International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 10, No.1 (Aug-2021)

http://dx.doi.org/10.12785/ijcds/100181

Review on Human Pose Estimation & Human Body Joints Localization

Miral M Desai[1] and Hiren K Mewada[2]

[1] *Department of EC Engineering, CSPIT, CHARUSAT, Charotar University of Science & Technology (CHARUSAT), CHARUSAT campus, Changa 388421 Anand, India*
[2]*Electrical Engineering Department, College of Engineering, Prince Mohammad Bin Fahd University, AL Khobar, Saudi Arabia*

Received 08 Jan. 2021, Revised 11 Apr. 2021, Accepted 15 Apr. 2021, Published 5 Aug. 2021

Abstract: Human Pose Estimation (HPE) is a relatively new and significant computer vision field and its applications. HPE is the process of estimating the location of the human body joints from the image or video. The correct estimate of human body joints is used to track people's minimal activities in real-time applications. HPE is an extensive research area that relies on many individuals being monitored. HPE can be categorized in two ways, e.g. (a) based on the number of humans whose pose to be estimated, i.e. single-person or multi-person pose estimation and (b) based on the environment used, i.e. 2-dimension (2D) or 3-dimension (3D). Initially, this paper presents a traditional approach in brief and later paper focus on recent advancement in HPE using deep learning approaches. A rigorous review of deep learning approaches using both top-down and bottom-up approaches is expressed and compared using various evaluation matrices and models' accuracy. It is observed that most models succeed to perform well on MPII dataset in comparison to COCO dataset. The Distributed aware architecture gives the best performance among all models providing 97% Percentage of Correct Key (PCK) on the MPII dataset and 78.9% average precision (AP) on the COCO dataset. For multi-person HPE, AP is limited to 78.6% using the AlphaPose model.

Keywords: Human Pose Estimation, Oclussion, Clothing Variation, 2D Pose, Multi-person Pose Estimation

1. INTRODUCTION

Human pose estimation involves localizing the human body joint localization like elbow, wrist, heap, shoulders, knee and ankle in static image or video. It is also considered as the identification of a specific pose from all articulated poses. Pose estimation via body joint localization can be done either in 2-dimension pose estimation or 3-dimension pose estimation. The 2-dimension pose estimation estimates 2D pose as (x,y) coordinates for each joint from color image or video. The 3-dimension pose estimation estimates 3D pose as (x,y,z) coordinates for each joint from color image or video. HPE has many applications like action recognition, animation, gaming, human activity recognition, video surveillance, people tracking system, clothing parsing, human-object interaction etc [1].

As depicted in Figure 1. [(a) to (e)], image input data for HPE face a variety of difficulties, including occlusion, clothing variation, body part skipping, illumination variation, and background clutter. For video input data, blur motion is one of the known challenges. Occlusion is the condition where a person or multi-person interacting with one another or with surrounding elements. Body part foreshortening is when human body joints disappeared due to the projection of body parts being of higher angle concerning the image plane. Clothing variation is the challenge where human body joints covered because of fashionable clothing. In the illumination variation challenge, body joints do not appear clearly due to the effect of the image's background. In some cases, the environment is either so dark or very bright. Background clutter is when the background of the image appears in a wide range of color distribution.

(a) Occlusion [3]

(b) Clothing Variation [3]

(c) Body part skipping [2]

E-mail: miraldesai.ec@charusat.ac.in, hmewada@pmu.edu.sa

(d) Illumination variation [4]

(e) Background clutter [5]

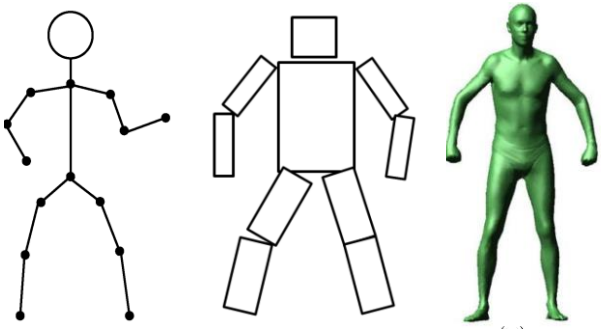
Figure. 1. Challenging environments to find the key points in human pose estimation

The second category for input source is depth images, in which the value of the pixel of the image is related to the distance measured from the camera. Depth images are also known as time-of-flight images. A separate database is required for input images in the depth image category. The popular low-cost device Microsoft Kinect made this thing easier and provide a large number of depth data [6]. The depth image processing requires complex models. The type of cameras for capturing depth images also not available readily in the market. Near-infrared images whose pixel intensity is determined by the amount of infrared light sensed by the camera can also be used to assess pose. This also requires a complex computer vision model. Microsoft Kinect provides IR images [6]; however, there is no specific IR images dataset to date.

The human pose estimation process is designed and estimated on static images as well as on video data. Supportable datasets for both types of input sources are readily available. In both cases, pose estimation can be estimated by drawing the skeleton on the human body. The first key points of body joints are required to draw the skeleton on the human pose. Identifying the correct key points of body joint is the human pose estimation in the true sense. Compared to the video data source, it will be easy to locate the right key points in the image data source. Video data source is nothing but a collection of images that is consecutive in two respective frames. Maintaining the right key points of body joints in successive frames is a major challenge in human pose estimation in video data sources. As a result, in the video database-based input source, it is assumed that the estimated human pose will remain consistent from frame to frame. The designed algorithm should also be computationally efficient for handling a large number of frames. In comparison to static images, the issue of occlusion would be easier to solve in the case of video source input due to the existence of past and future frames in which the body position is not obscured.

Every human pose estimation algorithm gives the output in three different forms, i.e. skeleton, shape or mesh-based pose representation, as shown in figure 2. These classes formulate the problem of HPE to estimate body model parameters. The most typical algorithm produces an n-point rigid Kinematic Skeleton representation as output. For the final result, n is usually ranging from 13 to 30. The Kinematic model is represented as a line or graph, with each vertex corresponding to a joint. In the earlier study, a shape was used as a body model in estimating the pose. In the shape model, human body parts are depicted using geometric forms such as rectangles and cylinders. One of the human pose estimation applications is character animation, in which a more elaborated body model is required. In this case, the m*esh-based model* representing the whole body in 3D with a point cloud is needed.

The HPE system uses the setup of single or multiple cameras. A standard HPE algorithm involves a single camera for human pose estimation. A specific algorithm involves multiple cameras to take multiple viewpoints and combine them to create an accurate human pose. The problem of occlusion is easily handled in multiple cameras. The research on multiple cameras is limited as the dataset is limited in this particular field.



(a) Kinematic Skeleton Model [7]

(b) Shape based body model [7]

(c) Mesh Based Model [7]

Figure. 2 Classes of pose representation

HPE approach is bifurcated into four different categories. Pre-processing, Feature extraction, Inference, and Post-processing are the steps in the HPE process, as shown in Figure 3. Pre-processing approach is further classified into two different categories: Background subtraction and bounding box generation. In the background subtraction step, a person(s) is segregated from the whole image by removing the extra noise or additional part of the image. A bounding box is drawn over the segmented human blob after it has been segmented from the image. For the multi-person pose estimation case, the bounding box is created separately for each individual in the image. In multiple cameras, the pre-processing approach is further extended into the camera calibration and image registration step. In this step, image registration is required because numerous inputs are collected from multiple cameras. Camera calibration is also essential in the 3D human pose estimation to get accurate coordinates of human body joints [8].

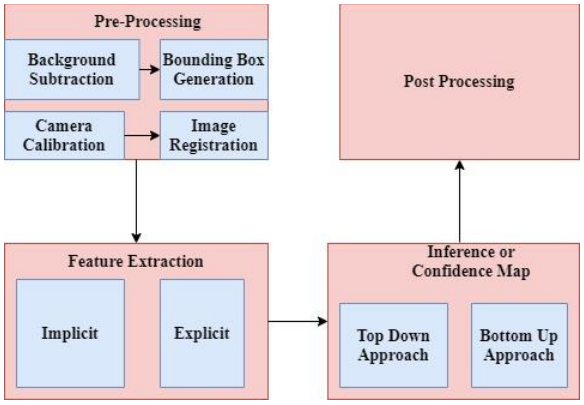


Figure. 3. Human Pose Estimation Approach

Feature extraction is one of the fundamental processes of machine learning and deep learning in which resulted image is generated from the raw input sources like image and video. This resulted image is used as an input parameter to learn a specific algorithm. The method of Feature extraction can be either explicit or implicit. Without providing any input to the learning algorithm, distinct features are created as a separate class. In traditional computer vision algorithms, scale-invariant feature transform and histogram of oriented gradients (HOG) can generate explicit features [9, 10]. Implicit features are never generated in advanced rather, they are generated as a part of a process. Implicit features are the part of a complex convolutional neural network (CNN) in the form of a feature map generated as an output of the network. The inference is the process of generating confidence maps for key points on the human body joints. The confidence map predicts the human body's key points or calculates the probability distribution over the human body image representing the joint's confidence located on each pixel of the key point. Confidence maps can be processed in two ways: a top-down approach and a bottom-up approach. The bottom-up approach is the process where human body joints are first detected. After detecting the body joints, all the body joints are assembled. The assembled parts are then associating with the respective human body image. Part confidence map and Part affinity field techniques are used to detect the body joints [11]. In the top-down approach, each human body is first segregated from the whole image. Then pose estimation algorithm is applied to individual human body bounding box for estimating the body joints. The top-down approach can be either the generative body model or the deep learning model. The Generative body model tries to match the body model on the image where it assumes that the final prediction is as Human body. The deep learning model tries to predict the body joints as a part of the network. So it might happen that the final output is not as the predicted human body. The bottom-up and top-down approach does not give the final output as the human body every time. Sometimes it requires to reject or correct the outcome of both approaches. Post-processing is the approach where the output as the unnatural human pose is to be rejected. The post-processing algorithm decides the correct human pose by observing some thresholds of the body joints' key point. The post-processing algorithm is most required and affected for the complicated poses [12].

HPE is one of the vital field of computer vision as well as Machine learning and deep learning. There is numerous application of pose estimation available. Some of the applications are listed below:

A. Human Activity Recognition

Detection and tracking of key points or human body joints provide the human pose. This perfect pose is used to track the daily activities of human. Gesture recognition is also one of the important branches of human activity recognition. Human activity recognition is used to track if the person falls or lying on the plane. Human body language recognition can also be done with the help of activity recognition. Sport and dance activities can be tracked using pose estimation. Human activity recognition can be used for security and surveillance [13].

B. Motion Capture

Human pose estimation tries to give the correct key point of body joints. These key points of body joints are used to capture the motion of human activities. Augmented reality and computer graphics technology can add to the stylish artwork once the human pose is estimated. This technology can be useful for animation [14].

C. Robotics

The human pose estimation technique can improve the manual design of the humanoid robot. Instead of designing manual body parts and joints, corrected key points can help create the humanoid robot. We can also train the humanoid robot's key points, which is used to teach specific action to the humanoid robot [15].

D. Gaming

Human pose estimation is widely used for mobile and desktop gaming. As the correct key points of the body joints are one of the focused areas of human pose estimation, it can easily control the person in games. The Homecourt mobile application designed by Nex Team, San Jose, California, is the basketball playing game in which the tracking the activities of the player according to the basketball movement is design using a HPE [16, 17].

This paper aims to provide a rigorous review of HPE methods. A brief review of traditional methods is

expressed. Then, advancement in HPE using convolution neural network (CNN) is presented in details. Evaluation parameters and datasets play a vital role in the comparison of models. Therefore, all parameters and datasets are introduced. Later various CNN based HPE are explained and compared using these parameters. According to the authors' knowledge, this is the first paper introducing a comprehensive analysis of all HPE methods and the comparative evaluation of parameters among various models and various datasets.

2. EVALUATION OF MATRICES

The human pose estimation method's performance is measured using evaluation matrices that are dependent on various parameters depending on the dataset. Some of them are based on human body size, the upper part of the body, or the full-body part estimation. Some are based on an analysis of a single person or multi-person pose estimation. Commonly used Evaluation Parameters are as follows:

A. Percentage of Correct Parts (PCP)

PCP is a tool for estimating human pose in two dimensions. PCP reports the accuracy of limb localization. If two limb endpoints are within the threshold of corresponding ground truth endpoints, PCP implies the limb is perfectly localized. The 50 % limb length threshold is used here.

B. Percentage of Correct Key-Points (PCK)

PCK is used to measure the accuracy of body-joint localization of 2D human pose estimation. PCK is considered as corrected if key points are in the range of the threshold pixel of ground truth joints. The threshold is defined as the fraction of the pose bounding box where the mean distance between predicted and true joints should be less than the torso height radius. Higher the percentage of correct key points seems the model is accurate.

C. Average Precision (AP)

Assume that given a set of samples, where each sample is the set (ci, gi) where $ci \in R$ is the confidence score and $gi \in 0, 1$ is the ground truth, the AP is computed way. First, all the sets are ordered based on the confidence score, starting from the highest to the lowest. At each sample *si* in this ordered list, precision *pi* and recall *ri* are computed on the first *i* samples. These precision and recall values are plotted with recall as X-axis and precision as Y-axis to obtain a precision-recall (PR) curve. The higher the value of AP (i.e. area under the PR curve), the algorithm performs better.

D. Average Recall (AR)

Evaluation of multi-person pose estimation is as equal as an extension of object detection problem. Object detection problem considers similarity measurement. In the multi-person pose estimation, object key point similarity (OKS) is the same evaluation parameter for similarity. Average Recall is reported as a similarity of a key point in the COCO dataset.

3. DATASETS

Datasets are one of the important parameters for human pose estimation algorithms. Datasets give the source of inputs. A brief note on some of the common dataset is given as follows:

A. MPII

MPII is the abbreviation of Max Planck Institute Informatik. It is a state-of-the-art benchmark for evaluating coherent HPE. It consists of 2D pose estimation for multi-person, which inculcates around 25K images in which there are about 40K people who are having body joints annotated on it. It incorporates 410 human activities, which are performed on a regular basis with corresponding labels. All images in the dataset were extracted from youtube video with un-annotated frames. It achieves excellent results on its 3D pose test set, especially for orientation of head and torso, and also on occluded image part proper key point annotation has been achieved [18].

B. COCO

COCO is the abbreviation of **C**ommon **O**bjects in **Co**ntext. This 2D pose estimation dataset was collaboratively prepared by Lin et al. from Google Brain, MSR, Caltech, TTI-Chicago, WaveOne, Cornell Tech, Facebook AI Research, Georgia Tech and CMU fellows by collecting data from Flickr. It is highly concentrated on object detection, captioning and segmentation-based data features. It has some enticing features such as context recognition, super pixel stuff and object segmentation. It consists of 300K images in which more than 200K are labelled images. It has 80 object categories and 91 stuff categories. More exciting features are that each image has five captions, and 250000 people images are annotated with key points. It shows up its efficient Dense Pose, Semantic segmentation, panoptic segmentation, Detection, keypoints annotation and image captioning task [19].

C. FLIC

FLIC is the abbreviation of **F**rames **L**abelled **I**n **C**inema. It consists of 5003 images which are taken from Hollywood movie scenes. It was collected using canonical person detector on each consecutive tenth frame of 30 movies. Overall 20K people were taken under consideration for confidence maps which were sent to highly populated Amazon Turk to obtain ground truth

labelling. With the help of turkers image upper body joint annotations were made and five labelling points' median was taken under consideration for outlier annotation. Test data consists of 1016 data images. It also has a FLIC-full dataset with ample frames set from amass movies whose hand joint annotations were made by Mechanical Turk [20].

D. LSP

LSP is the abbreviation of **the L**eads **S**ports **P**ose dataset. It consists of 2000 annotated pose images related to eight different sports activities from Flickr. All images are scaled down to approximately 150 pixels in length. All the sports images can detect 14 key point joint locations. Training and testing data are divided into a ratio of 50% [21].

E. HumanEva

Max Planck Institute prepared this dataset for intelligent systems Perceiving Systems. It consists of two datasets: (a) HumanEva-I: This dataset is synchronized using software by keeping seven video cameras at a time. Video cameras were of two types: three-color and four grayscale cameras, in which six cameras were motion capture cameras. It has training, validation and testing datasets. (b) HumanEva-II: This dataset is synchronized using hardware by keeping four video cameras at a time. There were four-color video cameras in which eight cameras were motion capture cameras. It has only a testing dataset.

It is a 3D pose estimation for single-person, consisting of video sequences discussed in two types of datasets. Marker-based motion capture cameras are used to prepare 3D poses ground truth images. It consists of 4 subjects who are performing six common actions. Error matrices in 2D and 3D pose are made available [22].

F. Human3.6M

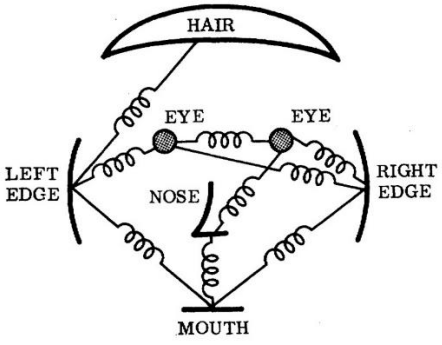
This dataset contains 3.6 million images of human pose images in 3D. It was created with eleven lionized and professional actors in which there are six male and five females on 17 various activities. Video frames at a frequency of 50Hz and high resolution were captured using highly calibrated and time of flight based four cameras. 24 body part labels for each configuration and pixel-level are measured. Actors' 3D laser scans are utilized. It has proper background subtraction and bounding box on the person.

4. DIFFERENT HPE APPROCHES

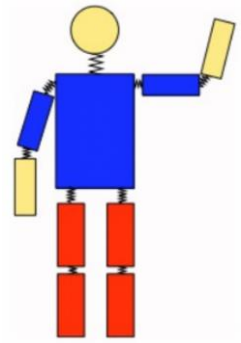
The main objective of HPE is to estimate the correct key points of the body joints or body parts. The right key points of body joints or body parts can be estimated with different approaches. These approaches contain the traditional approach and deep neural network based approaches. The traditional approach includes a pictorial structure framework and deformable parts model-based approach. Deep Neural Network approach incorporates DeepPose, OpenPose, AlphaPose, ConvNets, DeepCut, HRNet etc. A brief discussion of all mentioned algorithm is given as follows:

A. Traditional Approach for HPE

The idea behind the Traditional Approach is shown in Fig. 4. Fig. 4(a) represents an object by collecting the body parts, and the arrangement of these body parts in the deformable structure is shown in fig 4(b). In the Pictorial structure framework, the object represents detectable body parts estimated from the person's image. All estimated body parts are connected with the spring. The spring indicates the spatial connection between two body parts. Here body parts are evaluated according to the pixel location as well as orientation. The resulting articulation model looks like a human body. The pictorial framework structure has a limitation that the approach is not suitable for specific image datasets. Yang and Ramanan introduced the deformable part model. The Yang and Ramnan model represents human body parts as a complex joint relationship [23]. Here each body part is represented in the form of matched templates. The matched templates are a combination of global templates as well as part templates. All these templates are arranged in the form of Human pose estimation.



(a) Pictorial Structure Framework



(b) Deformable Part Model

Figure. 4. Traditional Approach for HPE [24]

B. HPE using Deep Neural Network

The traditional approach for HPE is limited in the form of a good dataset and detection accuracy for correct key points of body joints. The accuracy of correct key points of body joints can significantly improve with the deep learning approach. The deep learning approach is based on the convolutional neural network. The deep neural network is separately designed for a single person and multiple person poses estimation. Recent papers are summarized as follows:

DeepPose: Deep Neural Networks for HPE [25]: Toshev and Szegedy presented DeepPose as an applied deep learning approach to estimate the key points of body joints. A cascaded CNN based regression approach was used to retune the key pints of body joints. CNN based regression model was effectively used as it can also estimate certain hidden body joints.

Functionality of Model: The DeepPose model works in two stages, and HPE was formulated in the DNN regression model. One of this model's essential key features is that it refines key points using a cascaded regression and feedback system. In the first stage of the model (i.e. Fig. 5), the specific key points are estimated as coarse tuning. The key point estimated image is cropped around the predicted joint and given as input of the next stage in Fig. 6, which is used to fine-tune the estimated pose [25].

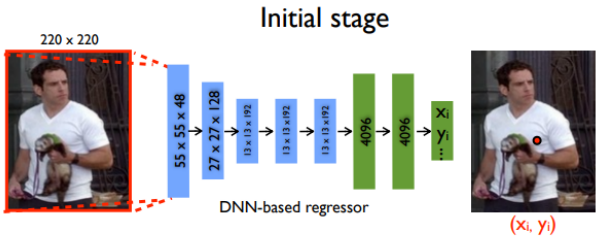


Figure. 5. First Stage Regression Model using CNN [25]

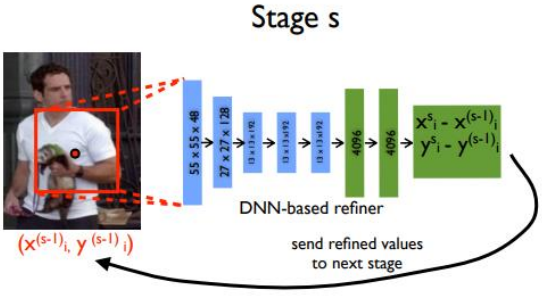


Figure. 6. Second Stage Regression Model using CNN [25]

Comment: 3-Stage DeepPose gives 56% and 38% correct parts for the upper and lower arm, respectively. Maximum PCP is achieved on the upper and lower leg is up to 78% and & 71%, respectively. The average percentage of the correct part is up to 61%.

Limitation: As regression of body joint location increasing complexity, it gives weak performance on specific region.

Efficient Object Localization Using Convolutional Networks [26]: In this approach, the author presented heat map regression in comparison with the previous direct regression model. The heat map is a discrete output than the continuous regression at a single point. The image is run through several resolution points in a parallel method to produce the heat map. This heat map predicts the likelihood of joints at each pixel.

Functionality of Model: Heat map regression with sliding window and Heat map regression model is shown in Fig. 7 and Fig. 8, respectively. This model is also known as the multiresolution CNN, where it uses a sliding window to produce coarse heat map output. In this approach, additional CNN is used to refine the localization points estimated by coarse heat-map. The simple cascaded model generates a coarse heat-map. The trainable parameters are fine-tuned by feeding the coarse heat map cropped output as input to the extended network. The spatial accuracy of the body joints for pose estimation is improved by this model [26].

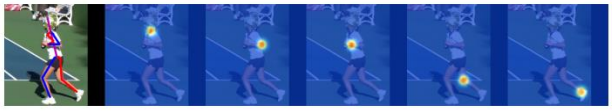


Figure. 7. Heat map Regression Sliding Window [26]

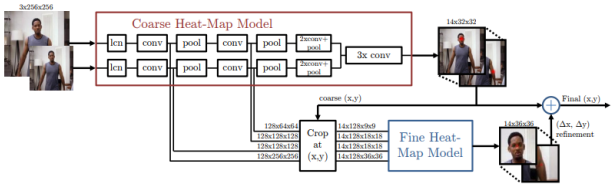


Figure. 8. Heat map Regression Model [26]

Comment: As the name suggests, this model is efficient compared to DeePose, because for specific body joint detection, heatmap works better than body-joint regression.
Limitation: Model fails to localize occluded body joints and body part skipping scenario.

Convolutional Pose Machines [27]: Wei et al. invented this Convolutional Pose Machine, also known as Pose Machine. A pose machine is the multistage CNN used to compute the respective body joints' heat map. In the first stage of the model, the pose machine predicts the heat map and give appropriate output accordingly. The first stage's output is fed into the second stage, along with the input image, and it generates another heat map. The second stage heat map is more clear output as compare to the first stage output. Usually, three stages are required for overall corrected key points of body joints.
Functionality of Model: Fig. 9 gives the outline, and Fig. 10 presents the architecture of the pose machine. As shown in Fig. 9, for the first stage, *x* is the input image, *g1()* is the predicted heat map, and it will generate a resultant heat map as *b1*. This resultant heat map *b1* is the input for the next stage along with the input image, and it will create the second stage predicted heat map as *g2()*. As shown in Fig. 10, the right knee joint of a sportsman person is predicted in stage 1, and it is refined in stage 2 of the pose machine. In this algorithm, intermediate monitoring is performed after each stage using the prediction process, avoiding the issue of vanishing gradients [27].

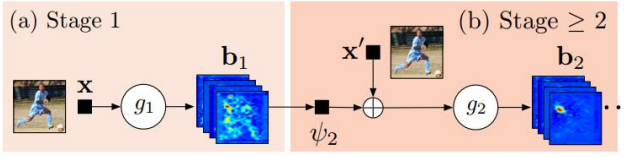


Figure. 9. Outline of Convolutional Pose Machine [27]

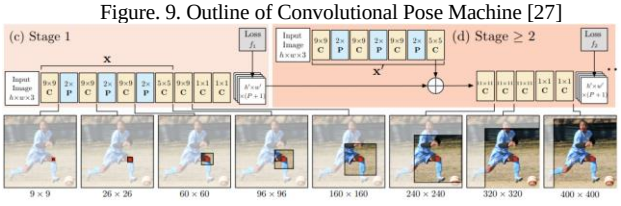


Figure. 10. Architecture of Convolutional Pose Machine [27]

Comment: This model suggests a novel approach as Convolutional Pose machine along with FLIC, MPII and LSP Dataset.
Limitation: Model fails to recognize body joints of multiple persons in the scenario when they are localizing close to each other.

HPE with Iterative Error Feedback [28]: Carreira et al. presented an iterative error feedback model using Convolutional Networks (ConvNets). ConvNets has been used for sequentially extracting its features using feed-forward processing for the variegated classification task. A multi-layered hierarchy is used to represent images via ConvNets. Provision of a generic framework for modelling structure 2D for input and output has been made. A top-down approach to feedback is used. The network would not directly predict the result but would firstly estimate the error, and through iterations, it would retrain it until the expected threshold value. This model has made 17 key point annotations.
Functionality of Model: The iterative feedback model and its mechanism are shown in Fig. 11. When the image *I* is feed into the network, it would estimate some key points y0 in the 2D points set. The model works on several equations as per the diagram in which function f is treated as ConvNet. Function g would convert each key point position into one gaussian heat map. This model can learn over body joint space configurations and images.

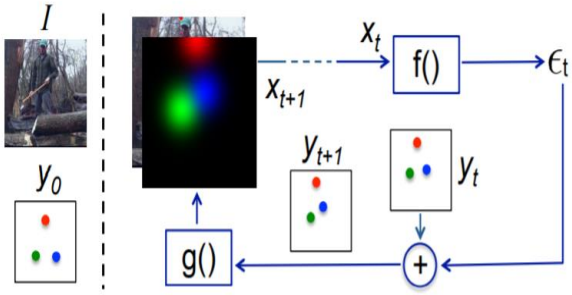


Figure. 11. Iterative Error Feedback Model [28]

ConvNet architecture is pretrained on ImageNet. Filters were modified for 17 key points in the first convolution layer. Conv-1 layer has 20 different input channels in which three are of ImageNet with pre-trained weights while the remaining 17 channels were initialized by gaussian noise with variance (0.1. The model was trained on both occluded and visible images. For obtaining occluded points, a back-propagating gradient was made zero. Then, by ConvNet and Gaussian filters, those points are made visible [28].
Comment: IEF achieves maximum accuracy in head detection of 95.5 and shoulder detection of 91.6, and prediction of an upper and full body is 81.9 and 81.3, respectively.
Limitation: In the feedback session of model, it just feeding back the images with the Gaussian distribution up to first layer only. Model can give far better performance, if sophisticated feedback block like Deconvolution is used. Sophisticated feedback strategy can give better performance for 3D human pose.

HPE using Stack Hourglass Networks [29]: Newell et al. extended stack hourglass network using skip connection. Single hourglass network is consecutively

placed by multiple hourglass networks together end-to-end. The top-down and bottom-up approach is simultaneously applied. This module processes spatial information at various scales for dense prediction before stacking to fully connected ConvNet and other designed layers.

Functionality of Model: As shown in Fig. 12, convolution and max-pooling layers are used in a single pipeline with skip layers to retain spatial information at all resolutions. When features reach the lowest resolution, the network starts a top-down approach, a sequence of up-sampling and different combinations of features at various scales. Network performance has increased when standard convolutional layers applied with large filters. As shown in Fig. 13, each box is working as the residual module. Each hourglass stage output is supervised by using the intermediate supervision method [29].

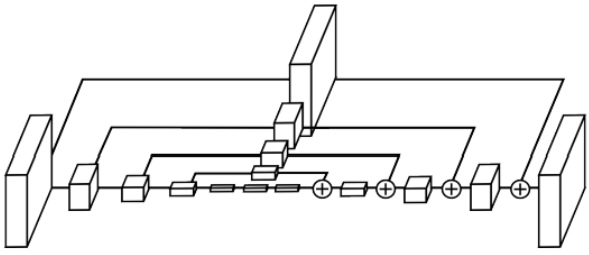


Figure. 12. Stacked Hourglass Module [29]

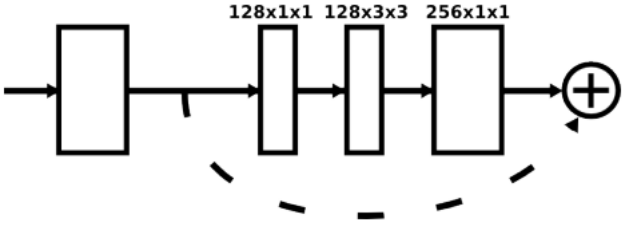


Figure. 13. Intermediate Supervision – Residual Module [29]

Comment: This model has a maximum prediction rate of 98.2% and 96.3% for head and shoulder respectively and minimum for ankle, i.e. 83.6% on MP-II Human Pose. On the FLIC dataset, 99% accuracy on the elbow and 97% for the wrist has been obtained.
Limitation: Annotations of body joints for multiple person is out of scope for the model. In case of multiple person, model fix only single person's body joint annotations. So it gives overlap body joints scenario for multiple person images.

Simple Baselines for HPE and Tracking [30]: Xiao et al. used a ResNet model with deconvolutional layers at the endpoint. The loss between expected and targeted heatmaps is estimated using the mean squared error method. The 2D Gaussian filters generate targeted heatmaps.

*Functionality of Model***:** The suggested network takes the reference of hourglass and cascaded pyramidal network. As shown in Fig. 14, the suggested network is very simple for design as compared to stacked hourglass network. Unlike stacked hourglass network, simple baseline network puts up sampling and convolution parameters in a single block. This combination works as the deconvolution layer for the suggested architecture. [30].

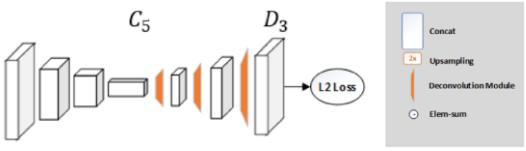


Figure. 14. Simple Baseline Network for Pose Estimation [30]

Comment: AP on COCO test-dev by training with ResNet-152 architecture is 73.7 and AP_{50} is 91.9.
Limitation: The performance of AP with respect to image size for Simple base line architecture is low as compare to HRNet architecture.

HPE using Deep High-Resolution Representation Learning [31]: Sun et al. proposed a top-down method based on a fine-to-coarse resolution network with one-by-one steps in the form of parallel stages. As with the stacked hourglass network, the proposed approach does not use intermediate heat map supervision. The Mean Square Error loss approach is used to evaluate heatmaps. *Functionality of Model*: As shown in Fig. 15**,** the suggested network is a combination of fine-to-coarse resolution sub-networks in parallel form to achieve more précised spatial heatmaps. Deep High-resolution [HrNet] model gives excellent performance in multi person pose estimation scenario. The approach of the other models are as they follows first high resolution, then low resolution and high resolution, HrNet maintain high resolution throughout the process. The model begins with a fine-resolution subnetwork and gradually introduces fine- to coarse resolution subnetworks to link multi-resolution subnetworks in parallel. [31].

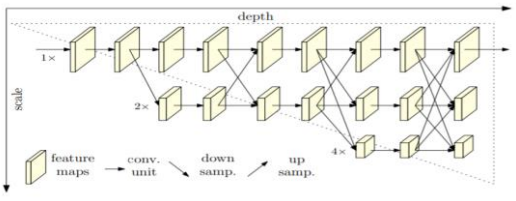


Figure. 15. High Resolution Network (HrNet) for Pose Estimation [31]

Comment: AP by training with HRNet-W48 architecture is 75.5% and AP_{50} is 92.5%. AP by training with HRNet-W48 + extra data architecture is 77% and AP_{50} is 92.7%.
Limitation: Computational cost is high.

OpenPose: Real-time Multi-Person 2D HPE using Part Affinity Fields [32]: Cao et al. used part affinity field to obtain multi-person pose estimation in real-time by training the network with human body parts on every frame.

Functionality of Model: This model works on an affinity-based approach, as shown in Fig. 16. Every part is applied affinity model, and confidence maps approach. In this network, the receptive field is preserved, and computation is reduced. Every pixel of an image is applied filter of part affinity to get proper key point directions [32].

Figure. 16. Multi person HPE using part affinity fields [32]

Comment: Results of AP on MP-II is getting maximum on head which is 91.2 and minimum on ankle which is 61.7.
Limitation: Model shows failure cases for complex poses, missing body parts, overlapping parts and false positive scenario like statue.

HPE using DeepCut [33]: Pishchulin et al., articulated pose estimation approach on a multiple scale by integrating subset partition and labeling approach in ConvNet. It is able to justify occluded images. CNN based part detectors are used for partitioning and labelling formation of set of body parts.

Functionality of Model: The formulation suggests joint subset partitioning and labelling problem [33]. The functionality of the model is classified in three different approaches. In the 1st approach the model produced the set of body parts. This set represents the body joint locations of each person of the image. The 2nd approach labialized body parts in to specific class. The 3rd approach generate set of body parts belongs to the same person. Fast R-CNN and Dense CNN methods are used to detect body parts. The procedure of identifying key point for occluded images using DeepCut algorithm is shown in Fig.17.

Figure. 17. DeepCut key point procedure [33]

Comment: Results of Percentage of Correct Parts (PCP) for LSP dataset is 96.0 which is maximum and 64.2 which is minimum.
Limitation: The clustering problem of Deepcut resolves with Integer Linear Programming (ILP) in which labializing of each joint represent to single class only. With reference to the same it might happen that more than one joints labialized by the single class and it's create false prediction of body joints.

Regional Multi-Person Pose Estimation (RMPE) [34]: Fang et al. presented top-down approach for multi person HPE. In this approach single-person pose estimator is used to extract the person in form of Bounding Box. When pose estimation is performed on the region where person is located, it generate errors like duplication in bounding box. To resolve this problem, author suggests Symmetrical Spatial Transformer Network (STTM) which is used to extract high-quality single person pose from improper bounded-box. The Spatial De-Transformer Network (SDTN) remap the estimated human pose according to original coordinates

Functionality of Model: This method uses VGG-based SSD-512 for human blob detection because it is efficient for recognition-based approach. Stack hourglass network is also used for single person pose estimation. For STN, ResNet-18 has been used. ResNet – 152 based Faster RCNN was used for pose estimation Replacement of pose network was done with PyraNet [34].

Figure. 18. RMPE key point procedure [34]

Comment: The average accuracy on MP-II was 72 mAP. Maximum accuracy has been observed when a network uses ResNet based on Faster RCNN with 91.3 for head and lowest for wrist 76.4.
Limitation: Here the process of human detection is to be done by Single Person Pose Estimator (SPPE) which is not the part of training network.

Mask R-CNN [35]: Authors in this paper presented lionize model to obtain semantic and instance segmentation. It is extended Faster RCNN model. Image is resized to mask size where data or feature extraction is imperative.
Functionality of Model: The rudimentary architecture firstly extracts feature maps from an image using CNN. It is also a top-down approach for generating masks on targeted images. ResNeXt is used as backbone architecture with the depth of 50 to 101 layers for feature extraction from the entire image [35].

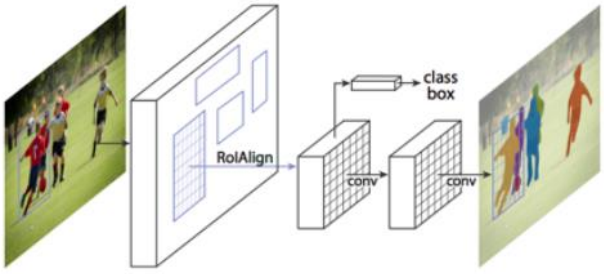


Figure. 19. Mask R-CNN Architecture [35]

Comment: Results of AP for Cityscapes dataset is 36.4 which is maximum among other methods.
Limitation: Computational cost is high.

Cascaded Deep Monocular 3D Human Pose Estimation [36]: Li et al. presented a 3D human pose estimation method using a new labialized dataset in this paper. In the 2D-to-3D pose estimation methodology, sometimes deep neural network cannot give better performance due to the data bias. Here the paper proposed a scalable approach that synthesized a large amount of training data of 3D human pose from 2D body joint locations.
Functionality of Model: The overall model is described in two stages: In the first stage of the model, 2D joints of the human pose is detected from the given image using the heat map regression technique. In the later stage of the model, 2D joint locations are passed through the 3D pose regression model and 3D pose refinement model, respectively, to estimate the 3D pose at the end of the network [36].
Comment: The result of MPJPE (Mean Per Joint Position Error) for the human 3.6M dataset is 62.9 mm, which is minimum among other methods.
Limitation: The model needs exploration for temporal information rather than still images and the multi-person scenario.

Pretraining boosts out-of-domain robustness for pose estimation [37]: Methis et al. presented the advantages of pretraining models in pose estimation in this paper. They explained how pretraining models could overcome challenges like robustness in out-of-domain data specifically for pose estimation. The authors used ImageNet as the pre-trained model and showed that transfer learning could better results in out-of-domain scenarios.
Functionality of Model: The feature extraction process is to be done on the DeepCut model. A deconvolutional layer follows the feature extraction process to localize pose heatmaps. The pose heatmap score latterly used to predict the confidence map of the relative pose [37].
Comment: Result of PCK for Horse-10 dataset is 59%, 67.7% and 74.7% respectively for MobileNet, ResNet and Efficientnet. However, accuracy is increased up to 1% for ImageNet in the case of the corruption scenario of the Horse-10 dataset.
Limitation: The suggested model can be extended to estimate the human pose in the occluded scenario for the Horse-10 dataset.

A generalizable approach for multi-view 3D human pose regression [38]: Kadkhodamohammadi and Padoy presented a novel approach for estimation 3D human pose for multi-view scenario. There are multiple methods available for the estimation of 3D human pose for a single view case. As the dataset is limited to the multi-view scenario, it is pretty challenging to estimate the 3D pose in that case. The authors suggested the state of art approach by separating a single pose from a multi-pose environment.
Functionality of Model: As shown in Fig. 20, the overall regression model is segregated into two stages. In the first stage, 2D estimation of the single-view pose is filtered out from the multi-view scenario. The second stage concatenate individual 2D pose for 3D pose regression [38].

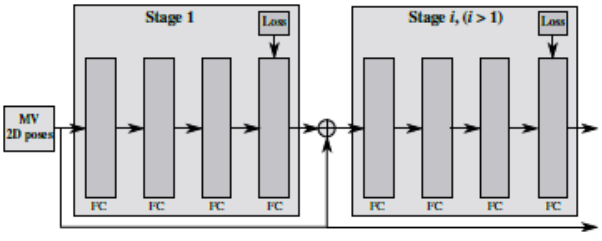


Figure. 20 Generalized Regression Architecture [38]

Comment: The result of MPPJE for the Human3.6 M dataset for single-view and multi-view scenario is 81.8 mm and 57.9 mm (minimum), respectively.
Limitation: Model is implemented on Human3.6M and Multi-view OR dataset (real-time dataset). The model can also be evaluated on other real-time datasets.

Single-shot 3D multi-person pose estimation in complex images [39]: Benzine et al. presented a novel approach of 3D human pose estimation for the complex environment. The model predicts human joints' location, the estimation of 3D pose for predicted joint location, and a complete human skeleton from the predicted 3D pose.
Functionality of Model: As shown in Fig. 21, the suggested model is designed to extend the stacked hourglass method. A stacked hourglass network provides a heatmap for human body joints. The combined approach of associative embedding network and resulted in heatmap from stacked hourglass will generate a 2D human pose. The occlusion robust pose maps (ORPM) generates 3D joint coordinates for individual 2D joint locations.

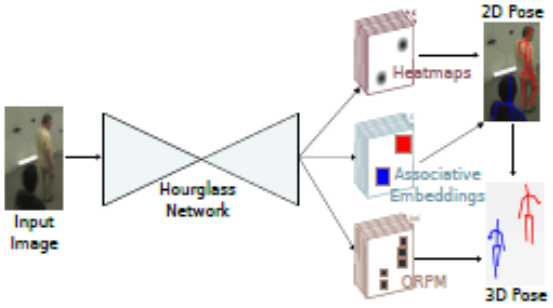


Figure. 21 Single-shot 3D Multi-person Pose Estimation Network [39]

Comments: The result of MPPJE for the Panoptic dataset is 68.5 mm (including ORPM) which is minimum among other methods.
Limitations: The suggested model gives key performance for complex images. The performance of the model can be extended as a challenge for the crowded people environment.

5. Comparative Analysis of Various Human Pose Estimation Models

The Comparative analysis of HPE using deep neural network approaches is done on multiple parameters of the model. Individual human pose estimation model is structured on various techniques like single person pose estimation, multi-person pose estimation, top-down approach, bottom-up approach. The various pose estimation models can also be compared based on the dataset used and test results achieved on evaluation parameters used to estimate the exact pose. Individual model is designed based on various deep neural network architecture like Convolutional Pose Machine (CPM), ConvNet, HrNet, Stacked Hourglass architecture, Simple Base Line, Body Joint encoding-decoding, Siamese, Pifpaf, CrowdPose, LSTM, Deep Grammar Network, Coupled Unet, Pose Box Fusion Network, RNN, OpenPose, AlphaPose etc.

Table I represents 22 recent HPE models based on deep neural network. Thirteen models are designed according to the top-down approach, seven models are designed according to the bottom-up approach, and two models are based on a mixed-model approach of top-down and bottom-up. Among all presented models, four models are designed explicitly for multi-person pose estimation, whereas the remaining models are designed for single-person pose estimation. Various datasets are used as an input source of models. Evaluation parameters used for multiple models are PCP, PCK, PDJ, AP, AR, MPJPE, PIE and Average Euclidean Distance.

The comparative analysis of the percentage of correct key Points (PCK) for various models on the different dataset is shown in Fig. 22. The primary datasets used as MPII, FLIC, LSP, EVAL and ITOP. As shown in Fig. 20, six models have used MP-II dataset as an input resource among eight different models. Distribution aware coordinate representation obtained maximum performance of 97% PCK for MPII dataset. In contrast, stack hourglass architecture obtained 99 % PCK for FLIC dataset. In another direction, stack hourglass with MP-II dataset has a PCK of 91%. PCK performance is average (as 50% Threshold) on EVAL and ITOP datasets. Most models succeeded to achieve maximum PCK performance using the MP-II dataset.

The Comparative Analysis of AP for various models on the different dataset is shown in Fig. 23. The primary datasets used as COCO, CrowdPose, EVAL and ITOP. As shown in Fig. 21, 7 models have used the COCO dataset as an input resource among eight different models. Distribution aware coordinate representation model having maximum performance as AP of 78.9. However, AP performance is reaching up to 75.5 and 74.1 for EVAL and ITOP datasets, respectively. The COCO dataset obtained maximum AP performance amongst all.

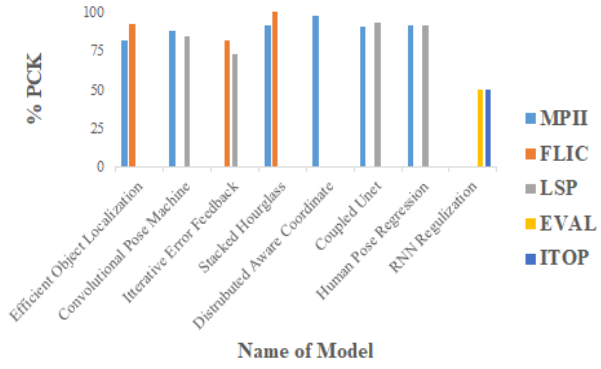


Figure. 22. Comparative Analysis of PCK Evaluation Parameter for Various Models on Different Datasets

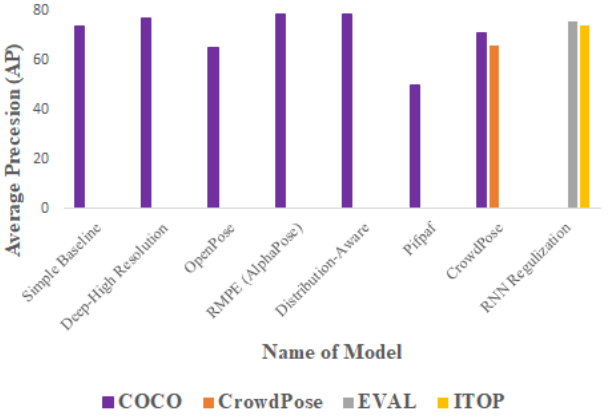


Figure. 23. Comparative Analysis of AP Evaluation Parameter for Various Models on Different Datasets

6. CONCLUSION

A complete survey of Human pose estimation using traditional and deep learning-based approach is presented here. Challenges faced in human pose estimation are occlusion, body part shortening, clothing variation, background clutter etc. The basic human pose estimation approach includes pre-processing, feature extraction, inference or confidence map, and post-processing. Human pose estimation has vast application like activity recognition, motion capture, gaming etc. The HPE can be executed using COCO, MPII, FLIC, LSP datasets. It is observed that PCK parameter on MPII dataset and AP parameter on COCO dataset gives quite good performance. It is also observed that stacked hour glass method achieved maximum PCK on FLIC dataset. In the same direction Distributed aware architecture gives highest performance for AP parameter on COCO dataset. The evaluation parameters for pose estimations are Percentages of correct key points, Percentage of correct parts, and Mean distance between tow joints using average precession. Pose estimation is classified into two different categories, i.e. traditional approach and deep learning based approach. There are various methods available for deep learning based approach. In this paper, single-person pose estimation and multi-person pose estimation using deep learning-based techniques are presented and reviewed. There are most favorable methods of deep learning-based approach using various evaluation parameters and different datasets. Comparative analysis of deep learning-based methods are shown in the form of various evaluation parameters of respective datasets.

TABLE I. COMPARATIVE ANALYSIS OF DEEP NEURAL NETWORK BASED APPROACHES

Name of Model	Name of Architecture	Estimate Single Person / Multi Person	Dataset Used	Evaluation Parameters	Test Results achieved (%)
DeepPose [25] [Top-down approach]	CNN + Regression	Single Person	LSP	PCP	61.0
			FLIC	PDJ	90.0
Efficient object localization using convolutional networks [26] [Top-down approach]	CNN	Single Person	MPII	PCK	82.0 (For Whole Body)
			FLIC	PCK	92.6 (For Head)
Convolutional Pose Machines [27] [Top-down approach]	CPM	Single Person	MPII	PCK	87.95
			LSP	PCK	84.23
Human Pose Estimation with Iterative Error Feedback [28] [Top-down approach]	ConvNet	Single Person	LSP	PCK	72.5
			FLIC	PCK	81.3
Stacked Hourglass Networks for Human Pose Estimation [29] [Mix-model approach]	Hourglass	Single Person	**FLIC**	**PCK**	**99.0**
			MP-II	PCK	90.9

TABLE I. COMPARATIVE ANALYSIS OF DEEP NEURAL NETWORK BASED APPROACHES (CONTINUE)

Name of Model	Name of Architecture	Estimate Single Person / Multi Person	Dataset Used	Evaluation Parameters	Test Results achieved
Simple Baselines for Human Pose Estimation and Tracking [30] [Mix-model approach]	Simple Baseline	Single Person	COCO	AP	73.7
Deep High-Resolution Representation Learning for Human Pose Estimation [31] [Top-down approach]	HrNet	Single Person	COCO	AP	77.0
OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields [32] [Bottom-up approach]	PAF	Multi-Person	COCO	AP	65.3
RMPE (AlphaPose): Regional Multi-Person Pose Estimation [34] [Top-down approach]	**RMPE**	**Multi-Person**	**COCO**	**AP**	**78.6**
Cascaded Deep Monocular 3D Human Pose Estimation [36] [Top-down approach]	Cascaded 3D Pose Estimation	Single Person	Human3.6M	MPJPE (in mm)	62.9 (in mm)
A generalizable approach for multi-view 3D human pose regression [38] [Bottom-up approach]	Generalized Regession architecture	Multi-person	Human3.6M	MPJPE (in mm)	81.8 (in mm) (For single-view)
					57.9 (in mm) (For multi-view)
Single-shot 3D multi-person pose estimation in complex images [39] [Bottom-up approach]	Single-shot 3D Multi-person Pose Estimation Network	Multi-person	Panoptic	MPJPE (in mm)	68.5 (in mm)
Distribution-aware Coordinate Representation for Human Pose Estimation [40] [Top-down approach]	Body Joint Coordinate Decoding-Encoding	**Single Person**	**COCO**	**AP**	**78.9**
				AR	83.5
			MPII	**PCK**	**97.2**
3D Human Pose Estimation with Siamese Equivariant Embedding [41] [Top-down approach]	Siamese architecture	Single Person	Human3.6M	MPJPE (in mm)	38.2 (in mm) (During Direction)
Pifpaf: Composite Fields for Human Pose Estimation [42] [Bottom-up approach]	PIFPAF	Multiperson	COCO	AP	50.0
				AR	55.0
Crowdpose: Efficient Crowded Scenes Pose Estimation and A New Benchmark [43] [Top-down approach]	CrowdPose	Multiperson	CrowdPose	AP	66.0
				AR	72.7
			MSCOCO	AP	70.9
				AR	76.4

TABLE I. COMPARATIVE ANALYSIS OF DEEP NEURAL NETWORK BASED APPROACHES (COTINUE)

Name of Model	Name of Architecture	Estimate Single Person / Multi Person	Dataset Used	Evaluation Parameters	Test Results achieved
Exploiting Temporal Information for 3d Human Pose Estimation [44] [Top-down approach]	LSTM [45]	Single Person	Human3.6M	Mean Per Joint Position Error (MPJPE) (in mm)	36.9 (in mm) (During Direction)
			HumanEva		13.6 (in mm) (During Walking)
Learning Pose Grammar to Encode Human Body Configuration for 3D Pose Estimation [46] [Top-down approach]	Deep Grammer Network	Single Person	Human 3.6M	Average Euclidean Distance (mm)	38.2 (in mm) (During Direction)
			HumanEva		19.4 (in mm) (During Walking)
CU-net: coupled U-nets [47] [Bottom-up approach]	Coupled UNet + Intermediate Supervision	Single Person	MPII (Top)	PCK	90.8
			LSP (Bottom)	PCK	93.4
Human Pose Regression by Combining Indirect Part Detection and Contextual Information [48] [Bottom-up approach]	CNN	Single Person	LSP	PCK	91.1
			MPII	PCK	91.2
Pose-Invariant Embedding for Deep Person Re-Identification [49] [Bottom-up approach]	CNN + Pose Box Fusion Network (PBF)	Single Person	Market-1501	Pose Invariant Embedding (PIE)	77.97
			VIPeR	PIE	54.49
			CUHK03	PIE	67.10
Recurrent Neural Network Regularization [50] [Top-down approach]	CNN + RNN [50]	Single Person	EVAL	PCK	Above 50 (Threshold)
				AP	75.5
			ITOP	PCK	Above 50 (Threshold)
				AP	74.1

REFERENCES

[1] H.B. Zhang, Q. Lei, B.N. Zhong, J.X. Du and J. Peng, "A survey on human pose estimation", Intelligent Automation & Soft Computing, 22(3), pp.483-489, 2016

[2] *Potrait Drawing*, accessed on 20 March 2021, < https://in.pinterest.com/pin/794463190503881404/ >

[3] D. Singh, "Human pose estimation: extension & application", International Institute of Information Technology Hyderabad, September – 2016

[4] *Illimination (image)*, accessed on 20 March 2021,< https://en.wikipedia.org/wiki/Illumination_(image)>

[5] *8 Tips to take better pictures by losing the background clutter*, accessed on 20 March 2021, < https://hightphoto.com/take-better-pictures-loosing-background-clutter/>

[6] Microsoft corporporation. Kinect for Xbox 360, 2009.

[7] *ClipArtMag*, accessed on 08 April 2021, < http://clipartmag.com/tag/estimation>

[8] Q. Dang, J. Yin, B. Wang and W. Zheng, "Deep learning based 2d human pose estimation: A survey", Tsinghua Science and Technology, 24(6), pp.663-676, 2019

[9] M. Kachouane, S. Sahki, M. Lakrouf and N. Ouadah, "HOG based fast human detection.", In 2012 24th International Conference on Microelectronics (ICM) IEEE, (pp. 1-4)., December-2012

[10] T. Lindeberg, "Scale invariant feature transform.", Scholarpedia, 7(5), pp. 10491,2012

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

[12] W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu and E.H. Zahzah, "Human pose estimation from monocular images: A comprehensive survey". Sensors, 16(12), p.1966., 2016

[13] M. Vrigkas, C. Nikou and I.A. Kakadiaris, "A review of human activity recognition methods". Frontiers in Robotics and AI, 2, p.28, 2015

[14] T. Von Marcard, G. Pons-Moll and B. Rosenhahn, "Human pose estimation from video and imus". IEEE transactions on pattern analysis and machine intelligence, 38(8), pp.1533-1547, 2016

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

[16] A. Kaul, M. Kasam, S. Ganju "Practical Deep Learning for Cloud, Mobile, and Edge: Real-World AI & Computer-Vision Projects Using Python, Keras & TensorFlow". United States: O'Reilly Media, 2019.

[17] S. R. Ke, L. Zhu, J. N. Hwang, H. I. Pai, K.M. Lan and C. P. Liao, "Real-time 3D human pose estimation from monocular view with applications to event detection and video gaming". In 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance (pp. 489-496). IEEE, August-2010

[18] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3202-3212, 2015

[19] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C.L. Zitnick, "Microsoft coco: Common objects in context." In European conference on computer vision (pp. 740-755). Springer, Cham, September-2014

[20] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3674-3681), 2013

[21] M. Eichner and V. Ferrari, "Appearance sharing for collective human pose estimation." In Asian Conference on Computer Vision (pp. 138-151). Springer, Berlin, Heidelberg, November-2012

[22] L. Sigal, A.O. Balan and M.J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion." International journal of computer vision, 87(1-2), p.4, 2010

[23] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts." IEEE transactions on pattern analysis and machine intelligence, 35(12), pp.2878-2890, 2012

[24] Sudharshan Chandra Babu, 2019, Nanonets, accessed on 20 March 2021, <https://nanonets.com/blog/human-pose-estimation-2d-guide/>

[25] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks." In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1653-1660), 2014

[26] J. Tompson, R. Goroshin, A. Jain, Y. LeCun and C. Bregler, "Efficient object localization using convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 648-656), 2015

[27] S.E. Wei, V. Ramakrishna, T. Kanade and Y. Sheikh, "Convolutional pose machines." In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 4724-4732), 2016

[28] J. Carreira, P. Agrawal, K. Fragkiadaki and J. Malik, "Human pose estimation with iterative error feedback." In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4733-4742), 2016

[29] A. Newell, K. Yang and J. Deng, "Stacked hourglass networks for human pose estimation." In European conference on computer vision (pp. 483-499). Springer, Cham., October-2016

[30] B. Xiao, H. Wu and Y. Wei, "Simple baselines for human pose estimation and tracking." In Proceedings of the European conference on computer vision (ECCV) (pp. 466-481), 2018

[31] K. Sun, B. Xiao, D. Liu and J. Wang, "Deep high-resolution representation learning for human pose estimation." In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5693-5703), 2019

[32] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields." arXiv preprint arXiv:1812.08008, 2018

[33] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P.V. Gehler and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation." In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4929-4937)., 2016

[34] H.S. Fang, S. Xie, Y.W. Tai and C. Lu, "Rmpe: Regional multi-person pose estimation." In Proceedings of the IEEE International Conference on Computer Vision (pp. 2334-2343), 2017

[35] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask r-cnn." In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969), 2017

[36] S. Li, L. Ke, K. Pratama, Y. Tai, C. Tang and K. Cheng, "Cascaded deep monocular 3D human pose estimation with evolutionary training data." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6173-6183), 2020

[37] A. Mathis, T. Biasi, S. Schneider, M. Yuksekgonul, B. Rogers, M. Bethge and M. Mathis, "Pretraining boosts out-of-domain robustness for pose estimation." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1859-1868), 2021

[38] A. Kadkhodamohammadi and N. Padoy, "A generalizable approach for multi-view 3d human pose regression". Machine Vision and Applications, 32(1), pp.1-14, 2021

[39] A. Benzine, B. Luvison , Q. Pham and C. Achard. "Single-shot 3D multi-person pose estimation in complex images." Pattern Recognition, 112, p.107534, 2021

[40] F. Zhang, X. Zhu, H. Dai, M. Ye and C. Zhu, "Distribution-aware coordinate representation for human pose estimation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7093-7102), 2020

[41] M. Véges, V. Varga and A. Lőrincz, "3d human pose estimation with siamese equivariant embedding." Neurocomputing, 339, pp.194-201, 2019

[42] S. Kreiss, L. Bertoni and A. Alahi, "Pifpaf: Composite fields for human pose estimation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 11977-11986), 2019

[43] J. Li, C. Wang, H. Zhu, Y. Mao, H.S. Fang and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 10863-10872), 2019

[44] M. Rayat Imtiaz Hossain and J.J. Little, "Exploiting temporal information for 3d human pose estimation." In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 68-84), 2018

[45] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink and J. Schmidhuber, "LSTM: A search space odyssey." IEEE transactions on neural networks and learning systems, 28(10), pp.2222-2232, 2016

[46] H.S. Fang, Y. Xu, W. Wang, X. Liu and S.C. Zhu, "Learning pose grammar to encode human body configuration for 3d pose estimation." In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1), April-2018

[47] Z. Tang, X. Peng, S. Geng, Y. Zhu and D.N. Metaxas, "CU-net: coupled U-nets." arXiv preprint arXiv:1808.06521, 2018

[48] D.C. Luvizon, H. Tabia and D. Picard, "Human pose regression by combining indirect part detection and contextual information." Computers & Graphics, 85, pp.15-22, 2019

[49] L. Zheng, Y. Huang, H. Lu and Y. Yang, "Pose-invariant embedding for deep person re-identification." IEEE Transactions on Image Processing, 28(9), pp.4500-4509,2019

[50] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung and L. Fei-Fei, "Towards viewpoint invariant 3d human pose estimation." In European Conference on Computer Vision (pp. 160-177). Springer, Cham, October-2016

Miral M. Desai is working as an Assistant Professor at Department of EC Engineering – Faculty of Engineering & Technology, Charotar University of Science & Technology, Changa. He has done his M.Tech from Institute of Technology, Nirma University in 2014. He received his B.E. degree in Electronics & Communication Engineering from south Gujarat University in 2010. His Research Area is Machine Learning, Deep Learning, Computer vision algorithm, Embedded System Design, Embedded Linux based applications, Digital System Design. He has published eight papers including peer reviewed international journals and conferences. He is a life member of ISTE.

Hiren K Mewada is Assistant Research Professor at Prince Mohammad Bin Fahd University, Al Khobar, Kingdom of Saudi Arabia. He has more than 18 years of academic and research experience. He has completed his Ph. D. and M. Tech. in Electronics Engineering from S.V. National Institute of Technology – Surat, Gujarat, India. His research interest lies in the area of Signal/Image Processing ranging from Software simulation to hardware implementation including FPGA and ARM processor based design. He is a life member of IETE and ISTE. He has authored one book, four chapters in different book series and more than 60 papers including peer reviewed international journals and conferences. He has completed several research projects from different agencies including Board of Research in Nuclear Science (BRNS), Government of India, Gujarat Council of Science and Technology (GUJCOST), Government of Gujarat, Solution with Innovation, USA.