "blank" and "sad," were best recognized by the model.**

Keywords—2D pose estimation, 3D pose estimation, body modelling, sign language recognition, MediaPipe, Holistic Pipeline estimation

I. INTRODUCTION

Numerous researches in the field of pose estimation from videos are available. However, there is ample scope for improvement in this field. Human body pose estimation is the method of understanding how a human appears in an image. The estimation is calculated by understanding where each body part, like joints, is located on the image. The subtle movements of the joints store a vast amount of information. The joints, also known as keypoints, include facial points, wrists, elbows and knees. Human body pose estimation differentiates a particular type of pose from a set of other poses. This estimation can be processed for a single individual or a group of individuals in a crowd. The challenge in crowd estimation is that multiple numbers of individuals may appear in varying poses. Identifying the right body pose at the same time for all individuals can be a tedious task.

Although highly complex systems that captured human movements existed, their accuracy in detecting human pose was comparatively minimal[1]. Active and passive sensing were widely used until the advent of computer vision. The individuals were attached with sensors to monitor their pose and activities, which restricted mobility greatly. Besides, each individual needs particular attention, even if their motions are related.

The input to the pose estimation model can be of different types based on the colour representation, dynamic or static representation and the number of coordinates considered. Every kind of input is processed in another way as the information obtained from them is different. A brief explanation of a few varieties of inputs is given below.

A. Colour-based input

- Red-Green-Blue (RGB) images are the type of images seen in daily life. The images captured by standard cameras are in this mode, and hence this is the most commonly seen image for pose estimation. Most of the models developed for pose estimation can directly take in the RGB image.
- Depth Images are different from RGB images as there is one more dimension involved called depth. Depth is the distance between the object and the camera. Microsoft Kinect is one of the devices that makes use of depth information and is measured using Time-of-Flight.

B. Static and dynamic input

- Static images are still images captured using a camera. Here, the body pose of the individual is fixed with respect to time; hence the estimation is comparatively simple. The required information is obtained from the spatial coordinates.
- Dynamic images or videos are a collection of static images. The body poses keep on changing according to both the spatial and temporal coordinates. Hence, different poses can be estimated from a complete video. For high-resolution videos, a collection of frames will have the same pose.

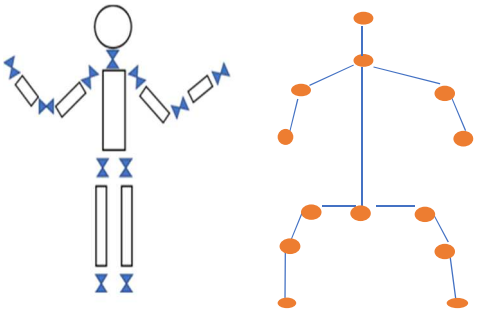
C. Coordinate based images (2D and 3D)

Two types of coordinate-based pose estimation approaches are available in computer vision- 2D pose estimation (X, Y) and 3D pose estimation (X, Y, Z).

2021 Innovations in Power and Advanced Computing Technologies (i-PACT) | 978-1-6654-2691-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/i-PACT52855.2021.9696513

- The functional approach in the two-dimensional (2D) pose estimation is representing the image in a pictorial form. The keypoints from the input are calculated based on the pixel values. The model can estimate the pixel values at each keypoint change for every individual movement and body pose. Fig.1. represents the 2D structure of the human pose. Each joint that has been marked gives a clear depiction. However, the pose model thus created has no dependency on the image, and hence the accuracy is not as high as expected. In [2], the authors used a mixture representation for the joints. This representation consists of complex joints in a deformable form with global and local templates. These templates can be used to match the parts of each image. A significant achievement of the proposed method is the efficient model articulation by sharing computation across similar warps; huge models are processed with the help of a collection of local templates. The global changes are reflected at the local levels as parts look different at different locations.

- The three-dimensional (3D) pose estimation model first calculates the 2D points and then converts them into 3D. The output required determines whether the process should be in 2D or 3D coordinates. In 3D pose estimation, the Z coordinate is usually the depth value present while capturing the image. Both static and dynamic input images can be processed using 3D pose estimation. The depth of the image is calculated without external devices like Leap Motion and Microsoft Kinect.



(a) (b)

Fig.1. (a) Planar representation of the Human Body Pose
(b) Kinematic representation of the Human Body Pose

II. Body Modelling

Before estimating the pose, every model adheres to a specific body model. Body modelling is one of the key aspects of pose estimation[3]. Choosing the body model helps in understanding the number of keypoints that are addressed. A human body is highly complex with few rigid and flexible characteristics like kinematic structure, body shape, surface texture, position, etc. The body models help to distinguish each part of the body separately for easy processing. The Human Pose Estimation models commonly use three body models: Skeleton-based, Contour-based, and volume-based models[3][4]. The following gives a brief description of the body modelling methods.

- Kinematic model
  The kinematic model, also known as the chain-based model, is a skeleton-based structure that connects the limbs and joints. The model forms a pictorial structure suggested by [5], where the rigid areas of the body are combined based on the spatial relationship[6].
- Planar model
  The Planar model gives more information about the shape and appearance of the human body. The use and importance of the model are presented in detail by the authors of [7]. The limbs of the model are planar patches.
- Volumetric/ volume-based model
  A volume-based model is a 3D-based pose estimation that uses human models in the form of volume-based structure[3][4]. In earlier pose estimations, the model used geometric shapes like cylinders, cones, etc., to understand 2D images better [4][8]. However, modern computation makes use of mesh structure.

III. Existing Architectures

There are a few publicly available architectures that are easy to use for custom pose detection. Some of them are explained in the following section.

- OpenPose
  OpenPose is one of the most commonly used multi-person pose estimation models [9]. This architecture is one of the best for detecting a person's hands, legs, body, and facial expressions. The API supports webcam feed and also surveillance camera feed on a real-time basis. It has an association with hardware architectures like CUDA, NVIDIA, etc. [10].
- High-Resolution Net (HRNet)
  HRNet is a neural network architecture used for estimating the human pose [10]. The main advantage of this model is that it can enhance the resolution of the actual video after estimating the pose. The authors [11] modified the vanilla network to deep HRNet, in which the model uses high-resolution video throughout the process. The authors also demonstrated the model's effectiveness using two benchmark datasets: the COCO keypoint detection dataset and MP II Human Pose Dataset.

- DeepCut

  DeepCut makes use of the bottom-up approach for pose estimation [10]. The DeepCut model can infer the number of persons in a scene, identify occluded parts, and disambiguate body parts of different people close to each other [12]. This model can process both videos and images. The model works in contrast to the other models that address the problem by first detecting the person and later identifying the pose.

- AlphaPose

  AlphaPose pose estimator is a top-down estimation model that can identify the poses of humans in inaccurate bounding boxes[10]. This model is the first accurate multi-person pose estimator with the open-source system that achieves 70+ mAP on the COCO dataset and 80+ mAP on the MP II dataset [13]. The PoseFlow model is used to track the pose of the same person in different frames [13].

IV. Related Work

The Human Pose Estimation has many applications in areas like sports, Human-Computer Interaction (HCI), Mixed Reality (MR), and others [14]. The input and output can be in the form of 2D or 3D images based on the context. The best way of estimation is to collect information from the key areas or the landmarks and ignore the rest.

As previously mentioned, 2D pose estimation obtains values from the (X, Y) coordinate. 2D Pose estimation has many applications in traffic safety, surveillance and other areas like human activity detection [15]. In 2D pose estimation, just a skeletal structure is visible. Even though the head movements were captured, complex features like facial expressions are not considered. A significant change in 2D pose estimation is that multi-pose estimation is possible[16].

3Dimensional Human Pose Estimation is gaining more popularity because there is no need for any external devices. However, much research is to be done to find models that can quickly and accurately convert 2D data to 3D. This section briefly reviews the various 3D human pose estimation applications that play a vital role in many fields. In [17], the authors propose an approach to estimate human pose from an RGBD image from which a robot can learn and imitate human activity. The robust keypoint detectors add the depth value.

In MR, hand pose estimation models are required as a solution for self-occlusion is yet to be found. As a solution, the authors of [18] have developed a system that accurately estimates the handpose in a 3D space. The authors designed a skeleton -difference loss function in a CNN network[18]. They have also modelled an object-manipulating loss function, which considers knowledge of the hand-object interaction to enhance performance.

In [19], the authors introduced MoVNect, a lightweight neural network that can capture 3D human poses using an RGB camera. The teacher-student approach used to train the neural network made it robust. The model has shown great performance in any platform like Windows and iOS without the need for servers.

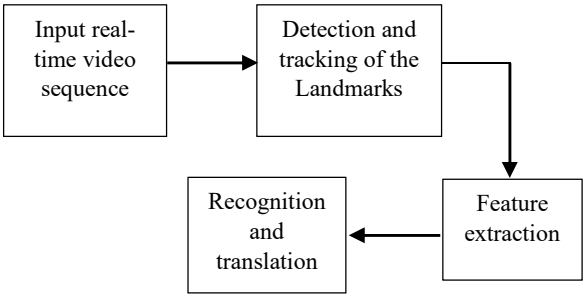


Fig 2. Steps involved in 3D pose based SLR system

In [20], the authors have proposed a multimodal framework for sign language recognition. Both Microsoft Kinect and Leap Motion sensors were used in their work. The sensors capture the features and feed them to HMM and BiLSTM – neural network classifiers from the raw data[20]. The dataset comprised 7500 Indian Sign Language (ISL) hand gestures that included single-handed and double-handed sign gestures. The overall accuracy of 97.85% was obtained using the model.

In [21], the authors have proposed a depth-sensor-based (Microsoft Kinect) SLR system. The invariant position framework recognized 2700 gestures with an accuracy of 83.77%. The dataset included occluded images, and the classification was performed using HMM classifier.

In [22], the authors introduce an SLR system developed with the help of human keypoints extracted from the face, hands, and other body parts. OpenPose, a freely available library, was used to detect and track the keypoints from the dataset. The features were then fed into an RNN classifier to obtain an accuracy of 89.5% for 100 sentences. The authors have also introduced a KETI dataset consisting of 10,480 sign language videos of high resolution and quality. The research utilizes the Human Pose Estimation system to recognize sign language from the given input video. The remaining part of the paper provides the details of the experiment that has been conducted.

V. Methodology

Sign language is one of the natural languages used by the hearing-impaired community of society. It is a complete language with its grammar and phonetics. Sign Language is conveyed with the help of all three modalities; facial expression, hand gestures and body pose. Hence, using just a simple computer vision approach might not be enough for capturing all the information. Another limitation of the computer vision approach is the inability to differentiate the ambiguity of the signs. Most of the time, the camera might not be able to catch slight movements of the fingers or the changes in the facial expression. This often leads to misinterpretation of the sign, especially in continuous sign language representation, which would be in a video format.

The optimal solution for such challenges would be to introduce 3D modelling of the input image. The machine learning architectures are so advanced that any type of

input can be captured and processed without any external support. During 2D processing, the input video undergoes various preprocessing stages, such as background removal and object segmentation. Once the signer gets separated from the video, the keypoints can be detected and tracked. The preprocessing step is wholly omitted in the case of 3D modelling. In this model, keypoints or landmarks, starting from the face, including all the features like nose, eyes, eyebrows, lips and teeth, extending up to the feet to understand various postures.

Fig.2. represents the basic steps in the 3D pose estimation-based SLR model. The video is captured using a normal video camera under varying backgrounds, not requiring any special devices. Once the landmarks are identified, the features can be learned, which finally helps the classifier identify the sign. A few advantages of 3D motion capturing include the invariance of hand movements and their shape. In a real-time scenario, the signer's hand movements cannot be restricted.

A. Human Keypoint Detection Using MediaPipe

The proposed model makes use of the existing framework called MediaPipe[13] [14]. This framework can be used to convert a perception pipeline into graphical components, including model inference, media processing algorithms and data transformations, among others [24]. Using Holistic MediaPipe live view of the facial landmarks, human pose, and hand tracking in real-time is possible in any device [23]. During the pipeline process, the values from all the three modalities are integrated at an optimum level. A pipeline can be defined as a directed graph of components where each component is a calculator[24].

In MediaPipe, all the calculations occur within the context of the graph [24]. A graph is a collection of nodes starting from the face up to the feet of a human. Each node in the graph is numbered and calculated separately. The numbering starts from the nose, which is assigned the value 0. Holistic MediaPipe is one of the best tracking models in an austere and cluttered environment. The experiment was conducted using both types of background, and the framework's performance in correlation with the backgrounds was extraordinary.

B. Feature Normalisation

The MediaPipe framework can extract values from all the 3D features and the fourth-dimension visibility available for the body pose model. The extracted features are then normalized using the Standard Scalar method.

$$\text{Feature Vector } \tilde{V} = [X1\ Y1\ Z1\ V1] \ldots\ldots [Xn\ YnZnVn] \quad (1)$$

For each sign, an array of feature vectors $\tilde{V}$ is collected. Depending on the landmarks' presence and absence, the values range between [0,1]. The features are then fed into the Random Forest classifier for translating the sign language into text. The train-test data were randomly split into a 70-30% ratio. The dataset collected was also tested with linear models like Logistic Regression and Ridge classifier, but the Random Forest model gave the best result with a training accuracy of 99%. Random Forest is an ensemble model of individual decision trees. Each tree makes its own decisions which can be right or wrong, and as a team, it takes the right decision.

VI. Experimental Results

Sign language can be represented either in an isolated word-by-word manner or as a sentence in which the words share a relationship. Here, for the experiment, only isolated signs are considered. The experiments were conducted for ten different most commonly used words with the help of five other signers under varying sign speed and natural background.

The predict probability function of the Random Forest classifier takes the vote from each tree of the model and returns the mean values ranging between the values 0 and 1. The class probability values of the class are calculated using the given formula.

$$\text{Probability(Class)} = \frac{\text{No. of votes for each class (from each tree)}}{\text{Total no. of trees in the forest}} \quad (2)$$

This function returns a set of predicted values and the maximum probability is taken into consideration and is stored in Table 1. When the model identifies each sign, the probability of the predicted class is displayed in the top left corner of the video. The predicted class of the recognized sign is displayed near the left ear of the signer. During each prediction, the model recalculates the changed sign along with the probability value. Fig.3. and Fig. 4. give an overall view of the predicted class along with the predicted probability.

Table 1 gives the maximum probability values of each class for different signers. The experiment showed that "Sad" and "Blank" are the two best detected signs. Whereas "Yawn" and "All the best" are poorly recognized signs. Fig.3. gives the graphical representation of the table where the percentage prediction of each sign is calculated. The dataset was collected using a built-in webcam. The class prediction with hand pose was much better when the elbows were visible. During testing, the signers maintained the exact distance between them and the camera.

TABLE I. Probability Values of Each Class for Different Signer

Class	**Class Prediction for Different Signers**				
	S1	***S2***	***S3***	***S4***	***S5***
Happy	0.81	0.75	0.85	0.91	0.81
Sad	0.95	0.98	1	0.99	0.95
Victory	0.85	0.9	0.87	0.75	0.81
All the best	0.75	0.78	0.74	0.8	0.78
Yawn	0.65	0.7	0.68	0.71	0.75
Hello	0.81	0.86	0.89	0.81	0.85
Good bye	0.85	0.89	0.89	0.9	0.88
Thank You	0.89	0.9	0.89	0.88	0.88
Sorry	0.89	0.85	0.9	0.88	0.88
Blank	0.98	0.99	1	0.99	1

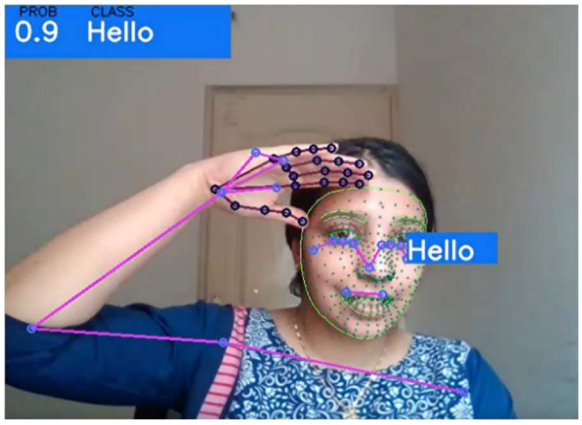


Fig. 3. Sign "Hello" recognized by MedidiaPipe and RF classifier with 0.9 class probability

Fig.4. Sign Language "sad" recognized using MediaPipe and RF classifier with 0.95 class probability

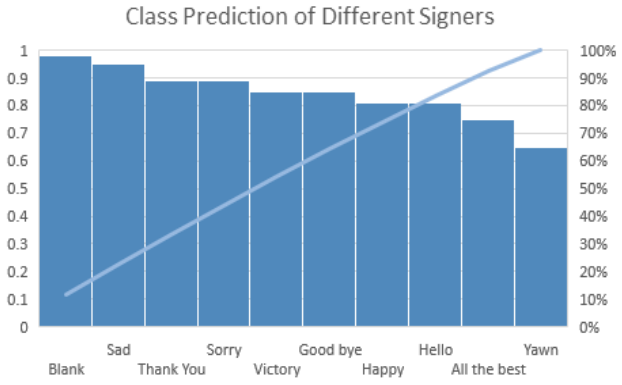


Fig. 5. Collective probability class of each category. Blanck and sad expression attained the highest value.

A. *Limitations of the model*

A few challenges were faced during the prediction. If the head angle is not proper, the model tends to misinterpret. An optimum distance must be maintained between the signer and the camera. If the signer is too close, certain keypoints are not recognised. Finally, when the hand pose was included; the signer should wait until the elbows are detected to get the maximum predicted value.

Conclusion and Future Work

Human Pose Estimation has several applications. This paper makes use of pose estimation for sign language recognition. The Holistic Human Pose Estimation model - MediaPipe was used as a keypoint detection and tracking system. The extracted landmarks are then fed into a Random Forest classifier to identify the different signs along with their probability values. The experiment was conducted with the help of five signers with a real-time dataset of ten isolated signs. The Random Forest Classifier predicted the class by calculating the maximum probability value. Gestures like "sad" and "blank" were the best-predicted classes. In contrast, classes like "yawn" and "all the best" were poorly recognised by the model. This paper mainly focused on recognising isolated sign language. The same model can be tested with a larger dataset with continuous sign language as future work.

REFERENCES

[1] T. B. Moeslund, and E. Granum, "A survey of computer vision-based human motion capture," Comput. Vis. Image Underst., vol. 81, no. 3, pp. 231–268, 2001, doi: 10.1006/cviu.2000.0897.

[2] Y. Yang, and D. Ramanan, "Articulated human detection with flexible mixtures-of-parts," IEEE transactions on pattern analysis and machine intelligence, vol. 35, pp.2878-2890, 2012

[3] C. Zheng et al., "Deep learning-based human pose estimation: A survey,"Deep Learning-Based Human Pose Estimation: A Survey. arXiv preprint arXiv:2012.13392, 2020.

[4] Y. Chen, Y. Tian, and M. He,"Monocular human pose estimation: A survey of deep learning-based methods," Comput. Vis. Image Underst., vol. 192, p. 102897, 2020, doi: 10.1016/j.cviu.2019.102897.

[5] S. Zuffi, O. Freifeld, and M. J. Black, "From pictorial structures to deformable structures."IEEE conference on computer vision and pattern recognition (pp. 3546-3553). IEEE, 2012

[6] M. A. Fischler, and R. A. Elschlager, "The representation and matching of pictorial structures" IEEE Transactions on computers,vol. 100, pp.67-92,1973.

[7] S. X. Ju, M. J. Black, and Y. Yacoob, "Cardboard people: a parameterized model of articulated image motion," in proceedings of the Second International Conference on Automatic Face and Gesture Recognition (pp. 38-44). IEEE, 1996.

[8] C. Sminchisescu, "3D human motion analysis in monocular video: Techniques and challenges," Springer, Dordrecht, pp. 185–211, 2008.

[9] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using part affinity fields," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[10] 'Pose Estimation: The ultimate overview in 2021 | viso.ai'. [Online]. Available: https://viso.ai/deep-learning/pose-estimation-ultimate-overview/. [Accessed: 30-May-2021].

[11] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5693-5703),2019

[12] L. Pishchulin et al., "DeepCut: joint subset partition and labeling for multi person pose estimation," in Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4929-4937) 2016.

[13] 'GitHub - MVIG-SJTU/AlphaPose: real-time and accurate full-body multi-person pose estimation&tracking system'. [Online]. Available: https://github.com/MVIG-SJTU/AlphaPose. [Accessed: 30-May-2021].

[14] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: video inference for human body pose and shape estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5253-5263),2020.

[15] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: people detection and articulated pose estimation, "in IEEE conference on computer vision and pattern recognition (pp. 1014-1021), 2009, IEEE

[16] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Real-time multi-person 2d pose estimation using part affinity fields," in Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7291-7299),2017.

[17] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3D human pose estimation in rgbd images for robotic task learning," in IEEE International Conference on Robotics and Automation (ICRA) (pp. 1986-1992), 2018, IEEE.

[18] M. Y. Wu, P. W. Ting, Y. H. Tang, E. Te Chou, and L. C. Fu, "Hand pose estimation in object-interaction based on deep learning for virtual reality applications," J. Vis. Commun. Image Represent., vol. 70, p. 102802, Jul. 2020, doi: 10.1016/j.jvcir.2020.102802.

[19] D.-H. Hwang, S. Kim, N. Monet, H. Koike, and S. Bae, "Lightweight 3d human pose estimation network training using teacher-student learning," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 479-488) 2020.

[20] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra, "A multimodal framework for sensor based sign language recognition," Neurocomputing, vol. 259, pp. 21–38, 2017, doi: 10.1016/j.neucom.2016.08.132.

[21] P. Kumar, R. Saini, P. P. Roy, D. P. Dogra, D. Prosad, and D. D. A. In, "A position and rotation invariant framework for sign language recognition (SLR) using Kinect, "Multimed Tools Appl, vol. 77, pp. 8823–8846, 2018, doi: 10.1007/s11042-017-4776-9.

[22] S.-K. Ko, J. G. Son, and H. Jung, "Sign language recognition with recurrent neural network using human keypoint detection', ACM Ref. format, 2018, doi: 10.1145/3264746.3264805.

[23] 'Holistic - mediapipe'. [Online]. Available: https://google.github.io/mediapipe/solutions/holistic. [Accessed: 29-May-2021].

[24] C. Lugaresi et al., "MediaPipe: a framework for building perception pipelines," Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172,2019.