International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 04 | Apr 2022 www.irjet.net p-ISSN: 2395-0072

A Survey on Human Pose Estimation

Vedank Pande[1], Anant Mokashi[2], Shantanu Patil[3] , Akshaj Singh[4], Nagesh Jadhav[5]

[1]*Student, Dept. of Computer Science and Engineering, MIT ADT University, India*
[2]*Student, Dept. of Computer Science and Engineering, MIT ADT University, India*
[3]*Student, Dept. of Computer Science and Engineering, MIT ADT University, India*
[4]*Student, Dept. of Computer Science and Engineering, MIT ADT University, India*
[5] *Asst. Professor, Dept. Computer Science and Engineering, MIT ADT University, India*

Abstract - *Human pose estimation (HPE) depicts the posture of an individual using semantic key points on the human body. In recent times, deep learning methods for HPE have dominated the traditional computer vision techniques which were extensively used in the past. HPE has a wide range of applications including virtual fitness trainers, surveillance, motion sensing gaming consoles (Xbox Kinect), action recognition, tracking and many more. This survey intends to fill in the gaps left by previous surveys as well as provide an update on recent developments in the field. An introduction to HPE is given first, followed by a brief overview of previous surveys. Later, we'll look into various classifications of HPE (single pose, multiple poses, 2D, 3D, top-down, bottom-up etc.) and datasets that are commonly used in this field. While both 2D and 3D HPE categories are mentioned in this survey, the main focus lies on pose estimation in 2D space. Moving on, various HPE approaches based on deep learning are presented, focusing largely on those optimised for inference on edge devices. Finally, we conclude with the challenges and obstacles faced in this field as well as some potential research opportunities.*

Key Words: **Pose estimation, pose estimation metrics, human pose modelling, pose datasets, 2D human pose estimation, 3D human pose estimation**

1. INTRODUCTION

Human pose estimation (HPE) attempts to locate spatial key points or landmarks on images containing humans. With applications in healthcare mentioned in [1] such as detecting early movement-based disorders, another application in sports where teams use pose estimation to monitor or analyse training footage to help improve sports players has recently come to light and many more are discussed in detail later in the survey.

HPE has been part of extensive computer vision [1] research for quite some time, and in the past decade has seen multiple deep learning approaches [2,3,4,5,6,7,8,9,10] attempting to provide more optimised and accurate solutions. For pose estimation to truly have an impact on human life, it needs to be easily deployed on edge devices and accessible to the common man. This is why, more recently, deep learning approaches [9,10] are being tailored and optimised to be deployed on edge devices.

HPE, like all domains of research, has its own challenges and hurdles that come with it. Many deep learning based methods require expensive hardware for training, thus increasing costs. Along with hardware expenses, gathering and maintaining human pose datasets is not a menial task. One of the biggest challenges for HPE today is optimising models to be deployed on edge devices, this supports real time applications.

HPE methods have various classifications based on the manner of the input images supplied and the architecture of the algorithm or model developed.

1.1 Previous Surveys and Contributions

Many surveys in the field of human pose estimation have been documented in the past, all contributing greatly in their own ways to recapitulate recent developments in the field. One notable survey, written by Zheng et al. [11], provides an intensive review of the most important topics in HPE at the time of publication. This review focused on explaining the taxonomy (classification) of HPE in-depth and also mentions important datasets and performance metrics used by researchers. Another review by Munea et al.[27] briefly introduces HPE nuances and majorly focuses on describing deep learning architectures designed to solve problems in this field. Although past surveys are very well

documented, we found certain gaps in surveys that we discuss in this survey. The majority of deep learning architectures designed for HPE are data-hungry and require expensive hardware for training. Such hardware is not available in edge devices that are used by consumers, this makes applying HPE to applications difficult as running heavy models on servers presents latency problems in applications deployed on edge devices. In this literature, we discuss recent developments and optimizations which allow lighter models to be developed. These models may be run on edge devices thus eliminating the problem of latency between a client and the servers. These advancements are crucial in making HPE accessible and deployable on edge devices.

2. HUMAN BODY MODELLING

The human body has a non-rigid structure with numerous variations in texture, poses and orientation. Human body modelling allows us to visualise the key points or landmarks extracted by HPE and thus, understand the pose of the individual or individuals in the frame. There are three major types based on the dimensionality used to model the human body, namely, skeleton-based model, planar model, and volumetric model

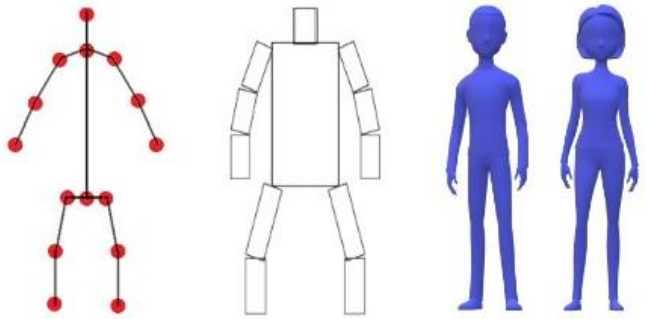

Fig -1: Types of human body modelling

2.1 Skeleton-based model

The skeleton model, commonly referred to as the kinematic model [11], visualises the pose by using the joint key points and the subsequently formed edges representing limbs in a one-dimensional space as shown in Fig.1(a). This computationally light approach is widely used for both 2D [1] and 3D HPE but falls short in scenarios where texture representation is required.

2.2 Planar model

The planar model, as opposed to the kinematic model, roughly provides structure and width information regarding the limbs of the body. Planar models use approximate rectangular shapes to represent the limbs as shown in Fig.1(b). An example is the cardboard model by Ju et al.[14], which showed promise in lower body tracking for activities such as walking and running.

2.3 Volume-based model

Volume-based models visualise in 3D space using geometric shapes or meshes. This approach provides rough estimations of the structure, width and texture of the human body. These models are useful for pose estimation in 3D space. Some examples of volumetric models [11] include the skinned multi-person linear model (SMPL) [12] and the tensor-based human modelling [13].

3. TAXONOMY OF HPE

In this section, we explore the various approaches and pipelines of HPE. Classifications of HPE are based on the spatial dimension the predicted key points and pose are in (2D/3D), the flow of the algorithm or method being used (top-down or bottom-up), the number of individuals, whose pose is to be approximated (single pose or multi-pose).

3.1 2D HPE

2D human pose estimation locates the landmarks of body joints on monocular images (images captured by single lens sensors). These joints are represented by a stick figure or skeleton model representation. Traditional HPE methods [1] on monocular images required handcrafting features and intense computer vision processing to determine the pose of the individual(s) in the frame. In this survey, we will focus on newer deep learning approaches to HPE.

3.2 3D HPE

3D HPE finds the pose of the individual(s) present in the frame in 3D space. It gives more information about the orientation and structure of the body compared to 2D HPE methods. This additional information is useful for the application of HPE in domains like 3D movies, animation, and even certain

military training areas. Although this field has had remarkable progress in recent years, 3D HPE has its own challenges, starting with the availability of datasets, while annotating joints in 2D images is trivial, doing so in the 3D space is not. While 3D pose estimation on monocular images has been successful, images taken from multiple perspectives (e.g. binocular camera images), are more suitable for the task as they provide a sense of depth in the image. Other sensors like LiDAR, make use of pulsed light waves, which bounce off of objects in the environment and are collected by the sensor. LiDAR sensors use the time taken for these light waves to return to estimate the positions of objects in the environment. This process is done millions of times, which ultimately produces an accurate 3D representation of the surroundings of the sensor.

3.3 Single/Multi Person HPE

This classification of HPE is based on the number of individuals who pose the model estimates. Single Person HPE detects poses for images that contain only one human in the frame. In frames where multiple bodies are present, the one in focus may be manually cropped out. Multi-person HPE as the name suggests is capable of localising the body joints of multiple individuals in the frame. While the latter type does require more compute power, it is useful since most real-world applications will consist of images with multiple humans in the frame. Most recent deep learning-based solutions are capable of performing both single and multiple person HPE.

3.4 Top-Down HPE

Top-down pose estimation consists of two main steps, human body detection and localization (using bounding boxes) and single person HPE. This method first finds all human bodies in the frame and then applies single person HPE to every detected individual. While this approach is simple, it is computationally expensive. This is because HPE is done once for everybody in the image, whereas the bottom-up approach is more efficient (discussed in 3.4). This approach also heavily relies on the human body detector; if it fails, joints will not be localised for those individuals.

3.5 Bottom-Up HPE

Bottom-up HPE as mentioned in section 3.3, is efficient, especially for multi-person pose estimation. Bottom-up HPE solutions first localise all joints in the entire frame and then proceed to associate and group body parts into their corresponding bodies. In certain situations, bottom-up approaches may fail to identify individuals by grouping the localised joints and limbs.

4. DATASETS

Deep learning-based HPE models are data-hungry and require correctly annotated, large datasets during training. Collecting suitable images and then properly annotating them is tasking and presents its own challenges. Datasets for HPE can be divided into those made for 2D HPE and those made for 3D HPE. Datasets made for 2D HPE generally consist of monocular images; certain 3D HPE methods that are capable of running on monocular images may use these datasets as well.

4.1 Datasets for 2D HPE

The MPII Human Pose Dataset gathered by Andriluka et al.[15] contains over 40,000 images of people taken from Youtube. The dataset has two hierarchies, the first level classifies the images in a broader space with groups like, “sports”, “Lawn and garden”, “Home”, etc. While the second hierarchy level narrows the classification down to groups like, “rock climbing” and “picking fruit”. Adriluka et. al. proposed a new evaluation metric which is a slight variation of the percentage of correct key-points (PCK) metric called PCK. PCK measures the accuracy of the localised key points and uses a threshold which is a fraction of the size of the bounding box on the body. PCKh changes this threshold to 50% of the size of the head segment length. Bulat et al. [7] in 2020 achieved a 94.10% score on the PCKh metric using the MPII dataset. This dataset is used widely as a benchmark for evaluating HPE models.

FLIC Dataset

The Frames Labelled in Cinema (FLIC) dataset composed by Ben Sapp and Ben Taskar [16], contains 5003 frames from 30 Hollywood movies and contains annotations of 10 upper body joints. A human body detector was used on every 10th frame

of a movie to extract images. Frames in which humans were detected with high confidence were sent to Amazon Mechanical Turk (MTurk) for annotation. MTurk is a platform to crowdsource tasks to individuals or organisations online. For evaluation, Sapp et al. use a metric similar to PCK explained in [16]**.** While this dataset does not quite have the volume that the MPII dataset does, it still remains a popular dataset used for benchmarking.

COCO Dataset

The COCO dataset [17] contains almost 330,000 images made for object detection. While the original COCO dataset is not specifically made for HPE, COCO Keypoints 2016 and COCO Keypoints 2017 are datasets for HPE originating from the original COCO dataset. The difference between the two lies in the train, test and validation split used.

4.2 Datasets for 3D HPE

HumanEva Dataset

The HumanEva dataset [18] by Sigal et al. contains 7 video sequences that provide 3D body poses of the subjects being recorded. It has two versions, namely, HumanEva-1 and HumanEva-2. The difference between the two is the number of subjects participating in recording, and the actions performed by these subjects.

HumanEva-1 used 4 subjects performing 6 actions in 3 repetitions each. The actions done were walking, jogging, gesturing, throwing and catching a ball, boxing, and combos. The dataset contains 74,600 frames extracted from videos taken by 7 synchronised video cameras. Motion capture (MoCap) systems from ViconPeak and video capture cameras from IO Industries and Spica Tech make up the hardware that was used in this version of HumanEva.

HumanEva-2 used 2 subjects (both who were part of HumanEva-1), performing a sequence of actions. Starting by walking along an elliptical path, then proceeding to jog around the same path and finally having the subject balance on each of his/her two feet in the centre of the frame. Training and validation data between the two versions of HumanEva remain the same, while HumanEva-2 contains a lower number of test frames (2,460 as opposed to 24,000 in HumanEva-1).

Human3.6M

The Human3.6M [19] dataset contains 3.6 million images of 3D human poses. The recording took place with 5 females and 6 males and the experimental setting consisted of 4 video cameras, 1 time-of-flight sensor, and 10 motion cameras. Reflective markers were attached to every subject's body which let the motion capture (MoCap) system track them. The dataset provides bounding boxes around humans in the frame, useful for top-down pose estimation methods.

4. PERFORMANCE EVALUATION METRICS

To determine how accurate and usable an HPE model is in real-world scenarios, various metrics are used by researchers. Different performance metrics are used for 2D and 3D HPE. The next sections describe some of these metrics.

5.1 2D Evaluation Metrics

Percentage of Correct Parts (PCP) [20] measures the percentage of localised body parts that match the ground truth to a certain extent. A localised body part is said to be correct if the segment endpoints lie within a fraction of the length of the truth segment. This fraction is also known as the PCP threshold and is varied for different tests. Decreasing the threshold leads to stricter criteria and decreases the chances of a predicted part being labelled as correct.

Percentage of Detected Joints (PDJ) [4] is similar to PCP but addresses a drawback of PCP where it penalises shorter limbs such as lower arms (since the length of the segment is small, the PCP threshold results in a fine margin for correct part evaluation). PDJ works around this by using the diameter of the torso to calculate the threshold for correct evaluation. This way, it overcomes the problem of overly strict evaluation of shorter limbs. A joint is said to be correctly predicted if it falls under the

length of a fraction of the torso diameter. This fraction can be varied for finer precision.

Percentage of Correct Key-Points (PCK) proposed by Yang et al. in [21], states that their evaluation metric overcomes a drawback of PCP where the metric does not penalise false-positive errors, giving an unfair advantage to models which localise a large number of joint keypoints. PCK accepts a keypoint as correct if the predicted keypoint is within a distance of α.max(h,w). Here, h and w are the height and width of the bounding box surrounding a keypoint. α is a threshold value that determines the strictness of evaluation, a lower value of α requires more precision in keypoint prediction compared to a higher value.

5.2 3D Evaluation Metrics

Mean Per Joint Position Error (MPJPE) [11] is an evaluation metric used for 3D HPE which uses the mean Euclidean distance between the predicted joint keypoint and the truth value for the joint.

3D Percentage of Correct Keypoints (3DPCK) is an adaptation of the original PCK metric [21] to a 3D space. It labels a localised joint as correct if the distance between it and the truth value joint is less than a certain threshold.

6. MAJOR APPROACHES TO HPE

In this section, we discuss various solutions discovered by researchers in this field. For each model, we discuss important topics like the model architecture, the dataset used, the performance metric applied and its results, etc.

6.1 Convolutional Pose Machines

Convolutional pose machines (CPM) proposed by Wei et al. [3], use a sequence of iterative convolutional network processing which finally outputs a belief map for each joint keypoint to be detected. These belief maps tell us the probability of each pixel in the image being part of that joint and can later be visualised using heatmaps. The prolonged use of convolutional networks may potentially cause the problem of vanishing gradients [22], the authors solved this problem by adding supervision in intermediate layers of the network.

The algorithm consists of two stages, stage one consists of only the first iteration of the algorithm. Here, only the input image is used and a classifier outputs a belief map for the image. The second stage, which deals with all iterations after the first (>=2), uses the input image passed through a convolutional feature extractor and the belief map from the previous layer.

The authors tested their model on benchmarks such as FLIC [16], LSP, and MPII [15].

Using the PCKh-0.5 metric on the MPII dataset, Wei et al. achieved a score of 87.95%. The ankle, which is the most challenging joint to detect in HPE according to [3], achieved a score of 78.28%, which was 10.76% higher than the closest competitor at the time of publication (recently, more advanced models have been brought about which have better scores on similar tests).

The model also achieved 84.32% on the Leeds Sports Pose (LSP) dataset and 97.59% and 95.03% score on the elbows and wrists respectively, of the FLIC dataset using the PCK metric.

6.2 Stacked Hourglass Networks for HPE

The stacked hourglass model proposed by Newell et al. [8], is based on the hourglass architecture shown in Fig.2 The hourglass design allows networks to capture information at different scales of the image. A smaller scale is required for identifying features such as faces and limbs, whereas spatial information such as orientation and the arrangement of the limbs requires a larger scale of the image.

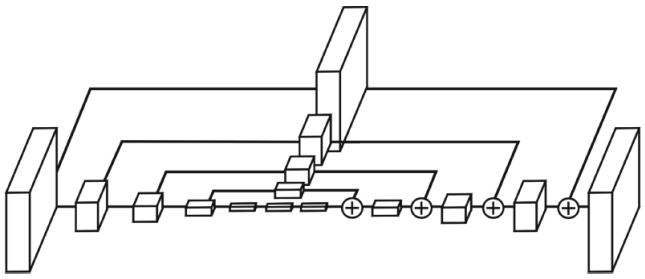

Fig -2: Representation of a stacked hourglass model

The hourglass architecture consists of a sequence of downsampling and upsampling operations along with the heavy use of residual networks. The design uses convolutions along with residual connections to pass on information to later layers, and max pooling to down sample the feature to a low resolution. After reaching the lowest resolution in the design, the nearest neighbour method for upsampling. Since both spatial (orientation and arrangement of limbs) and lower level features (limbs) are both important for HPE, skip connections are used to preserve spatial information from earlier layers in the latter layers of the network.

In the final model, these hourglass networks are stacked sequentially, this allows outputs of previous hourglass networks to be evaluated and processed again to achieve better results. Heatmaps are generated after every hourglass and passed onto the next hourglass network. This allows for intermediate supervision, to alleviate the network from the problem of vanishing gradients.

The stacked hourglass model was evaluated on the FLIC and MPII datasets using the PCKh@0.5 and PCK@0.2 metrics, respectively. Newell et al. achieved 99% PCK accuracy on the elbow and 97% accuracy on the wrist. On the MPII dataset, scores of 91.2% and 87.1% on elbows and wrists, respectively, using the PCKh@0.5 metric.

6.3 HPE via Soft-Gated Skip Connections

Bulat et al. in this literature question the use of residual connections in most state of the art deep learning approaches to HPE [8]. The authors proposed a new design that helps better both the accuracy and the efficiency of HPE models.

Firstly, Bulat et al. propose the use of gated skip connections, with learnable parameters which help control the flow of data across the network. These parameters learn how much information from previous stages should be passed onto the next.

The authors also use a hybrid structure which is a combination of the Hourglass [8] and U-Net [23] wherein they introduce skip connections between the encoder and decoder sections of the U-net architecture.

Evaluation of the model was done on the MPII dataset based on the PCKh metric, achieving an overall score of 94.1%. The authors used the PCK based metrics on the Leeds Sports Pose (LSP) dataset and achieved an overall score of 91.1%. The surprising part of these results is that, despite being a shallower (uses only 4 stacks as opposed to the SOTA 8 stacks) and computationally lighter model, it outperforms most heavy models trained on large datasets.

6.4 Movenet

Movenet, developed by researchers at Google, is tailored for applications that require low latency HPE. Movenet provides this by being an extremely lightweight and accurate model for pose estimation. Movenet is divided into two variants, namely, lightning and thunder. The lightning variant is optimised to decrease latency (from inference), i.e. runs faster than its alternative. The thunder variant is made for use cases requiring more accuracy.

Moving to the architecture of the model, movenet is a single person, bottom-up pose estimation model. The model is based on the MobileNetV2 [10] and CenterNet [24] architectures. It consists of two major components, the feature extractor (based on MobileNetV2) and the prediction heads (based on CenterNet). The feature extractor is attached to a

feature pyramid network [25], followed by the prediction heads. The four prediction heads give the following outputs: A person centre heatmap, keypoint regression field, person keypoint heatmap and a per keypoint offset field.

The model has been trained on the COCO [17] and an internal Google dataset called Active. Evaluation metrics used include keypoint mean average precision (mAP) and the inference time taken for a single image. This quick inference and low latency make Movenet perfect for applications such as real-time motion tracking in fitness use cases.

6.5 Posenet

Posenet developed and researched by Kendall et al. [6], is a pose estimation model which heavily makes use of convolutional neural networks and is a result of transfer learning using GoogleLeNet [26] as a base model.

The authors altered the original model by replacing all 3 softmax layers with regressors to output a 7-dimensional vector. Before the regressors, a fully connected layer of size 2048 is inserted to obtain a feature vector to be used in the latter stages.

An in-house dataset was used consisting of images from 5 scenes named Cambridge Landmarks from training and testing.

6.6 BlazePose

Blazepose, [9] developed by Google researchers, is a new lightweight model built for inference on edge devices similar to Movement. The model consists of 2 main sections, the detector and the estimator. The detector finds the human in the frame and returns a cropped portion of the original image to the estimator. The estimator then outputs the 33 localised keypoints.

The estimator uses a heatmap for training, but not during inference to reduce latency (time taken for inference). BlazePose is based on an encoder-decoder architecture to obtain heatmaps for the body joints. The architecture starts with a heatmap and offset maps; these two are used only in training and removed during inference. The model also uses skip connections to send information of high-level features from the shallow layers of the network to the latter layers.

BlazePose achieved a score of 97.2% on a re-annotated version of the AR dataset using the PCK@0.2 metric. Just like Movenet, BlazePose is built for deployment and inference on edge devices that do not have the hardware capabilities to run larger models. This on-device inference leads to lower latency and makes the model suitable for fitness-related applications where real-time analysis is required.

7 APPLICATIONS

7.1 Activity recognition

Activity recognition tracks the body for a certain amount of time to detect the action or activity being performed by the individual.

Some use cases where activity recognition can be applied include monitoring of ill or old age patients. In the event of the patient falling over, the model can alert the appropriate people about the incident.

Activity recognition can also be used in fitness applications such as workouts and dancing. Monitoring the pose of individuals in the frame allows for analysis and correction of posture in activities such as exercises and dance, where improper practice can lead to injuries.

7.2 HPE in CGI and Animation

Animation and computer graphics use HPE to pre-determine the pose of the actor on whom the animations are to be applied. Instead of manually applying the graphics to the actor, using a motion tracking bodysuit, the movement and posture of the

individual can be tracked and graphics will be fitted accordingly.

7.3 Motion tracking for gaming consoles

Most notably, the Xbox Kinect console has a collection of games that require the player to move in real-time, using a motion sensor provided by Microsoft. The user is required to perform actions such as swinging a bat, jumping, etc. These actions are detected by pose estimation models in the Kinect module, allowing for more interactive games that involve physical exercise.

8 FUTURE RESEARCHES

Most state of the art (SOTA) models and solutions to HPE are data-hungry and require expensive hardware to perform training and inference. This limits the application of HPE to servers and cloud computing machines that possess the appropriate hardware for the task. Extending the previous point, more recently, models optimised to be run on edge devices (hardware such as phones, laptops and wearable devices which can not perform heavy computation). Some models mentioned in this survey [9], are examples of development in this area of HPE. To allow the full potential of HPE to be unlocked and applied in real-world applications, further optimization and research can be conducted in the coming years.

8. CONCLUSIONS

In this survey, we provided a review of 2D and 3D HPE and the various other classifications that exist in current solutions such as single/multiple poses and top/bottom-up approaches. Popular datasets used as benchmarks are described, and later on, models and their performance on these benchmarks are mentioned. The most recent research area in this field pertains to models being optimised for deployment on smaller edge devices such as mobiles and wearable gear. Despite a large number of researchers working in this domain, developers still face many challenges in this field. We hope that future research helps humanity use the full potential of HPE in critical applications.

REFERENCES

[1] Z. Cao, G. Hidalgo, T. Simon, S. -E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172-186, 1 Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[2] Su, Zhihui, Ming Ye, Guohui Zhang, Lei Dai and Jianda Sheng. "Cascade Feature Aggregation for Human Pose Estimation." *arXiv: Computer Vision and Pattern Recognition* (2019): n. Pag.

[3] S. Wei, V. Ramakrishna, T. Kanade and Y. Sheikh, "Convolutional Pose Machines," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4724-4732, doi: 10.1109/CVPR.2016.511.

[4] A. Toshev and C. Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1653-1660, doi: 10.1109/CVPR.2014.214.

[5] Li, Wenbo & Wang, Zhicheng & Yin, Binyi & Peng, Qixiang & Du, Yuming & Xiao, Tianzi & Yu, Gang & Lu, Hongtao & Wei, Yichen & Su, Jian. (2019). Rethinking on Multi-Stage Networks for Human Pose Estimation.

[6] Kendall, Alex & Grimes, Matthew & Cipolla, Roberto. (2015). PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. 2938-2946. 10.1109/ICCV.2015.336.

[7] Bulat, Adrian & Kossaifi, Jean & Tzimiropoulos, Georgios & Pantic, Maja. (2020). Toward fast and accurate human pose estimation via soft-gated skip connections. 10.1109/FG47880.2020.00014.

[8] Newell, Alejandro & Yang, Kaiyu & Deng, Jia. (2016). Stacked Hourglass Networks for Human Pose Estimation. 9912. 483-499. 10.1007/978-3-319-46484-8_29.

[9] Bazarevsky, Valentin & Grishchenko, Ivan & Raveendran, Karthik & Zhu, Tyler & Zhang, Fan & Grundmann, Matthias. (2020). BlazePose: On-device Real-time Body Pose tracking.

[10] Sandler, Mark & Howard, Andrew & Zhu, Menglong & Zhmoginov, Andrey & Chen, Liang-Chieh. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. 4510-4520. 10.1109/CVPR.2018.00474.

[11] Zheng, Ce, Wenhan Wu, Taojiannan Yang, Sijie Zhu, Chen Chen, Ruixu Liu, Ju Shen, Nasser Kehtarnavaz,

and Mubarak Shah. "Deep learning-based human pose estimation: A survey." *arXiv preprint arXiv:2012.13392* (2020).

[12] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black,
"SMPL: A skinned multi-person linear model," ACM TOG, 2015.

[13] Y. Chen, Z. Liu and Z. Zhang, "Tensor-Based Human Body Modeling," 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 105-112, doi: 10.1109/CVPR.2013.21.

[14] Ju, Shanon & Black, Michael & Yacoob, Yaser. (1997). Cardboard People: A Parameterized Model of Articulated Image Motion. Proceedings of the International Conference on Automatic Face and Gesture Recognition.

[15] M. Andriluka, L. Pishchulin, P. Gehler and B. Schiele, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3686-3693, doi: 10.1109/CVPR.2014.471.

[16] B. Sapp and B. Taskar, "MODEC: Multimodal Decomposable Models for Human Pose Estimation," 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3674-3681, doi: 10.1109/CVPR.2013.471.

[17] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In *European conference on computer vision*, pp. 740-755. Springer, Cham, 2014.

[18] Sigal, Leonid & Balan, Alexandru & Black, Michael. (2010). HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. International Journal of Computer Vision. 87. 4-27. 10.1007/s11263-009-0273-6.

[19] C. Ionescu, D. Papava, V. Olaru and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 7, pp. 1325-1339, July 2014, doi: 10.1109/TPAMI.2013.248.

[20] Eichner, M. & Marín-Jiménez, Manuel & Zisserman, A. & Ferrari, V.. (2012). 2D Articulated Human Pose Estimation and Retrieval in (Almost) Unconstrained Still Images. International Journal of Computer Vision. 99. 10.1007/s11263-012-0524-9.

[21] Y. Yang and D. Ramanan, "Articulated Human Detection with Flexible Mixtures of Parts," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2878-2890, Dec. 2013, doi: 10.1109/TPAMI.2012.261.

[22] Hochreiter, Sepp. (1998). The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 6. 107-116. 10.1142/S0218488598000094.

[23] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234-241. Springer, Cham, 2015.

[24] Duan, Kaiwen, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. "Centernet: Keypoint triplets for object detection." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6569-6578. 2019.

[25] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936-944, doi: 10.1109/CVPR.2017.106.

[26] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[27] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang and C. Yang, "The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation," in IEEE Access, vol. 8, pp. 133330-133348, 2020, doi: 10.1109/ACCESS.2020.3010248.

[28] Yang, Wei, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. "Learning feature pyramids for human pose estimation." In *proceedings of the IEEE international conference on computer vision*, pp. 1281-1290. 2017.

[29] Artacho, Bruno, and Andreas Savakis. "Unipose: Unified human pose estimation in single images and videos." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7035-7044. 2020.

[30] Yang, Sen, Zhibin Quan, Mu Nie, and Wankou Yang. "Transpose: Keypoint localization via transformer." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11802-11812. 2021.

[31] Bulat, Adrian, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. "Toward fast and accurate human pose estimation via soft-gated skip connections." In

2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 8-15. IEEE, 2020.

[32] Tompson, Jonathan, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. "Efficient object localization using convolutional networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 648-656. 2015.

BIOGRAPHIES

Vedank Pande

Computer Science and Engineering student, MIT ADT University, Pune

Anant Mokashi

Computer Science and Engineering student, MIT ADT University, Pune, India

Shantanu Patil

Computer Science and Engineering student, MIT ADT University, Pune, India

Akshaj Singh

Computer Science and Engineering student, MIT ADT University, Pune, India

Nagesh Jadhav

Asst. Professor, MIT ADT University, Pune, India