IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Human Pose Detection and Estimation

[1]Srividya Inampudi, [2]Romik Amipara, [3]Rutvik Kokate

[1]Student, [2]Student, [3]Student

[1]Computer Engineering Department,

[1]Fr. C. Rodrigues Institute of Technology,

Navi Mumbai, India

***Abstract:*** In an era of booming technology, the application of Artificial Intelligence and Machine Learning in day to day activities is increasing rapidly. In recent years, there has been a significant increase in the implementation of Artificial Intelligence in Human Pose Detection and Estimation, particularly pose tracking. This paper primarily focuses and summarizes the recent progress made in this field. The methodologies used, datasets encountered, and future research scope in pose estimation applications have also been presented.

***Index Terms*** **- Neural Network, Pose Estimation, Pose Detection, Machine Learning, Artificial Intelligence.**

## I. INTRODUCTION

Human Pose Estimation has myriad applications that can be used in a wide variety of fields, i.e., in healthcare, sports, fitness, criminal investigation. Currently, this technology is being used for Activity recognition, Motion Capture and Augmented Reality, Motion tracking, etc. Artificial intelligence is transforming the fitness industry with new innovative technologies. In today's world, Fitness and Technology are two major industries that are proliferating. The fitness industry witnessed an estimated $94 billion in 2018, with an annual growth rate of 6.1%. Whereas the AI industry of Technology dominated the market as one of the fastest-growing with 38.8% of the Compound Annual Growth Rate (CAGR) [1].

Human pose estimation is an essential technology under the computer vision community, which has been in use for the past two decades. It is a remarkable feat to understand and analyze people's gesture/posture in videos and images [2]. With access to 1.8 billion digital photos every day, a lot of data is available to open ourselves to opportunities. Pose estimation predicts the body part or joint positions of a person from an image or a video. It performs a complete biomechanical analysis of the human body in real-time; hence, it can benefit a vast number of applications. Sports analysis, Video surveillance and Assisted Living, Advanced Driver Assistance Systems (ADAS) are examples of implementations where Pose Estimation can make a difference [3].

Human Pose Estimation has been implemented using several approaches over the years. Initially, only the pose of a single person from an image was estimated, and this was usually done by identifying individual parts then linking these together to form a connection between the elements to create a pose [4]. These methods were not successful in detecting multiple people in an image and then estimating multiple people's pose simultaneously.

Even after many years of research in this field, human pose estimation remains a problematic and mostly unsolved problem with significant challenges such as ambiguity in human appearances and physique, obscurity in lighting conditions, occlusions due to various other objects in the scene, the complexity of the human structure, high dimensionality associated with the poses, 3D information loss from observing the posture from 2D planar images, etc. [5].

This paper presents the above-mentioned issues and handles the scalability and deployability of Human Pose Estimation. The paper also discusses the available datasets and required methodology to carry out multi-person pose estimation.

## II. LITERATURE REVIEW

Toshev et al., presented a method for estimating Human pose using Deep Neural Networks (DNNs), the formulation of their approach was based on a regression problem on the body joints of a human being. They carried out a delineated empirical analysis and procedure on the eclectic real-world images obtained from two large datasets, namely FLIC (Frames Labelled In Cinema) and LSP (Leeds Sports Pose) Dataset. A comparative study of different approaches to the metrics was presented by them and further demonstrated that a generic convolutional neural network applies to various tasks of localization. However, the primary purpose of CNN(Convolutional Neural Network) was classification. They presented a review of various approaches and applications of Human pose estimation [6].

Jo˜ao Carreira et al., proposed a method that incorporated both input and output by using top-down feedback. This feedback network's main aim is to increase the performance and accuracy of the Convolutional Networks, which are hierarchical feature extractors that have already yielded impressive results on various classification tasks based on feed forwarding. They tested their methods on the two most challenging datasets MPII Human Pose dataset and LSP (Leeds Sports Pose) dataset. They obtained promising results using the Iterative Error Feedback process, which progressively develops an output by feeding back the error predictions rather than directly predicting outputs [7].

Xiao Bin et. al., present the insights of the major work done on human pose estimation. They have compiled and documented all the previous work done in this field by comparing various methods and related work on datasets such as MPII Human Pose Dataset, COCO (Common Objects in Context) Dataset. They accumulated the results of The Pose Track Challenge of Multi-Person Pose Tracking event and compared the various methods used, such as ProTracker, PoseFlow, MVIG, BUTD2, SOPT-PT, and ML-LAB. Baselines on how to approach

Human Pose Estimation and work related to representation and tracking Human Poses considering various factors and metrics were discussed in detail in this paper [8].

Ke Sun et al., presented a methodology to represent high resolution learning of the human pose estimation in the COCO Dataset and the MPII Human pose Dataset. They formed the four stages and parallelled high-resolution networks by augmenting their rudimentary high-resolution subnetworks. This method helped maintain a high-resolution representation network throughout the process, rather than the conventional way of recovering high-resolution output from low-resolution representation. The performance and accuracy had increased, and then they demonstrated the superiority of their model in pose tracking competition on the PoseTrack dataset challenge. They were able to successfully obtain high-resolution representation for Human Pose Estimation [9].

## III. METHODOLOGY

Pose Estimation can be classified based on the number of tracked subjects, namely, Single Person Pose Estimation (SPPE) and Multi-Person Pose Estimation (MPPE). The latter consists of many edge cases and problems. In our paper, we will be focusing on the Multi-Person Pose Estimation Technique. Firstly, the most basic requirement of any model is data; this data can be of varied structure and form. The most typical data is RGB (Red-Green-Blue) images, which are highly modular and have an immense advantage over others in terms of the mobility of the input source. Others include Depth images, which represent the value of a pixel to the distance from the camera and Infra-red pictures in which the value of the pixel is determined by the amount of infrared light reflected to the camera. The most crucial input element is video. A video is nothing but an extensive array of images, where two consecutive frames share a massive portion of the information present in them. The estimated poses should be analogous across successive frames of video, and the algorithms should be computationally efficient to handle a large number of frames. The major problem of obstruction is eradicated to some extent in videos. The problem is further classified as 2D Pose Estimation and 3D Pose Estimation based on the output dimension requirement. The former is the prediction of body joints in the image, whereas the latter requires the projection of the three-dimensional spatial arrangement of all body joints. The most effective and efficient way is to predict the 2D pose first and then lift it to a 3D pose.

The Pose Estimation process comprises many crucial steps; they are as follows -

### 3.1 Body Model

Every pose estimation algorithm universally accepts a standard body pose model, which helps in normalizing the body model parameters. A structure where joints are represented in the form of vertices and edges can be encoded about the plane is implemented, such structure is termed as a simple N-joint rigid kinematics skeleton. This type of model is sufficient for most of the applications. However, some techniques accompanying complex mesh models are also used for character animation.

### 3.2 Pose Estimation Pipeline

The Pose Estimation Pipeline consists of four significant steps viz. Pre-processing, Feature Extraction, Inference, and Post-processing. The pipeline ensures the efficient processing of the images, filtering of images, noise removal from the data, etc. It helps in carrying out the task smoothly; the parts of the pipeline are as follows:

#### 3.2.1 Pre-processing

Background removal: Segmentation of humans from the background is necessary to remove the unwanted noise present in the frame. Bounding box creation: It is necessary to create individual boxes for multiple humans present in the frame for MPPE so that each box is individually evaluated for estimation.

#### 3.2.2 Feature Extraction

The process of feature extraction is very vital in pose estimation. It creates values from the raw data that can be used to train our machine learning model. The accuracy of these features will result in the precise estimation of human pose. There may be specific features that include conventional computer-vision based attributes like Histogram of Scale Invariant Feature Transform (SIFT) and Oriented Gradients (HoG). All these features are explicitly estimated before feeding in the input to the algorithm. The implicit features refer to deep-learning based feature maps. These features are never expressly created but are a part of a complete pipeline trained end-to-end.

#### 3.2.3 Inference

The most general way of predicting the location of joints is by producing confidence maps. They are considered the probability distribution over the image, representing the joint position's confidence at a given pixel.

There are two basic approaches to the discovery of joints in the frame. The bottom-up approach involves the detection of the parts or fittings for one or more humans in the image, and then assembling the pieces and associating them with the subject. In short, the algorithm predicts all the body joints present in the frame first. Then after a graphical formulation according to the body model, a final representation is generated, which connects the joints belonging to the same human. The top-down approach involves the segmentation of the subjects at first, where each human is first segmented into a bounding box and performing pose estimation on each block individually. Finally, by fitting the body-model onto the image a human like prediction is made.

### 3.3 Post Processing

Plenty of algorithms do not have any relation constraints on the final output. In simple terms, these algorithms lack the filtering of the rejection or correction of the ambiguous human poses; this may lead to aberrant human pose estimation. To cope up with this, a set of postprocessing algorithms that rejects such unnatural human poses are applied. The output is scored on the likeliness of occurrence, poses that get scored lower than the specified threshold is ignored during the testing phase.

## IV. DATASETS

Data sets are one of the most integral parts of the projects on pose estimation. The success of the project is based on the veracity and integrity of the data.

COCO (Common Objects in Context) is a large-scale object detection, segmentation, and captioning dataset, created by Microsoft and the largest 2D pose estimation dataset with multi-person 2-Dimensional poses. This dataset is considered as a benchmark for testing pose estimation algorithms. This is a vast dataset, containing approximately 1.5 million object instances [10]. In addition to this variety, the images are classified efficiently, making it easier for the user to interpret the dataset. The dataset has 80 object categories, 91 stuff categories, and five captions per image. Lastly, it contains 250,000 people with crucial points, making the dataset an ideal candidate for pose estimation algorithms.

LSP (Leed Sports Pose) dataset is one of the essential datasets that can also be used for pose estimation. It consists of more than 2000 pose annotated images of sports activities acquired from the internet using various labels [11]. Subsequently, the images are scaled to the desired length, and each image has been annotated with 14 different joint locations. The complex nature of the dataset makes it very helpful for the estimation of a pose.

MPII Human Pose dataset comprises 20,000 images containing 40,000 people with labeled body joints; it is a well-articulated dataset used as a benchmark for human pose estimation. The dataset covers 410 human activities such as “Rock climbing,” “Washing windows,” or “Picking fruit,” etc. [12]. The images are extracted from YouTube videos and have a relevant activity label. Also, body part occlusions and the 3D torso and head orientation are richly annotated.

## V. APLLICATIONS OF POSE ESTIMATION

Research in the field of Human pose estimation has seen a lot of changes, as it has evolved from the handcrafted feature-based approaches [13], [14], [15], [16], [17] to deep learning paradigm. This evolution of pose estimation and detection has seen a rise in its benefits in our quotidian life [18]. The various applications in which Human Pose Estimation is prevalent are:

Human-Computer Interaction (HCI): HCI has been a successful technological advancement in computers, and it has achieved effective integration of human factors with software technology. The human visual gesture is a critical interface that has the principle of Human Pose Estimation. For example, recognizing manufacturing steps to aid workers in learning and improving their skills [19], or using hand gestures to control the presentation slides [20].

Annotation of Videos: In today's world, a large amount of data is created on a daily basis, such as surveillance videos, movies, sports videos. To analyse and retrieve information from these videos, Human motion analysis can be used rather than scanning manually through the videos e.g., such as annotating video recordings of a soccer game [21] or broadcasting any sporting event [22].

Movies and animation: With advancements in technology, it is now possible to capture human motion. For example, in the Avatar movie, this technology was used [23], [24].

Video surveillance: Security is an important aspect, and video surveillance is collected in many places. However, it is becoming difficult to monitor these video surveillances manually. Therefore, automatic video surveillance analysis, e.g., [25], [26], will be required.

Facilitated living: Assistance to the elderly, disabled, and ordinary people can be improved with the help of pose estimation and activity recognition. For example, fall detection for elderly or in hospitals [27].

Virtual Reality Games: Games in which the players use gestures and body movements to play the games. For example, an Interactive balloon game [28] or the well-known Microsoft Kinect Xbox [29].

Sports Trainer: Several sports such as gymnastics, football, soccer requires correct body posture and movement; hence performance analysis and training can be carried out by Human Pose Estimation. It is extremely dangerous to perform yoga exercises or gym training without proper guidance. Therefore, Human pose estimation can be used to provide personalized feedback on fitness exercise form or pose. For example, an application [30] can be developed to act as a virtual fitness trainer.

## VI. CONCLUSION AND FUTURE RESEARCH SCOPE OF WORK

In this paper we have presented basic methodology and various applications of Human Pose Estimation because most of the issues addressed in the literature available on pose estimation is reasonably advanced, making it difficult for a new person to get acclimated to the topic. This technology will have considerable implications in shaping and providing a base to many industries such as Augmented Reality/Virtual reality, Healthcare Industry, Sports/Fitness Industry, and many more.

estimating pose of a moving person, evaluating the posture in cases of occlusions by objects, scanning a human pose beyond an opaque object, for example, behind a wall/gate, increasing the resolution, feature extraction from human poses and applications to other dense prediction tasks are the recent challenges which are yet to be addressed. For example, simultaneous object and pose detection, burglary identification, predicting a person drowning in a pool depends on his posture while swimming. One of the main areas where pose estimation can be instrumental is in the sports industry for training and scoring purposes for example, in gymnastics, boxing and various other sports events virtual judges could score the players based on the accuracy of their posture, movement, gestures, and various other factors.

## VII. ACKNOWLEDGMENT

Authors would like to express their sincere thanks to M. Kiruthika Associate Professor, Fr C Rodrigues Institute of Technology, Vashi for her constant support, guidance and encouragement. We also thank Dr. Lata Ragha, HOD Computer Engineering Department, Fr. C. Rodrigues Institute of Technology for providing the facilities.

## REFERENCES


**[1]** Kang, Nahua, and Issac Wu. "2 Fit-Tech Trends Are Leading Us towards Interactive AI Fitness Trainers." Revue, July 25, 2019. http://www.embodiedai.co/issues/2-fit-tech-trends-are-leading-us-towards-interactive-ai-fitness-trainers-182946.

**[2]** Babu, Sudharshan Chandra. "A 2019 Guide to Human Pose Estimation with Deep Learning." AI & Machine Learning Blog. AI & Machine Learning Blog, August 5, 2019. https://nanonets.com/blog/human-pose-estimation-2d-guide/.

**[3]** Ganesh, Prakhar. "Human Pose Estimation: Simplified." Medium. Towards Data Science, December 9, 2019. https://towardsdatascience.com/human-pose-estimation-simplified-6cfd88542ab3.

**[4]** Raj, Bharath. "An Overview of Human Pose Estimation with Deep Learning." Medium. BeyondMinds, May 1, 2019. https://medium.com/beyondminds/an-overview-of-human-pose-estimation-with-deep-learning-d49eb656739b.

**[5]** Sigal, Leonid. "Human Pose Estimation." *Computer Vision, A Reference Guide* (2014).

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

**[7]** J. Carreira, P. Agrawal, K. Fragkiadaki and J. Malik, "Human Pose Estimation with Iterative Error Feedback," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 4733-4742, doi: 10.1109/CVPR.2016.512.

**[8]** Xiao, Bin, Haiping Wu and Yichen Wei. "Simple Baselines for Human Pose Estimation and Tracking." *ArXiv* abs/1804.06208 (2018): n. pag.

**[9]** ke, Sun & Xiao, Bin & Liu, Dong & Wang, Jingdong. (2019). Deep High-Resolution Representation Learning for Human Pose Estimation.

**[10]** Lin, Tsung-Yi, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár and C. Lawrence Zitnick. "Microsoft COCO: Common Objects in Context." *ArXiv* abs/1405.0312 (2014): n. pag.

**[11]** Johnson, Sam and Mark Everingham. "Learning effective human pose estimation from inaccurate annotation." *CVPR 2011* (2011): 1465-1472.

**[12]** Andriluka, Mykhaylo, Leonid Pishchulin, Peter V. Gehler and Bernt Schiele. "2D Human Pose Estimation: New Benchmark and State of the Art Analysis." *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014): 3686-3693.

**[13]** M. Andriluka, S. Roth, and B. Schiele, ''Pictorial structures revisited: People detection and articulated pose estimation,'' in Proc. CVPR, Jun. 2009, pp. 1014–1021.

**[14]** M. Andriluka, S. Roth, and B. Schiele, ''Monocular 3D pose estimation and tracking by detection,'' in Proc. CVPR, Jun. 2010, pp. 623–630.

**[15]** X. Chen and A. L. Yuille, ''Articulated pose estimation by a graphical model with image dependent pairwise relations,'' in Proc. NIPS, 2014, pp. 1736–1744.

**[16]** L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, ''Poselet conditioned pictorial structures,'' in Proc. CVPR, Jun. 2013, pp. 588–595.

**[17]** Y. Yang and D. Ramanan, ''Articulated human detection with flexible mixtures of parts,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

**[18]** F. Zhang, X. Zhu and M. Ye, "Efficient Human Pose Estimation in Hierarchical Context," in IEEE Access, vol. 7, pp. 29365-29373, 2019, doi: 10.1109/ACCESS.2019.2902330.

**[19]** A. Postawa, M. Kleinsorge, J. Krueger, and G. Seliger, "Automated image based recognition of manual work steps in the remanufacturing of alternators," *Adv. Sustain. Manuf.*, vol. 5, pp. 209–214, 2011.

**[20]** H. Lee and J. H. Kim, "An HMM-based threshold model approach for gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 10, pp. 961–973, 1999.

**[21]** J. Assfalg, M. Bertini, C. Colombo, A. Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: Automatic highlights identification," *Comput. Vis. Image Understand.*, vol. 92, no. 2–3, pp. 285–305, 2003.

**[22]** J. Kilner, J.-Y. Guillemaut, and A. Hilton, "3D action matching with key-pose detection," in *Proc. Int. Conf. Comput. Vis. Workshops*, 2009, pp. 1–8.

**[23]** J. Geigel and M. Schweppe, "Motion capture for realtime control of virtual actors in live, distributed, theatrical performances," in *Proc.FG'11*, 2011, pp. 774–779.

**[24]** A. Alatan, Y. Yemez, U. Güdükbay, X. Zabulis, K. Müller, C. Erdem, C. Weigel, and A. Smolic, "Scene representation technologies for 3dtv—A survey," *EEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1587–1605, Nov. 2007.

**[25]** S. Park and M. Trivedi, "Understanding human interactions with track and body synergies (TBS) captured from multiple views," *Comput. Vis. Image Understand.*, vol. 111, no. 1, pp. 2–20, 2008.

**[26]** A. Utasi and C. Benedek, "A 3-D marked point process model for multi-view people detection," in *Proc. Computer Comput. Vis. Pattern Recognit.*, 2011, pp. 3385–3392.

**[27]** C. Rougier, J.Meunier, A. St-Arnaud, and J. Rousseau, "Fall detection from human shape and motion history using video surveillance," in *Proc. 21st Int. Conf. Adv. Inf. Netw. Applicat. Workshops*, 2007.

**[28]** C. Tran and M. Trivedi, "Introducing XMOB: Extremity movement observation framework for upper body pose tracking in 3D," in *Proc. IEEE Int. Symp. Multimedia*, 2009, pp. 446–447.

**[29]** J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. Computer Comput. Vis. Pattern Recognit.*, 2011.

**[30]** Chen, Steven & Yang, Richard. (2018). Pose Trainer: Correcting Exercise Posture using Pose Estimation.