Study on Deep Learning Models for Human Pose Estimation and its Real Time Application

Conference Paper · March 2023

DOI: 10.1109/ISCON57294.2023.10112004

2 authors, including:

Kanojia Sindhuben Babulal
Central University of Jharkhand
42 PUBLICATIONS **121** CITATIONS

SEE PROFILE

2023 6th International Conference on Information Systems and Computer Networks (ISCON)
GLA University, Mathura, India. Mar 3-4, 2023

Study on Deep Learning Models for Human Pose Estimation and its Real Time Application

Jyoti Jangade*
Dept. of Computer Science and Engineering
Central University of Jharkhand
Ranchi, India
sindhukanojia@gmail.com

Kanojia Sindhuben Babulal
Dept. of Computer Science and Engineering
Central University of Jharkhand
Ranchi, India

2023 6th International Conference on Information Systems and Computer Networks (ISCON) | 979-8-3503-4696-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCON57294.2023.10112004

***Abstract*-In computer vision, human pose estimation details the posture of the person's body structure that can be Kinematic, Planer, and Volumetric in an image or video. However, pose detection is often critical to be driven by distinct human actions. Thus, this survey report analysis the recent progression of the bottom-up and top-down human pose evaluation models. This survey report focuses on 2D and 3D skeleton-based human pose detection from the captured Red Green Blue(RGB) images. We have condensed the performance of the recent pose recognition, tracking, and detection techniques that utilize pose estimation from colour images as captured and then exhibit room for much more refinement in this domain. In this paper, scrutinize the study of human pose estimation models like 2d and 3d HPE for identify human movements such as running, dancing, sport so on and recent computer vision-based advances. This study has included various methods for detecting in two and three dimensions. This paper summarises the deep learning models for HPE, dataset, and challenges.**

***Index Terms*—Human pose estimation, Two Dimensional HPE, Three Dimensional HPE, Human Body Models.**

I. INTRODUCTION

The process of estimating the body parts of humans from input data in photos and videos can be done in 3D or 2D. It is a dominant study in the computer vision domain. For example, [1] sports activity recognition requires a highly accurate estimation of skeletal joints. There is no doubt that human pose estimation is a crucial and demanding task in computer vision. It is possible to recognize, classify or detect actions from human poses using pose estimation techniques; however, both tasks are unrelated. Future advances in human pose estimation by action recognition technology may be a suitable illustration methodology. We must know the position of the body joints to create a pose grading system so that the system can accurately assess the postures. Several techniques for estimating human poses have become a prerequisite in recent years. The estimation of human poses is not an essential case for computer vision. Still, it performs a vital role in various real-time applications such as Healthcare, Monitoring videos, man-machine interaction etc. these applications are further discussed. In this paper, we have discussed two main sections: 2D-3D human pose estimation and reviewed recently published research related to these fields.

II. HUMAN POSE ESTIMATION

Human pose estimation is a technique to identify the joints in the human body. This process is currently being used extensively in computer vision, and it transforms from two-dimensional (2D) to three-dimensional (3D) Human pose estimation. Fig.1 shows the division of HPE with their subdivisions.

A. 2D Human Pose Estimation

Calculating the position of a person's joints from visuals such as photos or videos is known as 2D human posture [2]. This technique is divided into single-person and multi-person pose estimation [10].

1) Single-person human pose estimation:

A single-person poses estimation method defines the human pose skeleton on the located keypoints that work on an individual. It is referred to as the computer vision job of drawing a skeleton through keypoints of a human. [4].

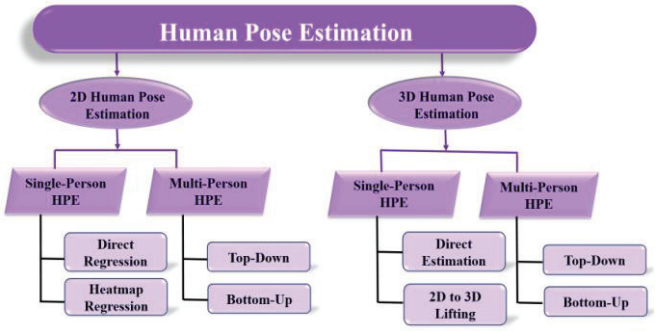

Fig. 1. Human Pose Estimation.

- Direct Regression :

This method directly maps the joints of the body or the features of human body models. If the model is used to predict 17 key points for a specific person, the result would be a 17/2 vector that contains the X and Y coordinates of every expected landmark shown below in Fig.2.

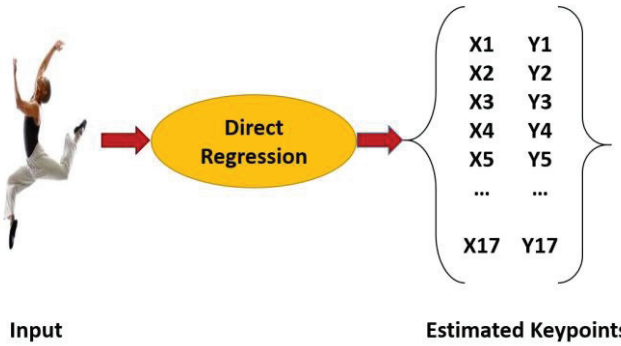

Fig. 2. Direct Regression

- Heatmap Regression:

Heatmap regression is extensively used for 2D Human Pose Estimation and grouping keypoints location: [27] for example - hands, faces, and bodies. In the heatmap framework, the pixel values are typically used as probabilities of the relevant pixel being mapped to the landmark within this framework. Putting this technique into practice is simple, and it could achieve pixel-level precision. Thus, heatmap regression is Human Pose Estimation's primary technique. It could be effective with a top-down process that resizes everyone to the same size. But in bottomup techniques, when people are of different scales, it is preferable to change the standard deviation for every landmark by the scale of the relevant people shown in Fig.3.

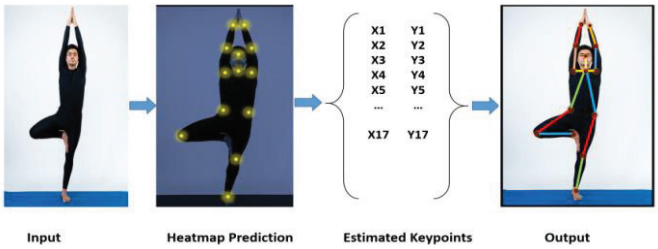

Fig. 3. Heatmap regression

2) Multi-person human pose estimation:

Multi person human pose estimation ware designed for drawing the landmarks of multiple person. The goal of 2D Multi-person Pose Estimation (MPPE) is to find and pinpoint the keypoints for every person visible in a input image.

B. 3D Human Pose Estimation

Three-dimensional(3D) Human Pose Estimation\cite{b35} is the process of identifying the 3D keypoints areas of a human body from an image or video as shown in \textbf{Fig.4}. It computes the 3D posture of a RGB image using x, y, and z coordinates.

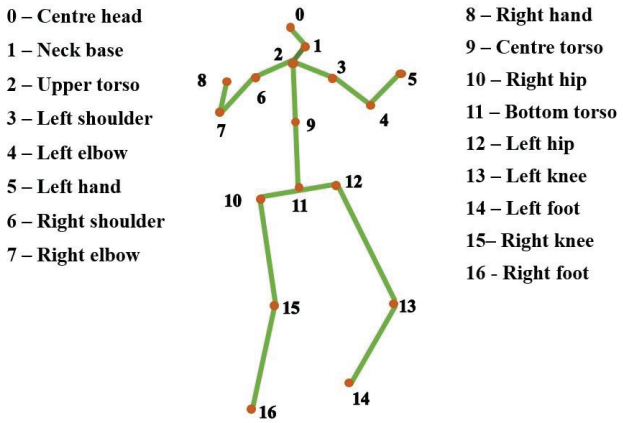

Fig. 4. 3D Human Pose Estimation

1) Single person HPE:

Generally, studies that estimate a human posture for a single person utilise just one image or video.

- Direct estimation approach: These technique directly map on images to draw 3D body joint position.
- Two dimensional to three dimensional lifting approach: Two dimensional to three dimensional lifting procedures that reckoner three dimensional human posture from approximated two dimensional human poses have grown in popularity as three dimensional HPE techniques. To estimate two dimensional posture in the first step, commercial 2D HPE techniques are used. The second step is used to lift two-dimensional to three-dimensional to create the three-dimensional position.

2) Multi Person HPE:

Multi-person processes are better suited for practical applications. Since the human pose contains plentiful structural and locomotion data, MPPE has gained interest in various applications, including animation, the comprehension of human activity, human-robot interaction, and more. Multi-person pose estimations are categorised into two schemes:

- Top-Down approach: This technique of three-dimensional multi-person pose estimation (3D MPPE) is first to carry out person detection to detect an individual. It is possible to collect keypoints using the top-down technique by selecting human subjects which may be subjected to a posture assessor through the use of a module. The topdown process comprises a human applicant indicator, a single-person posture assessor, and a human posture tracker, in Fig.5(a) show the process of this approach.
- Bottom-up approach: By computing key points, the bottom-up method recognizes many persons simultaneously and combines them with PAF (Part Affinity Field) and other techniques, in {Fig.5(b)} show the process.

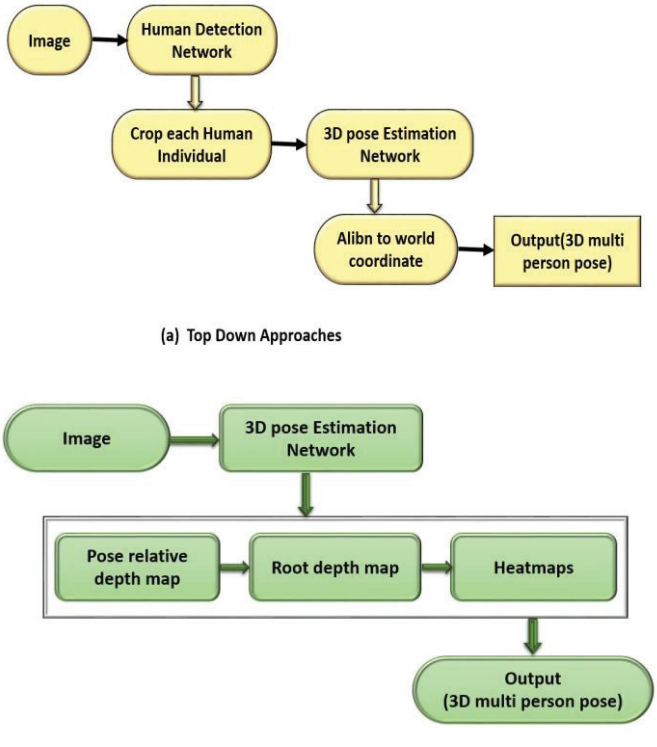

Fig. 5. Multi person HPE a) Top-Down Approach b) Bottom-Up Approach

III. HUMAN BODY MODELING

The human body model is beneficial for drawing keypoints leading to the accurate extraction of features. There are three kinds of Human body models: Kinematic, Planar, and Volumetric, showed in Fig.6.

Kinematic : This model is based on a skeleton where it draws the keypoints on a human's body posture by using pose estimation methods. This body model includes several keypoints, and all these keypoints are associated with sticks to match the structure of the human body, which helps to understand the position of the human body.These models are

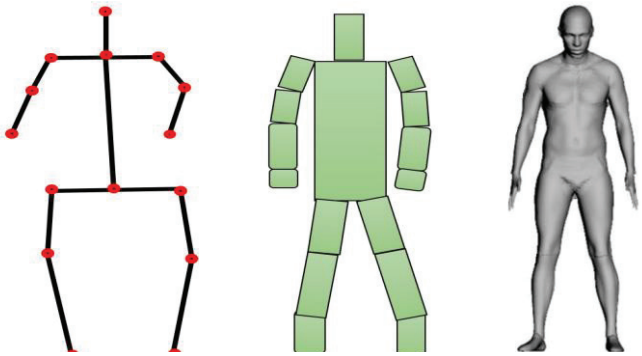

Fig. 6. Human Body Modeling: a) Kinematic b) Planer c) Volumetric

used to find the angle for detecting correct human activities. This model is frequently employed to estimate human poses Fig.6 (a) shows the model. Kinematic body model work on both 2D and 3D depictions based on a skeleton-based model. Planar : As seen in Fig.6 (b), the human body's shape and appearance are depicted using a planar model (b). Body components are typically depicted in the planar model as rectangles that roughly approximate the contours of the human body. The whole structure of the human body is bifurcated into chunks of rectangular body parts, as depicted in fig 6(b). Volumetric : With growing interest in 3D person rebuild, numerous volumetric person body structures have been developed for various formation of human body (one example representation is shown in Fig.6 (c)). A popular model in 3D HPE is the SMPL (Skinned Multi-Person Linear) which may be modelled with organic pose-dependent deformations displaying soft-tissue dynamics.

IV. DATASET

Dataset is set of raw data, collected for the research study it can be images, videos, text, etc. For testing human posture estimation algorithms, a lot of datasets have been made available recently.

A. 2D Human Pose Estimation Dataset

The aim of two-dimensional pose estimation is to locate keypoints in images or videos. Pose Estimation is used to infer human poses. We summarise 2D datasets in Table I.

1) MPII:

A dataset for pose estimation was developed using Max Planck Institute for Informatics for the first time. These are images from YouTube videos. In terms of the number of images, the MPII dataset consists of about 25k images, of which about three-quarters are available for training. In addition, there are about 40 thousand people who have been annotated. This document highlights the possibilities for the MPII to be used for both analyses of single- person pose estimations and multi-person pose estimations [1]. We provide a summary of some popular datasets for action recognition.

2) MSCOCO:

Over 200,000 images and over 250,000 instances of people are included in the COCO dataset, which is categorized by 17 keypoints. The COCO datasets for pose estimation have the following two versions: MSCOCO landmarks-2016 and MSCOCO landmarks-2017 [5].

3) CROWDPOSE:

Estimating the pose of a multi-person is a very challenging task because the images contain a large crowd, and it is complicated to find the pose of each person in that group. CrowdPose is the dataset [7] commonly used for various person pose prediction tasks. It comprises 20,000 images chosen from 30,000 another images on the grounds of crowd index.

4) POSETRACK:

Posetrack [6] is a brand-new, widely used benchmark for tracking and estimating human poses from videos. This dataset is highlighted in the following task:

- Estimating multi-person poses in One-frame.
- Estimation of multi-user poses by using videos.
- Tracking poses of multi-person.

TABLE I. TWO DIMENSIONAL HUMAN POSE ESTIMATION DATASET

SI	Year	Dataset	Train Dataset	Val Dataset	Test Datase t	Single/Multi Person	Keypoints
1	2014 [3]	MPII (Images) 3800	-	11,701	Single Person		16
2	2017 [5]	AI Challenger 21000030000 (Images)			30000	Multi person	14
3	2017 [6]	PoseTrack 2017(Videos)	20	20	20	Multi person	15
4	2018 [6]	PoseTrack 2017(Videos)	292	50	208	Multi person	15
5	2019 [8]	MSCOCO (Images)	57000	5000	20000	Multi Person	17
6	2020 [7]	CrowdPose (Images)	10,000	2,000	8,000	Multi person	14
7	2021 [9]	MSCOCO (Images)	57k	5k	20k	Multi Person	17

B. 3D Human Pose Estimation Dataset

Three-dimension pose estimation is based on computer vision technology that detects, tracks and analyses the person's posture. The neural network video detects the three dimensional coordinates of the number of joints. We have summarized the 3D HPE datasets in Table II.

TABLE II. THREE DIMENSIONAL(3D) HUMAN POSE ESTIMATION DATASET

SI	Year	Dataset Type	size	keypoints	Properties
1	2016 [14]	Videos	13k frames	15	There are RGB camera and 10 sensors used
2	2017 [15]	Image	1,892,176	26	There have 5 persons 5 action IMU data and inside environment
3	2017 [16]	Image	1.3M	15	There have indoor and out door environment
4	2018 [17]	videos	60,51k frames	18	3D human pose is taken 10 video clips.
5	2018 [18]	video	500k frames	14	There are mostly 10M poses
6	2020 [24]	video	2k frames	20	Contain 90 subjects and set of 20 action performed

V. DEEP LEARNING MODELS IN HPE

Deep Neural Networks (DNN) are excellent for predicting pose estimation of the individual pose but in case of multi person faces some problems: An image contains a number of people with different poses. Calculation complexity will be increasing and also time taking in the real world. We have summarized the deep learning models in Table 3.

A. OpenPose

OpenPose is a multi-individual human pose detection method in the real world. It has once demonstrated its ability to find landmarks on the user body, foot, hand, and face in a picture. Total 135 keypoints [25] can be detected by OpenPose. A picture is first passed over a CNN method to extract the feature maps from the input. The model employs the first 10 layers of the VGG-19 network. Using multi-stage CNN, the Part Confidence Map (PCM) and Part Affinity Fields are produced using feature maps. Forecasts from each branch are evaluated throughout time in phases. Part confidence maps are used to generate bipartite graphs between pairs of pieces. In a multi-phase Convolutional Neural Network (CNN), the Part Affinity Field clarify Lt in the first set of phases using the base network F extracted features.

$$L^t = \varphi^t(F, L^{t-1}), \forall 2 \le t \le T_p,$$

The confidence map recognition is refine in the following stage using the PAF output from the earlier layers.

$$S^{T_p} = \rho^t(F, L^{T_p}), \forall t = T_p,$$

$$S^t = \rho^t(F, L^T{}_p, S^{t-1}), \forall T_p \le t \le T_p + T_c,$$

The final S (CM) and L are then analysed using the greedy technique (PAF). (1)

B. DeepCut

This technique follows the bottom-up method [?] for estimate multiple-person pose. DeepCut's goal is to tackle the challenge of multi-person human pose estimation by modelling the poses of each person in an image together.

TABLE III. DEEP LEARNING MODELS FOR HPE

	OpenPose	Mask-RNN	AlphaPose	DeepCut	Iterative Error Feedback
Detection Part	Body, Foot, Hand, Face	Body, Face	Hand, Face, Body	Body	Body, Face
Keypoint	137 [25]	17 [11]	136 [12]		16 [13]
Operation	Real-time	Real-Time	Real-time	Real-time	Real-time
Approach	Bottom-Up	Bottom-Up	Top-Down	Bottom-Up	Top-Down
Single/Multi Person	Single, Multi Person	Multi Person	Multi Person	Multi Person	Single, Multi Person

C. AlphaPose

AlphaPose is a actual-time multi-person posture estimation used to resolve the estimation of multi-person posture problem. It is the 1st open-source system. There have following components [12].

- Symmetric spatial transformer network (SSTN).
- Parametric pose non-maximum-suppression (PNMS).
- Pose-guided proposals generator (PGPG).

It possesses the B-Box suggestions made by a person detectoion. As stated, the authors employed a VGG-based SSD512 detector to find persons in their research. After that, the Symmetric Spatial Transformer Network (STN) and SPPE module receives these bounding box recommendations. The posture suggestions are generated at this stage. There might be a lot of duplicate detections in the identified postures. The authors reduced these by removing the unnecessary postures using a parametric Pose-NMS.

D. Mask R-CNN

Mask-Regions with Convolutional Neural Networks is a DNN developed to resolve the instance segmentation [19][28] issue in machine learning or computer vision. It can specifically tell between different items in a picture or video [11]. The feature maps are used by a region proposal network (RPN) to detect B-Box options for the presence of objects. Since that the B-Box persons might be distinct sizes. A Region of Interest (ROI) align is applied to reduce the dimension of the extracted features, because they all have a similar size. After ROI process the features are extracted and send to the CNN for the final segmentation mask and bounding box prediction.

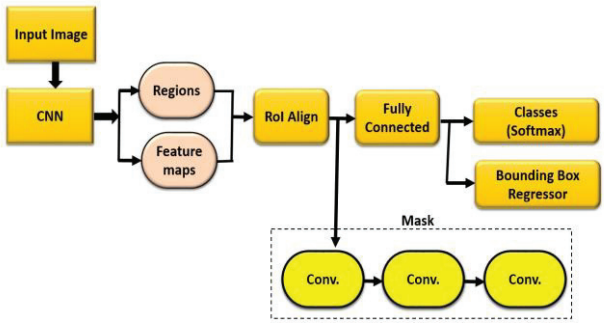

Fig. 7. Mask RCNN

E. Iterative Error Feedback

To collate two shapes using learning-based methods is challenging because accuracy for one-step prediction is difficult. This kind of issue is solved using the iterative error feedback method. Iterative Error Feedback (IEF) [13] was utilized to adjust the skeleton of the human pose to match the structure in 2D images.

VI. APPLICATION

Human posture estimation has been used in the following areas:

1. Human Pose Recognition and Detection: Various applications, including pose identification, prediction, and detection, have used pose details as indicated. Suggest a pose-based system for real-time pose recognition. Used the pose's skeleton style to recognise actions. A posture recognition system that can identify the pose the person is holding and then retrieve training materials for that pose from the net to call attention to it [23].
2. Yoga Posture Correction: Yoga posture correction for self-training is responsible for posture correction. Pose adjustment and action direction are typically handled in the presence of instructors. With the aid of three dimensional Pose Estimation and posture detection, Artificial Intelligence (AI) individual instructors can facilitate training by putting up a webcam without personal instructors present.
3. Healthcare: Different applications are there for guidelines of the diet and healthcare to do yoga, and it is also helpful for remedy of some gruesome diseases. So, yoga is the best way to physical challenges [1].
4. Monitoring on videos: Footmen are monitored through video in certain situations to track and monitor their position and movements. Human Pose Estimation technology was first employed in the field of video monitoring.
5. Man-Machine Interaction: Rapid human-computer connection systems [22] have been developed to estimate a person's poses. These technologies can appropriately analyse instruction by collecting positions of the human body. Researchers have been working in this area for several decades.
6. Sport: It uses human pose estimation methods to track players' key positions and movements in real time. Furthermore, the pose estimation can employ the instruction activities of their actions.
7. Social Distance: Social distance has proven a successful strategy for lowering the risk of viral infections; for example covid-19 [20]. Human pose estimation techniques [21] are used for detecting distance in a video.

VII. RESULT

Why do we need human pose estimation to recognition human activity?

The main factor is that in videos dataset sometimes it make difficult to recognise activity of human. Extracting poses from their context makes it possible to prevent misleading context completely. The skeleton based techniques are useful for human's real time action recognition. Human pose estimation perform better role for action recognition, with practical work like interaction between human and robots, social distance management so on.

VIII. CONCLUSION AND FUTURE WORK

This research work finding compares approximately 27 research manuscripts based of pose estimated deep learning technology. Following are some of the crisp outcomes of this manuscript:

- Discusses about different techniques used for the location of points of human body from image.

- Define variety of applications of Human Pose Estimation during action recognition.
- A comprehensive study of different classical as well as Deep Learning machanism for the estimation of 2D/3D single pose & multi pose method using HPE are reviewed.
- This analysis concludes that the algorithm of HPE changes according to its various environment.

The future work on Human Pose Estimation is bright and includes a wide range of applications that are crucial to daily living. High-dimensional datasets, which is advance than 2D and 3D, can also be effectively analyzed.

Inspite the extraordinary achievements in deep learning techniques, still research should be carried out to uproot the loophole regarding Crowd and occlusion dataset.

REFERENCES

[1] L. Song, Gang yu, Junsong Yuan, Zicheng Liu, ”Human pose estimation and its application to action recognition: A survey,” Journal of Visual Communication and Image Representation, Vol. 76, pp. 103-055, 2021.

[2] Q. Dang,J. Yin, B. Wang, W. Zheng,”Deep learning based 2d human pose estimation: A. survey,” Tsinghua Science and Technology, Vol. 24, pp. 663-676, 2019.

[3] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, ”2d human pose estimation: New benchmark and state of the art analysis,” Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pp. 3686-3693, 2014.

[4] Feng Zhang, Xiatian Zhu, Chen Wang, ”Single Person Pose Estimation: A Survey,” arXiv preprint arXiv, pp. 2109.10056, 2021.

[5] J. Wu, H. Zheng, B. Zhao, Yixin Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, others, ”Large-scale datasets for going deeper in image understanding,” 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 1480-1485, 2019.

[6] Umar Iqbal, Anton Milan Juergen Gall, ”Posetrack: Joint multi-person pose estimation and tracking,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2011-2020, 2017.

[7] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, L. Zhang, ”Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation,” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5386-5395, 2020.

[8] Ke Sun, Bin Xiao, Dong Liu, Jingdong Wang, ”Deep high-resolution representation learning for human pose estimation,” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5693–5703, 2019.

[9] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, Jingdong Wang, ”Bottom-up human pose estimation via disentangled keypoint regression,” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14676–14686, 2021.

[10] Y. Chen,Y. Tian, M. He, ”Monocular human pose estimation: A survey of deep learning-based methods,” Computer Vision and Image Understanding, Vol. 192, pp. 102-897, 2020.

[11] K. He, G. Gkioxari, P. Dollar, Piotr R. Girshick, ”Mask r-cnn,” Pro-´ ceedings of the IEEE international conference on computer vision, pp. 2961-2969, 2017.

[12] H.S. Fang, Jiefeng Li, H. Tang, Chao Xu, Haoyi Zhu, Y. Xiu, YongLu Li, Cewu Lu, ”AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[13] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, Jitendra Malik, ”Human pose estimation with iterative error feedback,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4733-4742, 2016.

[14] TV Marcard, Gerard Pons-Moll, Bodo Rosenhahn, ”Multimodal motion capture dataset TNT15,” Leibniz Univ. Hannover, Hanover, Germany, and Max Planck for Intelligent Systems, Tubingen, Germany, Tech. Rep,¨ 2016.

[15] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, John Collomosse, ”Total capture: 3d human pose estimation fusing video and inertial sensors,” Proceedings of 28th British Machine Vision Conference, pp. 1-13, 2017.

[16] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, C. Theobalt, ”Monocular 3d human pose estimation in the wild using improved cnn supervision,” 2017 international conference on 3D vision (3DV), pp. 506-516, 2017.

[17] M. Von, R. Henschel, M.J. Black, Bodo Rosenhahn, G. Pons-Moll, ”Recovering accurate 3d human pose in the wild using imus and a moving camera,” Recovering accurate 3d human pose in the wild using imus and a moving camera, pp. 601-617, 2018.

[18] M. Fabbri, F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, R. Cucchiara, ”Learning to detect and track visible and occluded body joints in a virtual world,” Proceedings of the European conference on computer vision (ECCV), pp. 430-446, 2018.

[19] K.S. Babulal, A.K. Das, ”Deep Learning-Based Object Detection: An Investigation,” Futuristic Trends in Networks and Computing Technologies, pp. 697-711, 2022.

[20] K.S. Babulal, A.K. Das, P. Kumar, D.S. Rajput, A. Alam, A.J. Obaid, ”Real-Time Surveillance System for Detection of Social Distancing,” Vol. 13, pp. 1-13, 2022.

[21] Gaku Nakano, Shoji Nishimura, ”=Real-time Social Distancing Detection System with Auto-calibration using Pose Information,” The First International Conference on AI-ML-Systems, pp. 1-3, 2021.

[22] F. Karray, M. Alemzadeh, Jamil Abou Saleh, Mo Nours Arab, ”Humancomputer interaction: Overview on state of the art,” International journal on smart sensing and intelligent systems, Vol. 1, pp. 137, 2008.

[23] A. Gahlot, P. Agarwal, A. Agarwal, V. Singh, A.K. Gautam, ”Skeleton based human action recognition using Kinect,” International Journal of Computer Applications, Vol. 975, pp. 8887, 2016.

[24] S. Ghorbani, K. Mahdaviani, A. Thaler, K. Kording, D.J. Cook, G. Blohm, N. FTroje, ”Movi: A large multipurpose motion and video dataset,” arXiv preprint arXiv, pp. 2003-01888, 2020.

[25] BeomJun Jo, SeongKi Kim, ”Comparative Analysis of OpenPose, PoseNet, and MoveNet Models for Pose Estimation in Mobile Devices,” Traitement du Signal, Vol. 39, pp. 119-124, 2022.

[26] N. Garau, N. Conci, ”CapsulePose: A variational CapsNet for real-time end-to-end 3D human pose estimation”, Neurocomputing, Vol. 523, pp. 81-91, 2023.

[27] S. Yang, Ze Feng, Z. Wang, Yanjie Li, S. Zhang, Z. Quan, Shu-tao Xia, W. Yang, ”Detecting and grouping keypoints for multi-person pose estimation using instance-aware attention”, Pattern Recognition, Vol. 136, pp. 109-232, 2023.