



A Survey on Deep Learning-Based 2D Human Pose Estimation Models

Sani Salisu^{1,2}, A. S. A. Mohamed^{1,*}, M. H. Jaafar³, Ainun S. B. Pauzi¹ and Hussain A. Younis^{1,4}

¹School of Computer Science, Universiti Sains Malaysia, Penang, 11800, Malaysia

²Department of Information Technology, Faculty of Computing, Federal University Dutse, Jigawa, 720211, Nigeria

³School of Industrial Technology, Universiti Sains Malaysia, Penang, 11800, Malaysia

⁴College of Education for Women, University of Basrah, Basrah, 61004, Iraq

*Corresponding Author: A. S. A. Mohamed. Email: sufril@usm.my

Received: 09 September 2022; Accepted: 29 January 2023; Published: 30 August 2023

Abstract: In this article, a comprehensive survey of deep learning-based (DL-based) human pose estimation (HPE) that can help researchers in the domain of computer vision is presented. HPE is among the fastest-growing research domains of computer vision and is used in solving several problems for human endeavours. After the detailed introduction, three different human body modes followed by the main stages of HPE and two pipelines of two-dimensional (2D) HPE are presented. The details of the four components of HPE are also presented. The keypoints output format of two popular 2D HPE datasets and the most cited DL-based HPE articles from the year of breakthrough are both shown in tabular form. This study intends to highlight the limitations of published reviews and surveys respecting presenting a systematic review of the current DL-based solution to the 2D HPE model. Furthermore, a detailed and meaningful survey that will guide new and existing researchers on DL-based 2D HPE models is achieved. Finally, some future research directions in the field of HPE, such as limited data on disabled persons and multi-training DL-based models, are revealed to encourage researchers and promote the growth of HPE research.

Keywords: Human pose estimation; deep learning; 2D; dataset; models; body parts

1 Introduction

HPE is among the fastest-growing research domains of computer vision used in medical imaging, virtual reality, sports analysis, human-robot interaction [1], activity recognition [2], object detection, surveillance, human-computer interactions and so on. It is among the attractive research domains of computer vision [3]. HPE is a task that intends to localize the human body joints in images and videos [4]. The impact in key points recognition and image segmentation was first discovered when the decision tree algorithm was applied to computer vision and became one of the key elements (and



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

with a huge volume of artificial training data) in the breakthrough success of human pose estimation from Kinect depth image [5].

The located human body joints or keypoints are used to build what is known as human body modelling (including skeleton pose) from the image or video input data. The features and joints extracted from the visual input data are represented by the body modelling. Usually, the model-based method is used to interpret and deduce human body poses and produce 2D or three-dimensional (3D) poses. (x, y) coordinates for each joint from the red, green and blue (RGB) image and (x, y, z) coordinates for every joint from the RGB image are estimated in the 3D and 2D pose estimation, respectively.

In very recent research works, there is an active interest shown in estimating the semantic keypoints of the human body which include the knee, upper shoulder, lower shoulder, and head for different purposes. In this context, research focus has been limited to traditional or classical approaches to articulate pose estimation using a pictorial structure framework. The classical approaches, however, are attached with the limitations of estimating a pose independent on image data and insufficient enough to determine the accurate location of the body keypoints. For these reasons, HPE research focused on enriching the representational power of the process. Recently, pose estimation has been greatly reshaped by deep-learning (DL) approaches. These DL-based approaches are efficient in extracting more sufficient and significant features from the input data. Such an approach has yielded promising outcomes and outpaced non-deep learning approaches [6]. Today, many researchers use machine learning (ML) and DL concept in different applications such as agriculture [7–10], environment [11–13], and cyber security [14]. The emergence of deep learning also presents new solutions to conventional computer vision branches like image classification and detection [15] as well as complex tasks which include image fusion [16] and image stitching.

Despite its outstanding performance in solving issues related to pose estimation, DL methods are facing challenges in detecting, capturing, and extracting the significant keypoints of the human body. Such challenges include occlusion (self-occlusion, inter-person occlusion, and out-of-frame occlusion), limited data (limited annotation, limited variation of pose, limited number of pose), bad input data (blurry, low resolution, low light, low contrast, small scale, noisy), domain gap, camera-centric, crowd scenes, speed, etc. Many researchers were inspired by [17] to employ the DL method to minimize such challenges facing HPE-accurate outcomes. Moreover, DL-based HPE has made a significant improvement recently in the tasks of single-person pose estimation in the top-down approach in an image [18] and videos [19] as well as multi-person pose estimation in monocular videos or images [20]. Table 1 shows the most cited HPE article from 2014 to 2022. All these improvements have been enabled by the use of a DL framework [21] and the availability of huge-scale benchmark public datasets which include “Leeds Sports Pose” (LSP) [22], “Frames Labelled in Cinema” (FLIC), “Microsoft Common Object in Context” (MS-COCO) [23], “Max Planck institute for informatics (MPII Human pose)” and so on. The most popular datasets used in 2D HPE research are shown in Fig. 1.

The availability of countless published research articles on HPE inspired researchers to publish a lot of informative surveys and review articles to serve as a guide for researchers in tackling HPE tasks. As the deep learning approaches are becoming more significant in solving complex tasks on

HPE, many researchers focused only on single-person DL pose estimation models, others focused on multi-person DL pose estimation only. As such, this survey highlighted both single-person and multi-person DL pose estimation [24]. Numerous surveys on using DL models to solve HPE problems were conducted [25]. But, to our knowledge there exists no survey that focuses on details of state-of-the-art DL-based 2D HPE. For example, a survey [26] only reviewed the DL-based 3D and 2D HPE and summarized the challenges and the benchmark dataset. A detailed review of the current DL-based articles for 3D pose estimation and a summary of the merit and demerit of those methods are presented [27].



Figure 1: Most popular 2D datasets for human pose estimation

Another study provides a review of current study on multi-person pose estimation and analyses the algorithms and compares their advantages and the disadvantages to fill in the gap of the existing surveys [28]. Most of the previous surveys focused on reviewing the DL approach in tackling HPE issues but did not consider summarizing and tabulating the development of DL in HPE from the year of breakthrough to date. In this paper the details component of HPE using the DL model are provided. This survey intends to highlight the limitation of published reviews and surveys in terms of presenting the systematic review of the current DL-based solution to 2D HPE. Besides that, the survey will guide the new researchers on computer vision 2D HPE. The following key points differentiate this survey from other surveys.

- Detailed components of HPEs are presented which include backbone, loss function, 2D datasets and evaluation metrics
- Summary of DL-based articles for HPE from the year of breakthrough to date (2014–2022)
- Overview of 2D HPE.

Table 1: Highly cited articles in deep learning based HPE

S/n	Study	Citations	Method/algorithm	Year
1	[29]	40	Revisiting skeleton-based action recognition	2022
2	[30]	17	Human-computer interaction	2022
3	[31]	52	High-resolution network	2021
4	[32]	43	Pose-guided representation learning	2021
5	[33]	40	Efficient pose:	2021
6	[34]	27	Human pose estimation	2021
7	[35]	269	Human pose estimation	2020
8	[36]	194	Human pose estimation	2020
9	[37]	1166	Deep high-resolution representation	2019
10	[38]	1892	Human pose estimation	2019
11	[39]	911	Human pose estimation	2018
12	[40]	903	Multi-person pose estimation	2018
13	[41]	19930	R-CNN	2017
14	[42]	7690	Part affinity fields	2017
15	[43]	4151	Human pose estimation	2016
16	[44]	2733	Convolutional pose machines	2016
17	[26]	1298	Convolutional networks	2015
18	[45]	555	Human pose estimation	2015
19	[19]	2653	Deep neural networks	2014
20	[46]	2026	Human pose estimation	2014

2 Human Body Modelling

As humans are different, so also their shapes and sizes. Human body modelling is one of the most significant aspects of HPE. To estimate the pose of a given body, the body must satisfy the attributes required for a particular task to create and define the human body pose. Three distinct categories of human body models commonly used in HPE [47] which include the kinematic-base model, planner model, and volumetric model as shown in Fig. 2.

2.1 Kinematic-Based Model

This type of model is popularly known as a skeletal-based model or stick figure and can be described as a simple and flexible human body structure frequently used in 2D [48] and 3D HPE [49]. This type of model is said to represent the joints' locations and limb orientation to present the human body, and skeletal structure that can be used to detect relationships between body parts.

2.2 Planer Model

The planer model, known as the contour-based model, and quite different from the kinematic-base model. In the contour-base model, keypoints are roughly symbolized with rectangles or boundaries of

an object's shape. The contour-based model is used to present the silhouette and form of the human body. The planner model is commonly used in classical HPE approaches [50] that used cardboard mode [51] and active shape mode to capture human body graph and the silhouette distortion using principal component analysis (PAC).

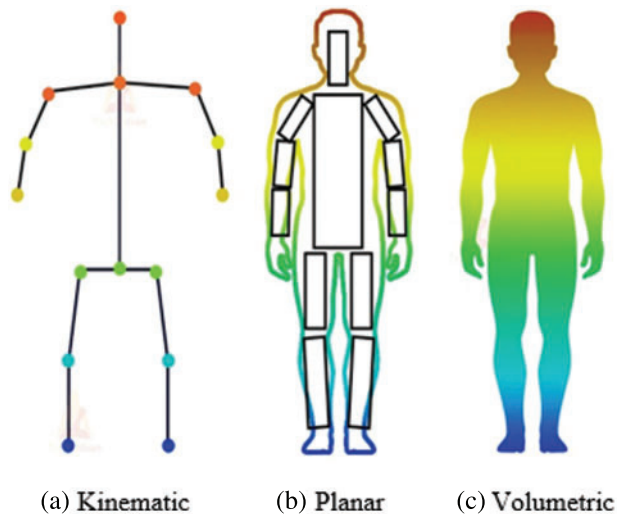


Figure 2: Three different human body model

2.3 Volumetric Model





















Volume-based model is used to represent three-dimensional object's silhouette and pose with geometric mesh. Traditional geometric mesh for modelling human body parts was cylindrical, conic, etc., while novel volume-based are characterised in mesh form captured with the 3D scans. The most popular volumetric model used for 3D pose estimation are stitched puppet model (SPM) and unified deformation model (UDM) [52], Frankenstein & Adam and generic human model (GHUM) & low-resolution generic human model (GHUML) (ite) [53].

3 Basic Stages in Human Pose Estimation

The key procedure of HPE is poached into two stages: i) human body keypoints/joints localization and ii) formation of valid human pose configuration by grouping the localized joints/keypoints [54]. Finding the location of key points on the human body (knee, ankle, shoulder, head, arms, hands.) is the focus in the first stage. Different human pose dataset format is used in gathering and identifying the key points stored in the selected datasets. The output of body key points of the same image may vary from the type of dataset format and platform adapted. For example, Max Planck Institute for Informatics (MPII) Human pose datasets provide only 14 body joints, while Microsoft Common Objects in Context (MSCOCO) dataset provide 17 body joints. Table 2 shows the output of MSCOCO and MPII datasets.

The second stage of pose estimation is grouping the localized keypoints into valid human pose structures to determine the pair organs of human body. Many researchers applied different techniques in joining the key points candidates [55].

Table 2: Keypoints out of MSCOCO and MPII datasets

MSCOCO	Output Format	MPII	Output Format
	0		0
	1		1
	2,3		2,3
	4,5		4,5
	6,7		6,7
	8,9		8,9
	10,11		10,11
	12,13		12,13
	14,15		-
	16,17		-
	18		18

4 2D Human Pose Estimation

A human pose can be estimated from a 2D image and video by locating the site of human body keypoints in an image or video. Pictorial structure techniques [56], handcraft feature extraction techniques [57], and other sophisticated body models [58] were used for 2D HPE. Most of these traditional approaches depend completely on human supervision and describe the human body as a stick figure in obtaining the local and global pose structure. In more recent years, pose estimation has been greatly reshaped by machine learning approaches. The DL-based approach has accomplished a great progress in HPE by enhancing the accomplishment considerably in categorizing the methodology into single-person pose estimation and multi-person pose estimation.

4.1 2D Single Person Pose Estimation

2D single-person pose estimation is used to predict the sets of 2D joints (keypoints) position (x, y) of the human body from the input image. In this approach, the bounding boxes of the person are presented before the estimation process. To address the issue of 2D single-person pose estimation, two categories of pipelines that used deep learning techniques such as regression-based model and detection-based model are provided.

4.1.1 Regression-Based Model

Since the main objective of the HPE problem is to locate and estimate the keypoints of human body parts, many researchers used a regression framework to accomplish that. Alex network (AlexNet) inspired many researchers [59–61] to adopt a regression-based method to predict human keypoints from the input image. Reference [19] was impressed by the wonderful performance of AlexNet and attempted to train a similar deep neural network known as DeepPose to predict a set of keypoints position on the input image. Furthermore, cascade regressors were employed to improve the precision of the location of each joint as shown in Fig. 3 once a joint position is identified, in the initial stage, the cropped image along with predicted joints are then fed to the network in the next stage to predict the refine value of the joints position in the patch and produce the finest joints position over multiple stages.

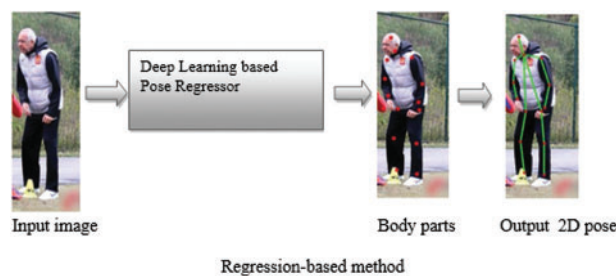


Figure 3: Regression-based method

Due to the remarkable performance of DeepPose in solving HPE, the DL approach has become more and more popular in HPE research. Reference [62] developed a model called Iterative Error Feedback where the whole prediction started with the mean pose skeletal which is then updated iteratively over many steps to match the ground truth. Reference [63] proposed a differentiable special to numerical transform (DSNT) which is an end-to-end regression method for HPE similar to the soft-argmax function to change feature maps into joint coordinates in a completely differentiable structure.

4.1.2 Detection-Based Model

A detection-based method is also known as a heatmap-based method. In this approach, the pose estimation network is trained to reduce the inconsistency between the target heatmaps and the predicted heatmaps. The target heatmap (ground-truth heatmap) is generated by a 2D Gaussian centered at the ground-truth joint location. Thus, a detection-based method for HPE is aimed to train the body part detector to predict the position of body joints and is used to address the estimation issue with a heatmap prediction approach as shown in Fig. 4.

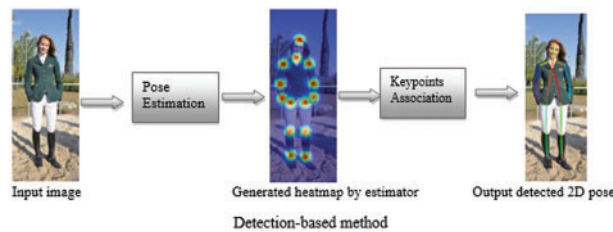


Figure 4: Detection-based method

One advantage of the heatmap-based method over a regression-based method is the provision of highly rich informative information by conserving the spatial position information to ease the training of the convolution neural network. This brings about the recent increasing interest in adopting the heatmap-based method to present the joint location and develop the effective convolutional neural network (CNN) architecture as in [64]. A novel loss function called heatmap weighting loss was proposed in [65] to generate weights for each pixel on the heatmap which makes the model more focused on keypoints. Balakrishnan et al. [66] integrated Transformer encoder with a recently proposed Bottleneck Transformer [67] and apply the model to the problem of 2D HPE.

Reference [68] proposed a novel approach known as “Dense, Multi-target Votes”, where every location in the image votes for the site of each keypoints using a convolutional neural net. The voting scheme enables the utilization of information from the whole image, instead of depending on a sparsely set of keypoints positions.

Reference [69] presented another method for pretraining 2D pose estimation networks in order to learn the position of each spot from an image composed of shuffled spots. Another achievement made in the heatmap-based method is the proposed novel model based on transformer architecture, improved with a feature pyramid fusion structure to predict the keypoints heatmap [70].

4.2 2D Multi-Person Pose Estimation

Multi-person pose estimation is a more complex and challenging problem that combines the task of identifying the number of persons, their positions and poses and then grouping their localized body keypoints together. To resolve these problems, multi-person pose estimation can be classified into Top-down and Bottom-up pipelines.

4.2.1 Top-Down Method

The top-down approach involved two main stages: detection of all persons bounding boxes in an image separately using a body detector and prediction of the location of keypoints within the detected bounding boxes using a single-person estimator. Therefore, the human body detector and single-person pose estimator remains the most significant components of the top-down HPE pipeline. Different HPE methods show some common pose errors until the design of PoseFix net [71] based on related pose error distribution from different HPE methods to improve the estimated pose from any framework. Overcrowding and complex pose are the main challenges of HPE using a top-down pipeline, as shown in Fig. 5.

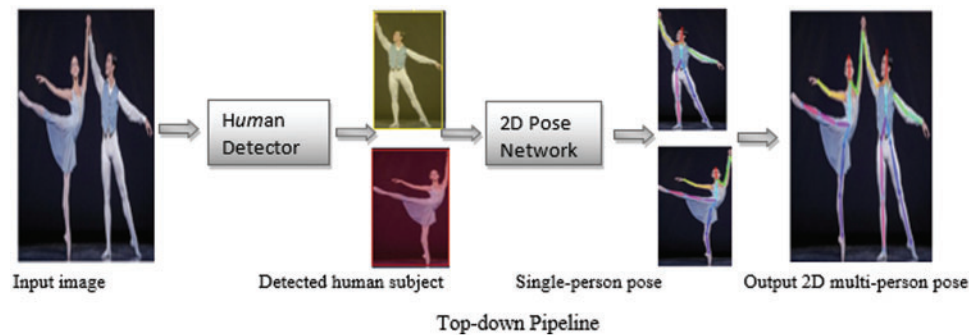


Figure 5: Top-down pipeline for multi-person pose estimation

4.2.2 Bottom-Up Method

Estimation of human poses using the bottom-up method can be achieved by first detecting the body joints and then grouping the joint candidate for a unique pose. Typically, there are two main components in bottom up HPE include body joints detection and joints candidate grouping, as shown in Fig. 6 these two components are tackled one after the other by most HPE Algorithms. One of the earliest bottom-up-based Algorithms known as Deep-Cut was proposed in [72]. The model will first detect all the body parts candidate and then labels each part to its corresponding part and assemble them using integer linear programming (ILP) to estimate the pose. DeepCut is computationally expensive, which is why [73] proposed DeeperCut to improve the DeepCut by applying a deeper, stronger, and faster body part detector to improve performance and faster computational speed. Bottom-up is faster compared to the top-down approach. However, bottom-up faced the main challenge of grouping the corresponding body parts when people are with large overlap.

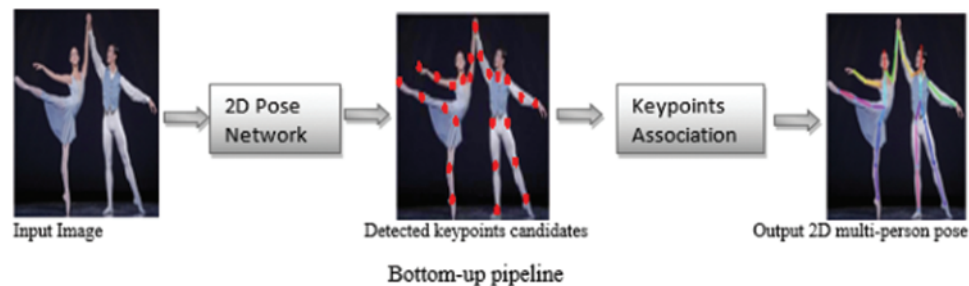


Figure 6: Bottom-up pipeline for multi-person pose estimation

5 Components of Human Pose Estimation Systems

Research in HPE resides in four main components which include backbone architecture, loss functions, the datasets and the evaluation metrics employed.

5.1 Backbone Architecture

Many researchers employ different backbone networks in solving HPE problems. These feature extractors networks are used by the model to extract features from the input image. AlexNet network architecture is the first popular backbone architecture to be implemented by DeePoPose while applying

DL to HPE. Due to the low receptive field, another popular backbone architecture called visual geometry group network (VGGNet) [74] is used in [75] to enlarge the receptive field for large-scale feature extraction. Other deep learning algorithms such as residual convolutional neural network (R-CNN), FastR-CNN, Faster R-CNN and MaskR-CNN have been used as backbone architecture for HPE research. In other research articles on HPE [76], ResNet has been use as a backbone architecture.

5.2 Loss Function

HPE models learned by use of loss function. The modelling of a given dataset by a specific algorithm is evaluated by the loss function. Reference [77] mentioned that the loss function calculates and decreased the error in the prediction process. There are three types of loss- function that is usually applied in the human pose estimation model [78] which include Mean squared Error (MSE), Mean Absolute Error (MAE) and Cross-Entropy loss. These types of loss functions are represented by the following equations. Eq. (1) is Mean Absolute Error (MAE), Eq. (2) is Mean squared Error (MSE), and Eq. (3) is Cross-Entropy loss, respectively.

$$L_1 = 1/n \sum_{i=1}^n |y_{i=i} - f(x_i)| \quad (1)$$

$$L_2 = 1/n \sum_{i=1}^n (y_{i=i} - f(x_i))^2 \quad (2)$$

$$\text{Logloss} = - (y \log(f(x_i)) + (1 - y) \log(1 - f(x_i))) \quad (3)$$

5.3 Datasets

Neural networks required a huge amount of data for training and testing purposes. For accurate hope estimation, a dataset of multiple poses is essential. These datasets are required to provide a fair evaluation among various algorithms. The widely used datasets for the 2D DL-based HPE method are quite many. However, only a few datasets are used in 2D HPE tasks due to various limitations such as limited data and lack of multiple object articulation. Among the popular large-scale 2D HPE datasets are:

5.3.1 Microsoft Common Object in Context

Popularly known as MS-COCO dataset [79]. A Microsoft product and one of the widely used datasets for HPE, consist of 200,000 labeled subjects with keypoints and more than 330,000 images and every person is labelled with 17 joints.

5.3.2 Max Plank Institute for Informatics

Max plank institute for informatics (MPII) HPE dataset [48]. This dataset contains about 25,000 images and annotated body joints of over 40,000 persons. This dataset includes about 410 labeled images of human activities. In MPII datasets, everyone's is labeled with 15 joints.

5.3.3 Leeds Sport Pose Datasets

Leeds sport pose datasets (LSP)and LSP extension (LSPe) datasets for HPE have a set of 11,000 training and 1,000 testing images from Flickr (an American image hosting and video hosting service). Most of the images in LSP are from sports activities having unusual poses with challenging appearance terminologies [80]. In LSP dataset, the full body of every person is labeled with 14 joints.

5.3.4 Frames Labeled in Cinema

Frames labeled in cinema (FLIC) dataset is another 2D HPE benchmark consisting of 5,003 images with about 80% images for training and 20% for testing. This dataset is formed from Hollywood movies. It contained several images with different poses and clothes. Every person is labeled with 10 body joints. This dataset is accurate for both single and multi-person pose estimation.

5.4 Evaluation Metrics

Several features and exigencies that need to be considered make it difficult to evaluate the performance of HPE. Such features and requirements include unusual poses, body size, single/multi-person, upper/lower or full body pose estimation. Consequently, researchers used different evaluation metrics for 2D HPE. Below are some popular evaluation metrics for 2D HPE.

5.4.1 Percentage of Corrected Parts

Percentage of Corrected Parts (PCP). This metric is the most powerful in earlier hope estimation research work. It is used to evaluate stick predictions to report the localization accuracy for the body parts. A body part is said to be detected if the distance between the detected joint and true joints is less than half the body part [81].

5.4.2 Percentage of Detected Joints

The percentage of Detected Joints (PDJ) detected joints is said to be accurate if the distance between the predicted and true joints is within a certain fraction of torso diameter 20. Eg. PDJ @ 0.2 implies that the distance between the predicted and true joint is less than $0.2 \times$ torso diameter. Where the torso diameter is the central diameter. This metric is proposed to address the drawback of PCP.

5.4.3 Percentage of Corrected Keypoints

Percentage of corrected keypoints (PCK), is used to measure the accuracy of localization of different keypoints. PCK is denoted as PCKh @ 0.5 when the threshold is set to 50% of the head segment length of each test image.

5.4.4 Average Precision and Average Recall

Average Precision (AP) and Average Recall (AR), AP measure is an index to calculate the accuracy of keypoints detection based on precision and recall.

6 Conclusion and Future Work

The DL-based 2D HPE models from past surveys, reviews and results articles, using systematic review, are investigated in this survey. A detailed and meaningful survey that will guide new and existing researchers on DL-based 2D HPE models is achieved. Despite all the achievements, there exist many challenges, which include occlusion (self-occlusion, inter-person occlusion and out-of-frame occlusion), limited data (limited annotation, limited variation of pose, limited number of poses), bad input data (blurry, low resolution, low light, low contrast, small scale, noisy), domain gap, camera-centric, crowd scenes, speed and so on. However, future research directions in the field of HPE to encourage researchers and promote the growth of HPE research were revealed. Limited data on disabled persons hinders HPE research on disabled persons. They can be improved by creating a

huge dataset of people with different disabilities and making it available for researchers. Considering the nature of different human body parts and shapes, unusual articulation and magical movement may occur that will require multiple training models to tackle these issues since most researchers are adopting a single model to tackle the normal situation. Our survey is limited to DL-based 2D human pose estimation only with the hope that this survey will serve as a guide for the existing researchers and motivation for new researchers in the domain of HPE.

Acknowledgement: We acknowledged the support given to us by the Universiti Sains Malaysia in carrying out our research work.

Funding Statement: This work was supported by the [Universiti Sains Malaysia] under FRGS Grant Number [FRGS/1/2020/STG07/USM/02/12(203.PKOMP.6711930)] and FRGS Grant Number [304PTEKIND.6316497.USM.].

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

- [1] C. Xu, X. Yu, Z. Wang and L. Ou, "Multi-view human pose estimation in human-robot interaction," in *Proc. IECON*, Singapore, pp. 4769–4775, 2020.
- [2] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue *et al.*, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1963–1978, 2019.
- [3] J. Cha, M. Saqlain, C. Lee, S. Lee, D. Kim *et al.*, "Towards single 2D image-level self-supervision for 3D human pose and shape estimation," *Applied Science*, vol. 11, no. 20, pp. 1–19, 2021.
- [4] W. Li, R. Du and S. Chen, "Semantic–structural graph convolutional networks for whole-body human pose estimation," *Information*, vol. 13, no. 3, pp. 1–14, 2022.
- [5] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook *et al.*, "Efficient human pose estimation from single depth images," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2821–2840, 2013.
- [6] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.
- [7] M. A. Haq, "Planetscope nanosatellites image classification using machine learning," *Computer Systems Science & Engineering*, vol. 42, no. 3, pp. 1031–1046, 2022.
- [8] I. M. Hayder, G. AL-Ali and H. A. Younis, "Predicting reaction based on customer's transaction using machine learning approaches," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 1, pp. 1086–1096, 2023.
- [9] M. A. Haq, "CNN based automated weed detection system using UAV imagery," *Computer Systems Science & Engineering*, vol. 42, no. 2, pp. 837–849, 2021.
- [10] H. A. Younis, A. S. A. Mohamed, M. N. Ab Wahab, R. Jamaludin, S. Salisu *et al.*, "A new speech recognition model in a human-robot interaction scenario using NAO robot: Proposal and preliminary model," in *Int. Conf. on Communication & Information Technology (ICICT)*, Basrah, Iraq, pp. 215–220, 2021.
- [11] M. A. Haq, "Smotednn: A novel model for air pollution forecasting and aqi classification," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 1403–1425, 2022.
- [12] M. A. Haq, A. K. Jilania and P. Prabu, "Deep learning-based modeling of groundwater storage change," *Computers, Materials & Continua*, vol. 70, no. 3, pp. 4599–4617, 2022.

- [13] M. A. Haq, G. Rahaman, P. Baral and A. Ghosh, "Deep learning based supervised image classification using UAV images for forest areas classification," *Journal of the Indian Society of Remote Sensing*, vol. 49, no. 3, pp. 601–606, 2021.
- [14] C. S. Yadav, J. Singh, A. Yadav, H. S. Pattanayak, R. Kumar *et al.*, "Malware analysis in IoT & android systems with defensive mechanism," *Electronics*, vol. 11, no. 15, pp. 1–20, 2022.
- [15] S. Wang, M. E. Celebi, Y. Zhang, X. Yu, S. Lu *et al.*, "Advances in data preprocessing for bio-medical data fusion: An overview of the methods, challenges, and prospects," *Information Fusion*, vol. 76, pp. 376–421, 2021.
- [16] Y. Zhang, Z. Dong, S. Wang, X. Yu, X. Yao *et al.*, "Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation," *Information Fusion*, vol. 64, pp. 149–187, 2020.
- [17] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. ICCVPR*, Columbus OH, USA, pp. 1653–1660, 2014.
- [18] F. Zhang, X. Zhu and C. Wang, "Single person pose estimation: A survey," arXiv:2109.10056, 2021.
- [19] H. Ullah, I. U. Islam, M. Ullah, M. Afaq, S. D. Khan *et al.*, "Multi-feature-based crowd video modeling for visual event detection," *Multimedia System*, vol. 27, no. 4, pp. 589–597, 2021.
- [20] Y. Tian, H. Zhang, Y. Liu and L. Wang, "Recovering 3D human mesh from monocular images: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2023.
- [21] X. Huang, X. Wang, W. Lv, X. Bai, X. Long *et al.*, "PP-YOLOv2: A practical object detector," arXiv:2104.10419, 2021.
- [22] X. Xu, H. Chen, F. Moreno-Noguer, L. A. Jeni, F. de la Torre *et al.*, "3D human pose, shape and texture from low-resolution images and videos," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4490–4504, 2022.
- [23] C. H. Nguyen, T. C. Nguyen, T. N. Tang and N. L. H. Phan, "Improving object detection by label assignment distillation," in *Proc. IEEE/CVF Winter, A.C.V, WACV*, Waikoloa, HI, USA, pp. 1322–1331, 2022.
- [24] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler *et al.*, "Efficient object localization using convolutional networks," in *Proc. IEEE/CVPR*, Boston, MA, USA, pp. 648–656, 2015.
- [25] C. Bisogni and A. Castiglione, "Head pose estimation: An extensive survey on recent techniques and applications," *Pattern Recognition*, vol. 127, pp. 1–14, 2022.
- [26] Y. Chen, Y. Tian and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Computer Vision and Image Understanding*, vol. 192, pp. 1–23, 2020.
- [27] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng *et al.*, "Deep 3D human pose estimation: A review," *Computer Vision and Image Understanding*, vol. 210, pp. 1–21, 2021.
- [28] A. Kamboj, R. Rani and A. Nigam, "A comprehensive survey and deep learning-based approach for human recognition using ear biometric," *Visual Computer*, vol. 38, pp. 2383–2416, 2021.
- [29] K. Chen, D. Lin and B. Dai, "Revisiting skeleton-based action recognition," in *2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, no. 1, pp. 2969–2978, 2022.
- [30] H. Liu, S. Member, T. Liu and Z. Zhang, "ARHPE: Asymmetric relation-aware representation learning for head pose estimation in industrial human-computer interaction," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 7197–7117, 2022.
- [31] C. Yu, C. Xiao, B. Gao, L. Yuan, L. Zhang *et al.*, "Lite-HRNet: A lightweight high-resolution network," in *Proc. IEEE/CSC/CVPR*, Nashville, TN, USA, pp. 10435–10445, 2021.
- [32] Q. Wu, A. Zhu, R. Cui, T. Wang, F. Hu *et al.*, "Pose-guided inflated 3D convnet for action recognition in videos," *Signal Processing: Image Communication*, vol. 91, pp. 1–9, 2021.
- [33] D. Groos, H. Ramampiaro and E. Af Ihlen, "EfficientPose: Scalable single-person pose estimation," *Applied Intelligence*, vol. 51, pp. 2518–2533, 2021.
- [34] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang *et al.*, "TokenPose: Learning keypoint tokens for human pose estimation," in *Proc. IEEE/ICCV, Montreal*, QC, Canada, pp. 11293–11302, 2021.

- [35] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang *et al.*, “HigherhrNet: Scale-aware representation learning for bottom-up human pose estimation,” in *Proc. IEEE CVF/CVPR*, Seattle, WA, USA, pp. 5385–5394, 2020.
- [36] F. Zhang, X. Zhu, H. Dai, M. Ye and C. Zhu, “Distribution-aware coordinate representation for human pose estimation,” in *Proc. IEEE/CVF/CVPR*, Seattle, WA, USA, pp. 7091–7100, 2020.
- [37] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Dang *et al.*, “Deep high-resolution representation learning for visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2021.
- [38] K. Sun, B. Xiao, D. Liu and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in *Proc. IEEE/CVF/CVPR*, Long Beach, CA, USA, pp. 5686–5696, 2019.
- [39] R. A. Güler, N. Neverova and I. Kokkinos, “DensePose: Dense human pose estimation in the wild,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake, UT, USA, pp. 7297–7306, 2018.
- [40] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu *et al.*, “Cascaded pyramid network for multi-person pose estimation,” in *Proc. IEEE/CVF/CVPR*, Salt Lake, UT, USA, pp. 7103–7112, 2018.
- [41] K. He, G. Gkioxari, P. Dollár and R. Girshick, “Mask R-CNN,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2017.
- [42] Z. Cao, T. Simon, S. E. Wei and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” in *Proc. IEEE/CVF, CVPR*, Honolulu, HI, USA, pp. 1302–1310, 2017.
- [43] A. Newell, K. Yang and J. Deng, “Stacked hourglass networks for human pose estimation,” in *Proc. ECCV*, Netherlands, pp. 483–499, 2016.
- [44] S. E. Wei, V. Ramakrishna, T. Kanada and Y. Sheikh, “Convolutional pose machines,” in *Proc. IEEE/CVF, CVPR*, Las Vegas, NV, USA, pp. 4724–4732, 2016.
- [45] T. Pfister, J. Charles and A. Zisserman, “Flowing convnets for human pose estimation in videos,” in *Proc. IEE Explore, ICCV Santiago*, Chile, pp. 648–656, 2015.
- [46] M. Andriluka, L. Pishchulin, P. Gehler and B. Schiele, “2D human pose estimation: New benchmark and state of the art analysis,” in *Proc. IEEE/CVF, CVPR*, Columbus, OH, USA, pp. 3686–3693, 2014.
- [47] W. Gong, S. Zhang, J. Gonzalez, A. Sobral, T. Bouwmans *et al.*, “Human pose estimation from monocular images: A comprehensive survey,” *Sensors*, vol. 16, no. 12, pp. 1–39, 2016.
- [48] S. Johnson and M. Everingham, “Clustered pose and nonlinear appearance models for human pose estimation,” in *Proc. BMVC*, Wales, UK, no. ii, pp. 1–11, 2010.
- [49] D. Mehta, H. Rhoden, D. Casas, P. Fua, O. Sotnychenko *et al.*, “Monocular 3D human pose estimation in the wild using improved CNN supervision,” in *Proc. IEEE/IC3DV*, Qindao, China, pp. 506–516, 2018.
- [50] H. Jiang, “Finding human poses in videos using concurrent matching and segmentation,” in *Proc., ACCV*, Queenstown, New Zealand, pp. 228–243, 2010.
- [51] O. Freifeld, A. Weiss, S. Zuffi and M. J. Black, “Contour people: A parameterized model of 2D articulated human shape,” in *Proc. IEEE/CSC, CVPR*, San Francisco, CA, USA, pp. 639–646, 2010.
- [52] H. Joo, T. Simon and Y. Sheikh, “Total capture: A 3D deformation model for tracking faces, hands, and bodies,” in *Proc. IEEE/CVF, CVPR*, Salt Lake, UT, USA, pp. 8320–8329, 2018.
- [53] H. Xu, E. G. Bazavan, A. Zanfiri, W. T. Freeman, R. Sukthankar *et al.*, “GHUM GHUML: Generative 3D human shape and articulated pose models,” in *Proc. IEEE/CVF, CVPR*, Seattle, WA, USA, pp. 6183–6192, 2020.
- [54] L. Pishchulin, M. Andriluka, P. Gehler and B. Schiele, “Poselet conditioned pictorial structures,” in *Proc. IEEE CVPR*, Portland, OR, USA, pp. 588–595, 2013.
- [55] B. X. Nie, C. Xiong and S. C. Zhu, “Joint action recognition and pose estimation from video,” in *Proc. IEEE/CVPR*, Boston, MA, USA, pp. 1293–1301, 2015.
- [56] M. Andriluka, S. Roth and B. Schiele, “Pictorial structures revisited: People detection and articulated pose estimation,” in *Proc. IEEE/CVPR*, Miami, FL, USA, pp. 1014–1021, 2009.
- [57] M. Dantone, J. Gall, C. Leistner and L. van Gool, “Human pose estimation using body parts dependent joint regressors,” in *Proc. IEEE/CVPR*, Portland, OR, USA, pp. 3041–3048, 2013.

- [58] G. Gkioxari, B. Hariharan, R. Girshick and J. Malik, "Using k-poselets for detecting people and localizing their keypoints," in *Proc. IEEE/CVPR*, Columbus, OH, USA, pp. 3582–3589, 2014.
- [59] W. Mao, Y. Ge, C. Shen, Z. Tian, X. Wang *et al.*, "Poseur: Direct human pose regression with transformers," arXiv:2201.07412, pp. 1–15, 2022.
- [60] W. Mao, Y. Ge, C. Shen, Z. Tian, X. Wang *et al.*, "TFPose: Direct human pose estimation with transformers," arXiv:2103.15320, pp. 1–15, 2021.
- [61] P. Panteleris and A. Argyros, "PE-Former: Pose estimation transformer," in *Proc. ICPRAL*, Paris, France, pp. 1–14, 2022.
- [62] J. Carreira, P. Agrawal, K. Fragkiadaki and J. Malik, "Human pose estimation with iterative error feedback," in *Proc. IEEE/CVF, CVPR*, Las Vegas, NV, USA, pp. 4733–4742, 2016.
- [63] A. Nibali, Z. He, S. Morgan and L. Prendergast, "Numerical coordinate regression with convolutional neural networks," arXiv:1801.07372, pp. 1–10, 2018.
- [64] L. Ke, H. Qi, M. C. Chang and S. Lyu, "Multi-scale supervised network for human pose estimation," in *Proc. IEEE/ICIP*, Athens, Greece, pp. 564–568, 2018.
- [65] S. Li and X. Xiang, "Lightweight human pose estimation using heatmap-weighting loss," arXiv:2205.10611, pp. 1–7, 2022.
- [66] K. Balakrishnan and D. Upadhyay, "BTranspose: Bottleneck transformers for human pose estimation with self-supervised pre-training," arXiv:2204.10209, pp. 1–24, 2022.
- [67] A. Srinivas, T. Y. Lin, N. Parmar, J. Shlens, P. Abbeel *et al.*, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF, CVPR*, Nashville, TN, USA, pp. 16514–16524, 2021.
- [68] I. Lifshitz, E. Fetaya and S. Ullman, "Human pose estimation using deep consensus voting," arXiv:1603.08212, pp. 246–260, 2016.
- [69] K. Zhang, R. Wu, P. Yao, K. Deng, D. Li *et al.*, "Learning heatmap-style jigsaw puzzles provides good pretraining for 2d human pose estimation," arXiv:2012.07101, pp. 1–13, 2020.
- [70] Z. Xiong, C. Wang, Y. Li, Y. Luo and Y. Cao, "Swin-pose: Swin transformer based human pose estimation," arXiv:2201.07384, pp. 1–6, 2022.
- [71] G. Moon, J. Y. Chang and K. M. Lee, "Posefix: Model-agnostic general human pose refinement network," in *Proc. IEEE/CVF, CVPR*, Long Beach, CA, USA, pp. 7765–7773, 2019.
- [72] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka *et al.*, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE/CVF, VCPR*, Las Vegas, NV, USA, pp. 4929–4937, 2016.
- [73] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. ECCV*, Amsterdam, Netherlands, pp. 34–50, 2016.
- [74] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, San Diego, CA, USA, pp. 1–14, 2015.
- [75] Y. Zhao, R. Han and Y. Rao, "A new feature pyramid network for object detection," in *Proc. IEEE/ICVRIS*, Jishou, China, pp. 428–431, 2019.
- [76] Z. Su, M. Ye, G. Zhang, L. Dai and J. Sheng, "Improvement multi-stage model for human pose estimation," arXiv:1902.07837, 2019.
- [77] J. Brownlee, "Loss and loss functions for training deep learning neural networks," *Machine Learning Mastery*, 2019. [Online]. Available: <https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks>
- [78] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang *et al.*, "The progress of human pose estimation: A survey and taxonomy of models applied in 2D human pose estimation," *IEEE Access*, vol. 8, pp. 133330–133348, 2020.
- [79] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Peron *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, Zurich, Switzerland, pp. 740–755, 2014.

- [80] K. Sun, C. Lan, J. Xing, W. Zeng, D. Liu *et al.*, “Human pose estimation using global and local normalization,” in *Proc. IEEE/ICCV*, Venice, Italy, pp. 5600–5608, 2017.
- [81] M. Eichner, M. Marin-Jimenez, A. Zisserman and V. Ferrari, “2D articulated human pose estimation and retrieval in (almost) unconstrained still images,” *International Journal of Computer Vision*, vol. 99, no. 2, pp. 190–214, 2012.