Chapter 2

THEORETICAL FOUNDATION AND LITERATURE SURVEY

This chapter presents the survey of the state-of-the-art techniques for HPE and HAR. Section 2.1 presents the theoretical foundation for HPE and HAR. The section 2.2 discusses the applications of HPE and HAR. Section 2.3 discusses the conventional and deep learning-based literature for HPE. Section 2.4 presents deep learning-based literature for HAR. Section 2.5 presents benchmarks and evaluation metrics for both HPE and HAR. Section 2.6 presents the comparative analysis of the recent state-of-the-art techniques for HPE and HAR. At last, Section 2.7 discusses the research gap and challenges for both.

2.1 Introduction

Computer vision is a versatile field related to artificial intelligence, machine learning, robotics, and signal processing. The purpose of computer vision is to program a computer to understand, process and analyze images. The fundamental tasks of computer vision are object detection, pose estimation, tracking, and segmentation. Object detection and pose estimation treated as one of the essential jobs due to its wide range of applications. Object detection has been used to identify and locate the object class in an image or video. Till day, Face and Human are the most examined objects. Human beings are usually considered as an articulated subject consisting of fixed moving parts linked to certain expressive points.

Depending on the requirement of the output dimension, the problem of HPE can be divided into 2D HPE and 3D HPE. The 2D HPE predicts the position of body joint locations in the frame (in the form of pixel value locations). To the contrary, the 3D HPE predicts a 3-D spatial ordering of every body joint location as the ultimate output.

From 2001 to 2019, scientific community has shown an evergrowing interest in 2D/3D HPE. As reported in Fig. 5.1, more than 5000 publications in this topic have been published and indexed in Web of Science , ranging from 2D HPE to 3D HPE and HAR with considering indoor and outdoor environments.

Due to the release of new advance datasets, the considerable increase of the scientific community on this field occurs. In spite of plenty of research, HPE and HAR remains to be a challenging and unresolved problem. The most acute challenges are: (1) high diversity in the human poses, (2) fluctuation in lighting situation, (3) partial occlusion

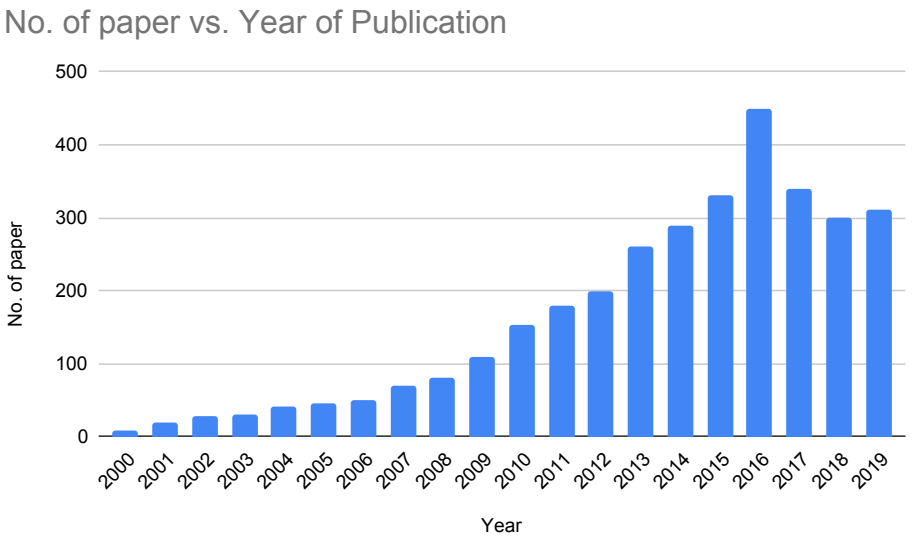

FIGURE 2.1: Number of research publications in the area of Human Pose Estimation (HPE) and Human Activity Recognition (HAR).

FIGURE 2.2: Some applications for HPE and HAR (from top to bottom and left to right): Human-robot interaction, human-computer interaction, gaming, video surveillance, sports, and proxemics.

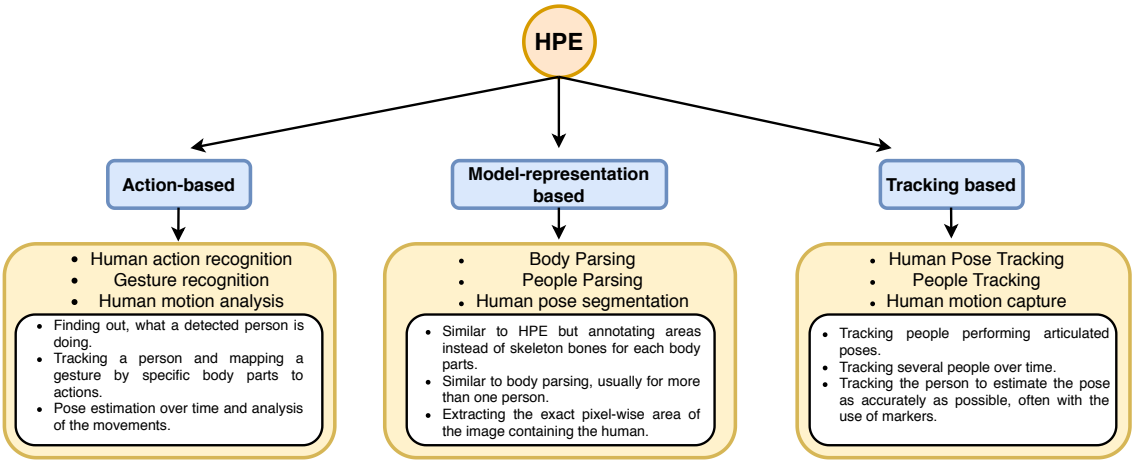

FIGURE 2.3: A brief overview of the related tasks with Human Pose Estimation (HPE)

because of body part of the object itself and through other objects in the background and, (5) complex skeletal structure of human, (6) variation in the person visual look in the images, (7) 3D information loss while 2D image projection for pose observation, and (8) Less availability of ground-truth data. Till now, no such method can give satisfying result while handling all the challenges above.

In literature, there are very few reviews present with the title "HPE" and most of them deal with particular specialization. The authors of [18], provide a study for 2D and 3D HPE, which is focused entirely on the model-based techniques. Liu et al. [19] presented a survey covering the body part parsing techniques for HPE. Zhong et al. [20] describe HPE by data-driven methods with their summary and conclusion. Dang et al. [21] covers the deep learning based 2D HPE approches.

The survey by Moeslund et al. [22], Ji et al. [23], Jung et al. [24], provide us the information in the field of "vision-based human motion capture and its analysis." Moeslund gives the advancement of this field from the year 2000 to 2006. Ji et al. [23] provide a review on the progress of view-invariant perspective of human motion analysis. Recently,

authors of [25] introduce a survey includes how to capture and analyze the human body motion and shape. The essential tasks of the motion analysis are initialization, tracking, pose estimation and recognition.

Similarly, we have gesture recognition, which focuses on tracking a person and mapping a gesture by specific body parts to actions. Reuteray et al. [26] and [27] provide a survey focus on gesture recognition to facilitate human-computer interaction. Poppe [28] and Dawn et al. [29] and [30] presented a study on human action recognition and gave a detailed overview of the advancement in this field.

In literature, there are very few literature survey that discuss the deep learning based HAR. So, in this chapter, we also present the survey for deep learning based HAR techniques. Although previous studies have examined the problems of human motion analysis, body parsing, recognition, and tracking analysis, they're still missing a study which outlines all the recent evolution for HPE and HAR.

To capture the advances in the field and to be persuasive in our methodology, we restricted this survey to a class of strategies that are currently the most widely used, in particular, the 2D and 3D body pose estimation and HAR from image and videos. The objective of this chapter is to present:

(i). Applications of HPE and HAR.

(ii). Provide a survey which covers both HPE and HAR.

(iii). We give a complete flow of the HPE with classical and deep learning based methods which describes the basic components used in pose estimation followed by summary table containing the latest work using deep learning for both 2D and 3D HPE.

(iv). We discuss recent state-of-the-art techniques that utilizes deep learning approach.

(v). Evaluation metric and datasets.

(vi). Also, we explain the challenges for both HPE and HAR.

This chapter is divided into two parts : (i) Conventional and deep based HPE techniques, and (ii) Deep learning based HAR techniques.

2.2 Applications

The growing interest in the vision based HPE and HAR systems is explained by several factors. According to us, the most significant cause is the evoluation of the related areas that utilize the pose estimation and recognition methods. In addition, current progress in reaserch in Augmented Reality (AR), Artificial Intelligence (AI) and medical imaging give contribution to the progress of research in estimation methods.

The three main area of application are determined individually by the authors are: Human-computer interation, survelliance and analysis.

2.2.1 Human-Computer Interaction

It is advantageous to develop the more advanced interface in between human and the intelligent systems over traditional computer keyboards and mouse for understanding the human gesture visually. The Human-computer interaction field of application deals with the tasks in which the estimated pose is utilized to give control over the functionality for

designing virtual game interface, and its remote control, animations and virtual environments. One of the examples is utilizing hand gesture to control the flow of presentation slides [31].

In current years, the attention is in utilizing Unmanned Aerial Vehicles(UAVs) to complete a sequence of tasks which can be dangerous or uncomfortable executed by the human has subject to substantially increase. For sports and game motive, this highly used with the large chance to buy the cheaper drone. This leads to push the scientist highly utilize the potential of UAV for video-based autonomous landing applications and object tracking. Besides this UAV landing, many research works have given a big contribution to support the application in self-driving cars.

Assisted Living: Mostly HPE and human activity recognition task used for helping the disabled persons. For example, a detection system when a person falls [32]and a robot controlled by blinking [33].

Gaming: It includes all interactive gesture-oriented Games. To which, all non-intruding body movement can be utilized by players to interact with games. The most illustrative example in the dissemination of the Kinect Sensor together with toolkit extensions, which enables unification of total-body control along with games [34] and Virtual Reality applications [35].

Intelligent Driver assistant systems: Some illustrative examples are supervising driver alertness using head pose tracking [36] [37] , integrating the driver's head and hand poses

tracking to make distraction alert [38] , modeling driver leg behavior to reduce pedal misapplication or predicting driver turn intent [39].

Sports Performance Analysis: For investigating performance and training in many sports like ballet, skating, or golf require to pose estimation and activity recognition task [40].

2.2.2 Video Surveillance

In video surveillance, pose estimation utilize to detect the abnormal crowd activity, perimeter breach detection, camera tampering, and loitering individuals detection [41].
The main distinction that distinguish the approches found in literature about video surveillance applications consist in the acquisition systems and the number of estimated human subjects (single or multi-person). Among the video surveillance one of the most active reserach topic is mainly focused on human pose estimation.

2.3 Conventional and deep learning based human pose estimation techniques

The four main components for conventional HPE are illustrated in Fig 3.5. It is not compulsory to follow all the mentioned steps for the estimation. Each flow denotes some different composition of the specific type of models.
The preprocessing includes mostly two parts camera calibration and body localization.

The primary goal for applying these techniques are the unification of multiple views, human shape estimation, human location, and size estimation. After preprocessing, the feature extraction techniques are used to acquire the main features from the human subject, that further supplied to next stage for estimation. We divided this stage according to the type of encoding used in feature selection, low-level, high-level and motion features. Then we are defining different kind of body models which are leading towards mainly two types of techniques for HPE called discriminative and generative.

Generative methods (also called model-based techniques), utilizes the human body model information for estimation HP. On the other hand, model-free or discriminative techniques use learning to map the appearance to the body pose. These methods are fast and more accurate compared to model-based but limited to use background subtraction as a preprocessing. Due to the advances in learning based approaches, using deep learning, these methods directly estimate the pose from the input images.

In this paper, we arrange the work as, in section 1, which we already described above the introduction of the HPE. In section 2, some related actions for HPE. Section 3, the steps involved in conventional HPE techniques like pre-processing methods required to move further for the estimation, the types of feature description for HPE, then the types of body model used for HPE, at last paper explains the main two categories of HPE methodology. Section 4, discuss all the types of deep learning based HPE.

The section 5, discusses the benchmarks and evaluation metric utilized for estimation. Section 6, provides the comparative study of the review of the recent advancement for both 2D and 3D pose estimation. Last section 7, give the challenges and conclusion for

further studies.

2.3.1 Analysis

The analysis of the estimated pose utilized in clinical applications that monitor patient activity like diagnosis of orthopaedic diseases [42], in the domain of rehabilitation medicine [43] and gait analysis, patient pose estimation in the bed is required in several field such as sleep laboratories [44], epilepsy monitoring and intensive care units [45].

Identifying anomalous human pose data is crucial to many emerging data-driven artificial intelligence systems. For instance, patient behavior monitoring systems can analyze patient behavior based on patient movement and pose predictions. Although pose tracking methods have improved over the years, anomalous pose estimates, even if infrequent, can result in troublesome events, such as error information on the patient behaviors, which can lead to false diagnosis and requires human labor intensive processes to identify those anomalous poses. This cost could be mitigated by correcting or identifying anomalous pose estimates in an automated fashion. Thus, there are many techniques presents an anomaly analysis framework for clinical human pose estimates to address these concerns.

In Fig. 6.6, few of the applications as mentioned above are represented, along with current technological advancement.

2.3.2 Input Modality

This section gives the discription of the different types of input present. The following types of inputs are easilly available for 2D/3D HPE:

2.3.2.1 RGB Image

The real-environment images, i.e., images which we people recognize around us on a regular base are the usually general kind of input for HPE. Methods that solely work with RGB inputs holds a big benefit over other models in the perception of the movement of the input source. The aforementioned is because of the easy availability of popular cameras (which take RGB images) so that these methods can be utilized for a variety of projects.

2.3.2.2 Depth Image

In this type of image, the measured spatial value of the pixel correlates to the range from the camera, estimated using the flight time. The advent and demand for inexpensive tools such as Microsoft Kinect have make it more accessible to get depth information. Depth images can complements the RGB images to produce numerous accurate and complex vision-based methods, while depth-based methods are often utilized when privacy is an issue.

2.3.2.3 Infra-red(IR) Image

The measured value of the pixel in an IR image is defined by the volume of infrared light returned back to the lense of the camera. Experiments in vision-based on these images are insignificant related to depth and RGB images. The Kinect system also delivers this type of image during the recording. Nevertheless, there are currently no records that include this type of image.

In this chapter, we only discuss the 2D and 3D HPE over RGB image.

2.3.3 Conventional Approach

2.3.3.1 Preprocessing

Information calibration In HPE, different camera views are used to capture the human poses, and because of the variation in the acquired data, camera calibration is used for pre-processing. Mostly utilized camera calibration techniques are, Stoll et al. [46] primarily coordinate then calibrates at the point to adjust the information video before building up 2D and 3D skeleton body. Likewise, Gall et al. [28]utilize a classifier fusion approach to combine the body pose information gathered from all the multi-view cameras. Kinect sensor [47][48] is also used for calibrating the human motion.

Recently, Shin et al. [49] presents an camera calibration technique for survelliance by utilizing the human periodic walk motion and the geometric rectangular model. Recently, Pavlakos et al. [50] and Simon et al. [51] utilizes calibrated multiview setup for HPE and

detection. On the other hand, there are many techniques which partially use the camera calibration such as Rhadin et al. [52] uses an only intrinsic parameter as required.

Localization Body localization locates human being in complex images. It contains human detection and background subtraction.

Human Detection Human detection is used to identify and localize the position of the human in the images or video sequences. Currently, a new version of semi-supervised boosting strategy [53]is presented for scene-oriented person detection. Aguilar et al. [54] use HAAR-LBP and HOG cascade classifier with saliency maps for pedestrian detection. Similarly, [55] develop a new method by integrating the mostly utilized HOG (Histograms of Oriented Gradients), Visual Saliency theory and multi-level deep network based saliency prediction model for human detection in video sequences. After the invention of deep learning, R-CNN [56] have been highly utilize for human detection. The deep network utilize by the system [57] which outputs a set of distinct detections by generating prediction jointly by using a recurrent LSTM layer for sequence generation and system end to end training with a novel loss function.

Background Substraction Background substraction mostly utilized for detection of moving object location. Many of the human detection algorithms use the pixel-wise human segmentation. Conventional segmentation algorithms use different features to model the background and foreground for each pixel. Saeed et al. [58], suggests a background

subtraction method that uses multiple thresholding procedures to detect an object of interest for a particular scene. Zhang et al. [59], introduced a method that initially used a median filter to retrieve the background image of the video. Subsequently, improved background subtraction was employ to localize and track an object. Guo et al. [60] , proposed a novel moving object detection procedure and combined it with the modified frame-difference and Gaussian mixture background subtraction.

Babaee et al. [61] introduced a new technique using deep learning for the background subtraction job. The method involves three processing action steps, namely the creation of background models, CNN for learning features, and the post-processing.

Summary:

The primary goal of these preprocessing technique is the unification for multiple views and human shape, location and size estimation (shown in Fig. 3.4). These methods make system leads to provide a decidedly fewer keypoints for learning, prevent to learn from the unnecessary information and many times predicting certain location keypoints even if they are occluded. The conclusion is that these pre-processing stages make the system efficient and robust.

2.3.3.2 Feature Discription

The human body is very complicated because of lots of limbs and joints. Realistically estimating the positions of joints and the length of limbs is a challenging job for both 2D

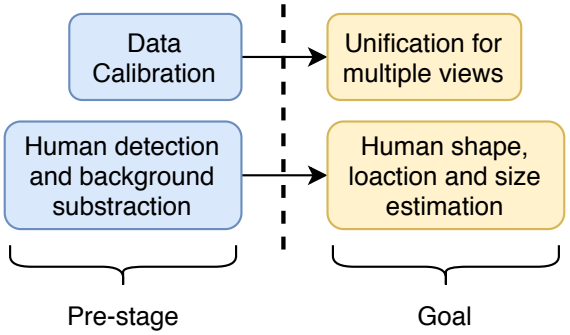

FIGURE 2.4: Goals of Preprocessing Techniques

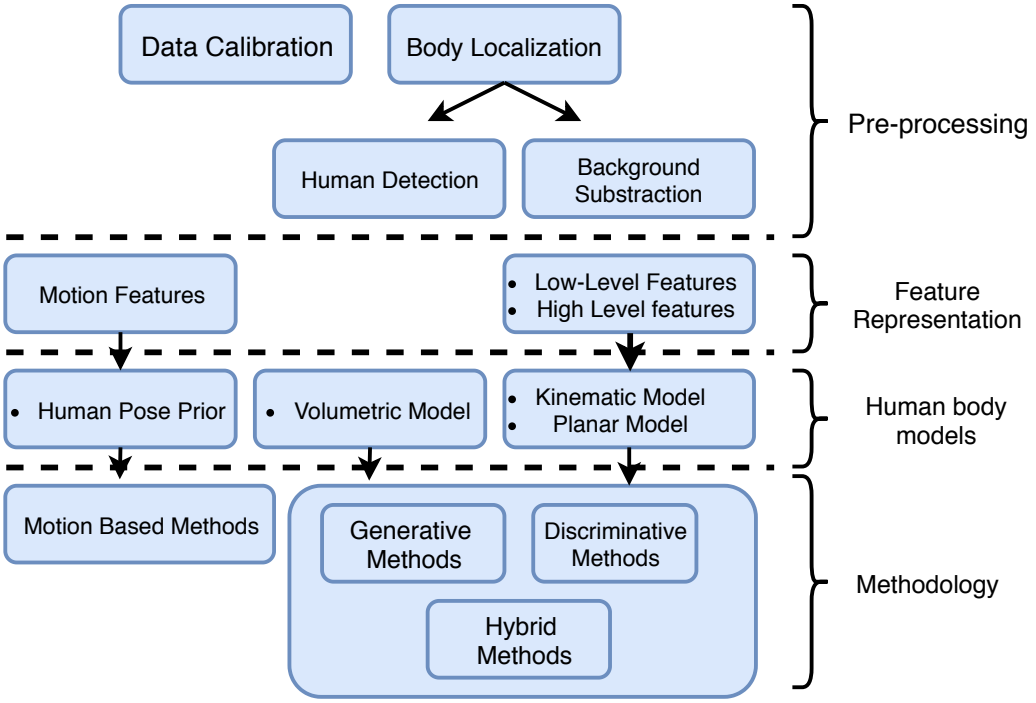

FIGURE 2.5: The framework of the review which mainly contains four stages (left to right): Pre-processing, Feature Representation, human body models and Methodology.

and 3D cases. For pose estimation, the most critical stage is to extract the key points and use them for the next stage. The key points or feature points of the image gives most of the representative information of the image but contains lots of redundancy and noise. That's why feature encoding is used. According to the type of encoded features, the following subsections are given below:

Low-level features: Earlier techniques in this area use local feature description such as edges [62] [63], contour [64] , silhouettes [65] , colour [66] , obtained after applying background subtraction. Edges capture the changing lines in an image and are usually computed using convolution. Additionally, silhouettes give global descriptors that contain the whole view of an object.

High-level features: Feature-based on edges and gradient are encoded in the histogram like HOG [67] , Scale-invariant feature transform [68], and edgelet [69] [70] features. Dalal et al. [67]prove experimentally that grid based HOG feature considerably outperforms the remaining feature descriptors for detecting human. Wang et al. [71]introduced a new HPE method using HOG feature with AdaBoost algorithm. Yaung et al. [72] proposed a method to estimate the face pose using 2D and 3D HOG feature with SVM and multi-layer perceptron (MLP) network. Bhuvaneswar et al. [73], present a human detection approach using silhouette-based descriptor named edgelet. Bo et al. [74], explain a static human detection procedure using part detectors with the new edgelet features to give a silhouette pattern's description. Sabzmeydani et al. [75], proposed a novel method utilizing learned shapelet features for detecting pedestrian detection. Chen et al. [76]recommended using the integration of template warping with SIFT- correspondence method for 3D pose estimation.

Motion features: In smart surveillance systems, pose tracking is performed by utilizing estimated pose from images. Spatial and temporal correspondences in videos are

extremely useful by correcting the estimation failure in a single frame. The motion features, for instance, optical flow [77] , robust optical flow [78], motion boundaries and edge energy and their fusion are used to improve the estimation performance [79] . Pfister et al. [77], introduced a new method for pose estimation, by using the direct process to regress the heat maps and improve its performance by integrating it with optical flow and the spatial fusion layers. The authors of [80] improve the HPE performance by utilizing the Pfister concept and ensemble CNN. In [81], they showed that after combining both RGB and motion features in the deep ConvNet system, the system outperforms the available state-of- the-art methods for human body pose estimation in the video. The authors of [82] , proposed a novel method for 3D human motion estimation from the monocular video images. They make use of optical flow to restore human motion at a time from an initialization frame.

Summary:

The efficiency of HPE system relies on the type of information or image features selected to represent human poses. Till now, the abundant number of features are introduced by many authors. Mostly utilized features and their goal are briefly described in Fig. 3.6. For improving performance and making system robust, dimensionality reduction and different encoding techniques have been proposed. The most famous dimension reduction method is a bag of words. Currently, after the exposure of deep learning much literature has been utilizing the CNN feature extraction technique for the estimation.

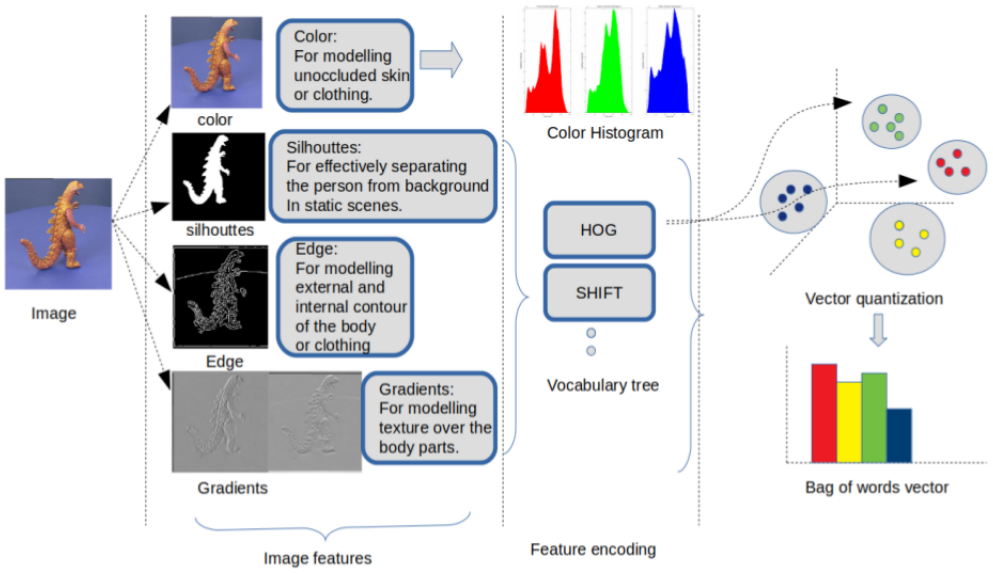

FIGURE 2.6: Image features

2.3.3.3 Body Modelling

The main problem in estimating the human pose is how to define a human body model (HBM). The HBM described by using kinematic structural, shape and texture information. The various types of body models will explain in the coming paragraphs.

HBM that use skeleton for representation is known as kinematic models (KM) [83][82]because the skeleton can easily define the kinematic property. We have two types of the KM; the first is the popular predefined model, and the other one is the learned graph model. The most commonly used graph model is the pictorial structure model [84] [85]. Other popular predefined models, is to learn from images the body part relations [86] . In addition to this, learned tree-based structure Bayesian networks [82] [87].

Most models explain the body like a kinematic tree made up of segments associated with the joints. Each joint have multiple degrees of freedom (DOF), which indicate the number

of directions in which the joint can locate. Every DOF in the body model together form a pose representation. These models can be described in either 2D or 3D.

Rather than acquiring the relationship between the body parts, the planar model is also used for learning appearance. Active shape models are one of the most used models for learning the appearance and shape of the body [88] . An additional case is of the cardboard model, formulated using the knowledge of body parts shapes and object foreground.

The Volumetric model realistically represents the 3D human body poses and shapes. Geometric body meshs and the shape are two useful volumetric models. Whereas you use geometric shapes for modeling components, body parts which are reassembled using a set of cylinders, comics, and many different shapes, to assemble body limbs. Such as, a human model can be designed in a composite form of cylinders, with every cylinder connected to one or more cylinders [89] . Another way to model human body using volumetric modeling approach is to make use of mesh. The meshes are easy to get deform and also triangulated, making them more suitable for displaying non-rigid human bodies [90] [91].

Many determinants limit the posture of human body, for example, behavioral patterns of the movement in certain activities, boundaries of the joints, and the kinematics. The kinematic dependencies along with a dynamic model give sufficient information for HPE.

The obtainability of motion capture methods [92] [93] [94]enables the pose priors learning through the data information. To learn body pose properties expertly, the authors of [95] collecting a set of motion capture information to examine all the human pose circumstances. By utilizing gathered data, a set of common joint angle training data could

be used that was marked with positive and negative examples of human poses. Although, present pose priors gained from one motion have issues generalize to new motions.

Summary:

Body configuration is represented by utilizing various body models. The kinematic model is used to show the skeleton of both 2d and 3d. Cardboard model gives the representation and shape of the body. Volumetric model mostly used for 3D human body shape and pose estimation. Largely utilized model for both 2D and 3D case is the skeleton, which has excellent kinematic property. The model contains approximately eight parameters for internal proportion and thirty for joints, used to encode the position of the clavicle, hip and skull joints. The model is widely utilized for both discriminative and generative techniques. In the discriminating method, kinematic models are usually used for the independent assembly of the body joints and parts. As an alternative in generative methods, the models are mapped onto the planes together with the pose, making a comparison with image evidence for checking the projected body pose.

2.3.3.4 Methodology

All the recent work on HPE are classified on the basis of the modeling procedure like it uses either geometric projection oriented approach or taken like a certain image processing job, the classified two main categories are called generative and discriminative methods.

Generative and Descriminative Methods: In generative model (also called top-down or model-based method), HPE modeled as a geometric projection of the real 3D-scene using computer graphics concept. The process compares the human model to the image information. The body model is created by utilizing the annotated data. After generating the body model, it is fitted to the picture with the goal that the case of the model is the nearest to the object of the image.

In the discriminative (or model-free approach), HPE is modeled as learning or mapping based method from image features to the human body pose. The mapping is learned by utilizing the training data and then adjusts the mapping using classification and regression techniques. The learning strategy is normally quicker since it only requires the picture information, while the model-based methods model the whole procedure of the given problem. Both the techniques conflict toward the path they pursue. The model-based strategies start with the human body model that projects the image in the plane to verify with the evidence of the image, while the second class begins with image information and gradually learn the whole concept that model the relationship between the image knowledge and the human poses using training data.

I. Descriminative methods: These techniques begin with image information, based on mapping or search oriented approach. The model learns the description of the relationship or map between image clue and human poses. The mapping utilized through the classification or regression technique and bundle of annotated data used to learn with the supervised approach.

At test time, model-based techniques minimize the error while model-free does not deal with any cost function. This usually indicates that the model-based techniques are faster because they fall into formulation calculations or restricted search problems rather than doing iterative optimization to fit the problem.

The regression-oriented techniques deal with HPE, by locating the object using the image feature, although the classification-oriented techniques do HPE by maximizing the score. Model-free or learning based methods seek out optimal solutions within their scope. Many studies have been carried out using methods of this category and clustered into two fundamental parts: learning-based and example-based methods. The sub-parts have been split in the following ways:

1. **Learning based methods:**

   (i) Mapping and Learning based methods: : Initial techniques in this approach utilize the probabilistic formulation to map the optimum-candidate search are Bayesian mixture of experts (BME) [96] [97], specialized maps [98], nearest neighbor [99] [100] . The techniques usually assume that a particular mapping function produced the image feature and focused on the probabilistic formulation of the problem. A well-known model for taking in these sorts of maps is a support vector machine (SVM) [101] [102], Which utilized hyperplanes to make a distinction in two classes. The techniques used for both classification and regression but mostly for classification. Likewise, a Bayesian kernel-based method called a relevance vector machine (RVM) uses a relevance vector for the classification purpose

[103] [104].

The learning-based method classified in two approaches one is direct, and other is boosting approach from 2D to 3D. The first class uses learning to directly map image feature to the 3D pose [105] [106] . Another class of methods firstly map image feature to 2D poses and then use modeling or learning to map 2D to 3D poses [107] [83]. Despite many advancements in learning-based approach for HPE, the evolution of the deep learning approach decreases the number of researches in the traditional machine learning methods. The next section describes the brief research using deep learning for HPE.

(ii) Deep laerning based HPE methods: This section comprises the approaches used for HPE using deep learning. We divide the HPE into four sections: (i) Direct joint prediction, (ii) Indirect joint prediction, (iii) One-stage approach, and (iv) Multi-stage approach. We observe that, deep learning method highly improve the performance of HPE.

Direct prediction: These techniques usually predicts the 2D/3D joint coordinates directly from the image or video. We also called it regression based techniques. The direct prediction of joint coordinate from the image is challenging in both the 2D/3D cases, as the data is extremely non-linear. For 3D cases, we have very less amount of 3D ground-truth data for deep learning. That's why the system gives less accuracy using this. Toshev et al. [108] proposed an DNN based direct regression technique for body joint prediction. Similarly, Compositional pose

regression [10] has introduced, where the authour utilizes the bone representation rather than joint data for HPE. Recently, Luvizon et al.[109] introduced an pose regression using deep learning with context information along with detecting body parts. Semantic graph convolutional network [110] has been proposed for 3D pose regression, where the local and global semantic learning has been performed for making system accurate.

Indirect prediction: On the other hand indirect prediction includes techniques where we utilize the joint heatmaps and the intermediate part patches for the final prediction. For 3D cases, we also utilize the 2D joint coordinate for final prediction. Largly utilized concept stackhourglass approach [111] uses repeated down and up sampling strategy with in-between supervision to improve the HPE performance. Similarly, Chou et al.[112] utilizes the two stackhourglass concept and named it as generator-discriminator, which hihgly improves the prediction result. Nibali et al. [113] proposed an 3D HPE technique, where they utilize the joint heatmaps for estimation. On the other hand few approches utilize the joint coordinate for 3D reconstruction. Martinez et al. [114] utilize the 2D joint coordinate for 3D HPE, which improves the performace by reducing the computational overhead and easy to understand the reason of error. Similarly Ramirez et al. [115] also utilize the predicted 2D coordinates for 3D reconstruction.

Single and Multi-stage approaches: The one-stage architecture usually have end-to-end learning, maps the image to poses. The multi-stage architecture uses

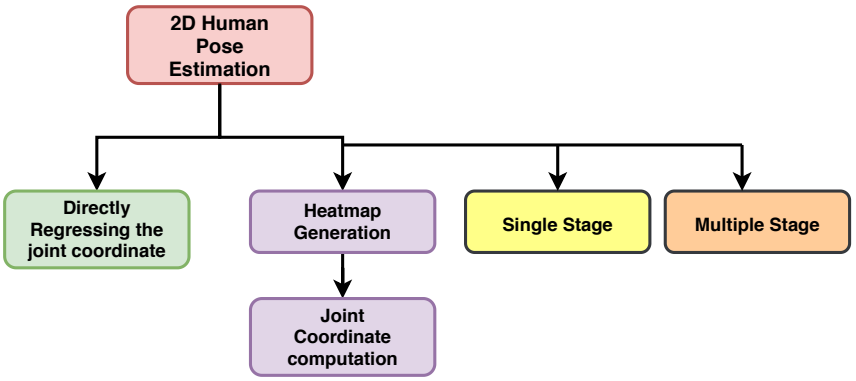

FIGURE 2.7: Direct method.

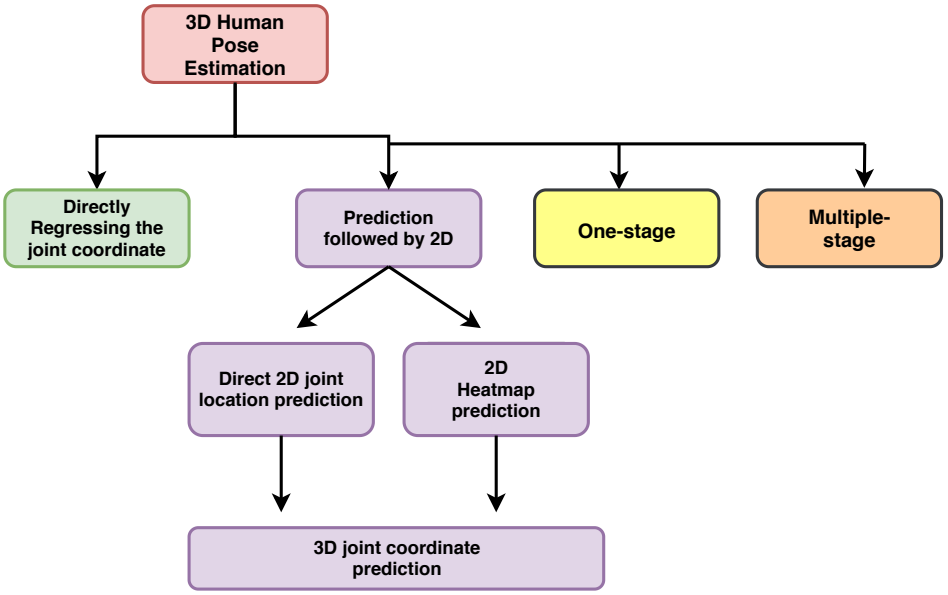

FIGURE 2.8: Indirect method.

intermediate supervision with many stages like multiple HPE method, first detect the area of interest having human and then estimate its pose. Similarly, for 3D case, first detect the 2D pose and then extend it to 3D HPE. The end-to-end training is easy compare to multi-stage training architecture.

2. **Example-based methods:**

   In this approach, the human pose is estimated using the different sets of particular

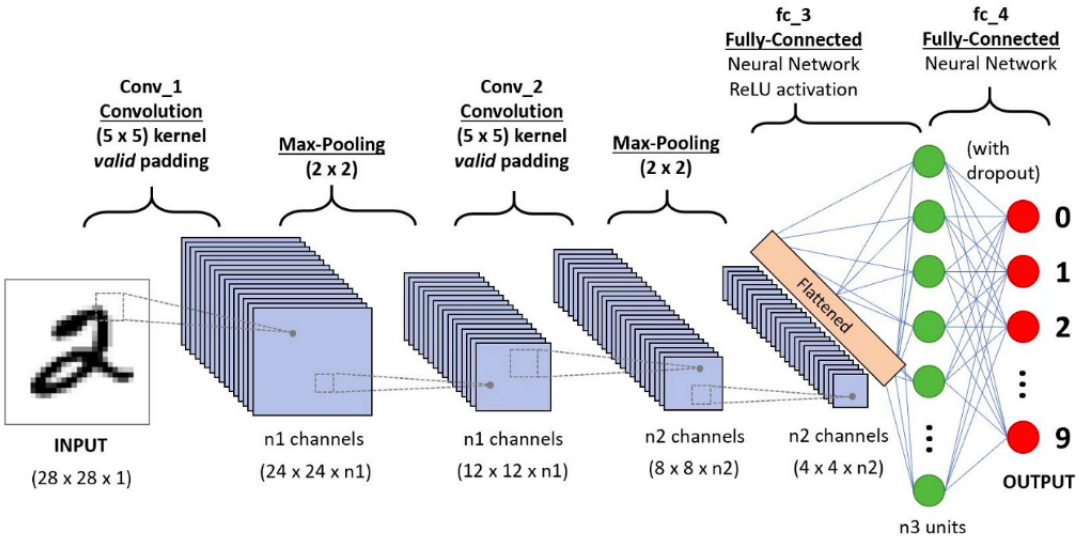

FIGURE 2.9: Convolutional neural network architecture.

poses along with their representations [116]. Mostly used classification techniques in this approach is random forest and randomized tree [117][118] because they provide speed and efficiency to the approach. The improved random forest was utilized in [7], they use two-layer random forests for the joint regressors; first one serves the discriminative body part classifier property and then another for identifying the joint localization by using the first layer output.

II. Generative based methods: The pixel-level predictions provide a collection of absolute local body pose hints that are absurd to meet kinematic restrictions. Usually, through placing the generative model to the cues, solve this problem. These techniques models the likelihood probability of observations for HPE. For locating likelihood probabilty peaks an intricate search is utilized on the state space. Model-based techniques are prone to local minima, therefore need a true initial pose estimate, irrespective of the used

optimization module. The body pose is usually derived from local optimization or from stochastic search.

Summary:

Both generative and discriminative methods are utilized for HPE. But generative configuration is very challenging for the high dimensional articulation space. The method gives a good result when observations from multiple cameras are available. For the more general articulation and monocular observations, that are often the focus of pose estimation algorithms, this class of methods has not been very successful to date. After the exposure of deep learning, the discriminative methods are widely used, even they have efficiently estimated the 3D pose from single image. Despite the popularity and lot of success, they also have some limitations. First, they are only capable of recovering a relative 3d configuration of the body and not its position in 3d space. The reason for this is practical, or reasoning about the position in 3d space would require prohibitory large training datasets that span the entire 3d volume of the space visible from the camera.

2.3.4 Post-processing:

Many algorithms, comprising generative and discriminative methods, does not have any relational constraint over the final output. In other words, the techniques which predicts joint coordinate using an input image, not having any filter to discard the false human pose.

To cope with this, there exist a set of postprocessing algorithms, which rejects unnatural human poses. The output pose from any Pose Estimation pipeline is passed through a

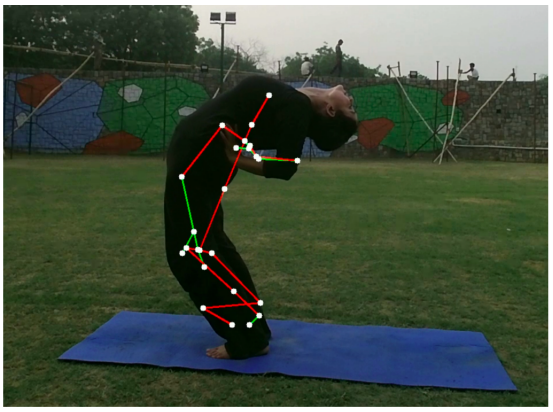

FIGURE 2.10: False human pose.

learning algorithm which scores every pose based on its likeliness. Poses that get scores lower than a threshold are ignored during the testing phase.

2.4 Deep learning based human activity recognition techniques:

Recent times possess the fast growth and progress of deep learning that reaches unmatched good performance in several fields such as natural language processing, visual object recognition, and logic reasoning [119]. Varying from conventional Pattern Recognition systems, deep learning can mostly reduce the effort to create features and easily get many added extra meaningful and high-level features by guiding an end-to-end neural network. We divide the HAR survey on the basis of differnt deep models:

2.4.1 Convolutional neural network

CNN has two benefits over other methods: scale invariance and local dependency when CNN is used in the time series type of classification methods like HAR. Scale invariance indicates the invariant of scale for different frequencies and poses, and Nearby signals in HAR are expected to be more correlated. Because of the robustness of the CNN network, the largest of the practice centered on this field. Recently cho et al. [120] proposed a technique where a one-dimensional CNN had used for HAR, which applies a divide and conquer oriented classifier method. Munzner et al. [121] discussed the problems of fusion of multimodal data and its normalization for HAR using CNN.

2.4.2 Autoencoder

The Almaslukh et al. [122] used a very well-familiar deep learning strategy identified as stacked autoencoder to improve the HAR accuracy and lower recognition time. As the basis for the HAR model, Balabka et al. [123] chosen the Adversarial Autoencoder and employ CNN for feature extraction.

2.4.3 Recurrent neural network

Singh et al. [124] proposed a deep learning algorithm to do HAR without utilizing some preliminary information. For this purpose, an LSTM RNN was applied.Mutegeki et al.

[125] introduced a holistic deep learning-based HAR method with CNN-LSTM. This method increases the recognition accuracy of human activities.

2.5 Benchmark and Evaluation metrics:

2.5.1 For human pose estimation method:

Because of the large difference present in various pictures, it is tough to create a generalized dataset for examine the outcome of HPE. Instead of it, the researchers have generated many datasets to analyze their introduced procedure for a certain task. This makes the comparison of the various techniques even more difficult.

The Buffy [126]dataset is collected from the TV clips having 748 images and 5 sections. The MPII Human Pose [127] dataset, cover various human activities with full-body and upper body PE. In addition to these most commonly used datasets, other records are gathered for different tasks. For example, the area of the image that has human is selected using PASCAL VOC [128]to analyze the HPE in an unconstrained situation. Most sports actions are difficult and variable; therefore they are continuously utilized in many benchmarks of HPE like UIUC stickmen dataset [129] and Leeds sports dataset [130].

In comparison to these above-mentioned full body-oriented datasets, Knapp et al. [131]suggested the FLIC dataset. Cherian et al. [132]proposed the "Pose in the wild" dataset for the poses in the wild scene.

Because of advances in HPE, there are many 3D HP datasets, like HumanEva [133]

[134]was proposed, which had gathered the video using multi-view cameras and contains many subjects. The dataset has two parts (I and II) with different properties. It contains 56 videos, 4 subjects, and 6 actions.

Lonescu et al. [135]proposed a 3D dataset called Human3.6M, which have 11 subjects and total 1376 videos. The subject contains 15 actions in total.

2.5.1.1 Evaluation metric:

1. **2D PE metrics:** For HPE from images, containing single and multiple humans, uses the demanding metric called percentage of correctly estimated parts (PCP) [126] [136] . The metric mainly depends on the limb detection rate. The correct limb detection is defined as the distance between the predicted and actual limb joint locations are at most half of the ground truth limb length. The PCP curve is drawn by varying the percentage of overlap in between 0.1 to 0.5.

   Recently, percentage correct keypoints (PCK) is mostly utilized to calculate the joint accuracy. The joint prediction is correct if it lies between the area around the radius of X from the ground-truth joint. Andriluka et al. [127] presented a modified version of PCK called "PCKh". Another metric is percentage of detected joints(PDJ). The PDJ is defined as, the detected joints have been taken correct if the distance between actual and predicted joint is within the certain fraction of torso daimeter.

2. **3D PE metric:**

   Sigat et al. [133] proposed the 3D Error(ϵ) metric, that is the mean square distance

in between the limb ends and a virtual marker associated to the joint centers.we also call it MPJPE (Mean per joint position error). The Per Joint Position Error(PJPE) is measured as a Euclidean distance between the predicted and ground-truth of a joint. MPJPE is the mean over PJPE for all joints.

$$\varepsilon(u, u') = 1/x \sum_{i=0}^{x} \|m_i(u) - m_i(u')\| \tag{2.1}$$

u: ground truth poses, u': estimated pose, x: no. of virtual markers.

The MPJPE has also calculated by alligning the root joint of the gound-truth and predicted 3D pose. The expression for alligned MPJPE is:

$$MPJPE_{allignd} = \frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N} \|(m_i(u) - m_{root}(u)) - (m_i(u') - m_{root}(u'))\|_2 \tag{2.2}$$

T is the number of samples and N is the total number of joints.

PCP is also utilized as a 3D evaluation metric [137], in which the correctly estimated part is detected using the formula given below:

$$\|y_e - y_e{}'\| + \|z_e - z_e{}'\| /2 \leq \gamma \|y_e - z_e\| \tag{2.3}$$

y_e : ground truth 3D coordinate of end points of part e

z_e : ground truth 3D coordinate of start points of part e

$y_{e'}$ and $z_{e'}$ is the corresponding estimated result and γ is a parameter used to control the threshold.

Mean Joint Angle Error(MJAE) is the evaluation parameter used to measure the predicted joint angles in degrees, here the mean (over all angles) absolute difference between the actual and predicted joint angles in degree, as shown below:

$$MJAE = \frac{\sum_{i=1}^{N} |(z_i - z_i' mod \pm 180°)|}{N} \tag{2.4}$$

N is the total no. of joints. $z_i^{'}$ and z_i is the ground-truth and predicted pose.

2.5.2 Evaluation metrics and datasets for HAR

Researchers have generated many datasets to examine the outcome of HAR. In the below paragraphs, we discuss details of few publically available datasets.

Cornell activity datasets (CAD-60) This dataset has 60 RGB videos. It has 4 main subjects: two female, two male, and one left-handed. These videos are captured in five different environment settings like kitchen, bathroom, office, bedroom, and living room. The subjects perform brushing teeth, rinsing mouth, wearing contact lens, drinking water, talking on the phone, opening pill container, cooking (stirring), cooking(chopping), relaxing on couch, talking on couch, working on computer, and writing on whiteboard.

CAD-120 This dataset has 120 RGB videos. It has same subjects as CAD-60. The 20 high level activities and sub level activities are taking medicine, making cereal, unstacking objects, stacking objects, picking objects, microwaving food, cleaning objects, arranging

objects, taking food, having a meal, moving, reaching, eating, pouring, opening, drinking, closing, placing,and scrubbing.

UTD-MHAD It is a multi-modal HAR dataset obtained by one wearable inertial sensor and one Microsoft Kinect camera. This dataset comprises 27 activities presented by eight different subjects (4 males and females), with all subjects doing all actions four times.

Many state-of-the-art techniques utilizes accuracy, recall, precision, and F1 score. Here TP, FP, TN, and FP denotes the true positive, false positive, true negative, and false positive.

$$Recall(R) = \frac{TP}{TP + FN} \tag{2.5}$$

$$Accuracy(A) = \frac{TP + TN}{TP + FN + TN + FP} \tag{2.6}$$

$$Precision(P) = \frac{TP}{TP + FP} \tag{2.7}$$

$$F - measure = \frac{2 \times R \times P}{R + P} \tag{2.8}$$

2.6 Comparative analysis of recent state-of-the-art techniques for human pose estimation.

A comparative analysis for both 2D and 3D cases (shown in 2.3 and 2.4). The brief description contains, the year of publication, author, work, dataset, evaluation metric, merits are highlighted. A few papers are considered for the comparison.

TABLE 2.1: Comparative analysis of state-of-the-art techniques on the basis of type of preprocessing, feature description, body model, method, dataset, and metric they use.

Year	Paper	Preprocessing			Feature Description		Body Model	Method		Datasets	Metric
		IC	HD	BS	Hand Crafted	Deep Learning		G	D		
2011	Stoll et al. [46]	✓	✗	✓	✓	✗	Kinematic skeleton	✓	✗		
2018	Rhodin et al. [52]	✓	✗	✗	✗	✓	Skeleton	✗	✓	Human3.6M, MPII-INF-3DHP	MPJPE, PCK
2017	Pavlakos et al. [50]	✓	✗	✗	✗	✓	Skeleton	✗	✓	KTH Multi-view Football II, Human3.6M	3D joint error PCP
2017	Fang et al. [138]	✗	✓	✗	✗	✓	Skeleton	✗	✓	MPII dataset,	Average accuracy
2020	Zhao et al. [70]	✗	✓	✗	✗	✓	Skeleton	✗	✓	MSCOCO 2016 dataset MSCOCO 2017,	in terms of mAP mAP
2020	Gartner et al. [139]	✗	✗	✗	✗	✓	Mesh	✗	✓	PoseTrack test dataset CMU Panoptic dataset	Per joint per reconstruction error
2018	Guler et al. [140]	✗	✗	✗	✗	✓	Mesh	✗	✓	DensePose-COCO	Ratio of correct point correspondance, Geodesic point similarity
2019	Sun et al. [141]	✗	✗	✗	✗	✓	Skeleton	✗	✓	COCO Dataset, PoseTrack, MPII Human Pose dataset.	mAP, PCK@0.5
2019	Pavllo et al. [142]	✗	✗	✗	✗	✓	Skeleton	✗	✓	Human3.6M, HumanEva-I	MPJPE, P-MPJPE, N-MPJPE

2.7 Challenges

2.7.1 2D Human Pose Estimation

- Because of the huge disparity in human poses, till now, there is no generalized model for pose representation.

TABLE 2.2: List and Details of largely utilized datasets for HPE.

Dataset	Size	Color/Gray	Type	Dim
Buffy [126]	Training: 472 frames Testing: 276 frames	720*405 Color	Upper Body	2D
PARSE [143]	Training: 100 images Testing: 205 images	Different Sizes Color	Full Body	2D
LSP [130]	Training: 1000 images Testing: 1000 images	Different Sizes Color	Full Body	2D
FLIC [131]	Training: 3987 images Testing: 1016 images	720*480 Color	Full Body	2D
PASCAL VOC [128]	Total 47186 images 110008 objects	Different sizes color	Upper Body	2D
MPII Human Pose [127]	410 activities 25000 images	Different sizes color	Full Body	2D
Pose in the wild [132]	30 sequences 900 frames	Different sizes color	Upper Body	2D
Human 3.6M [135]	1376 videos, 11 subjects	Different sizes color	Full Body	3D
HumanEva-I&II [133]	56 videos 4 subjects	Differnt sizes color	Full Body	3D

- The self and background occlusion leading toward a big confusion in 2D HPE.
- Many part-based methods use the sliding window approach for detecting the body part position from an image. It also requires a much amount of time for the execution. Also, not able to solve the body occlusion issue for HPE.
- Greatly varying the background scenario, illumination condition, scale, and complex body part appearances make HPE a difficult task.
- Along with, cameras shooting angle, vertical angle and vertical projection technique of imaging create more deformation in human body appearance.
- Part based method require expected pose let area should be very accurate.
- Learn to give a good result when a good amount of annotation is not present.
- Most of the methods do not handle multiple and non-frontal subjects.

TABLE 2.3: Comprative analysis of state-of-the-art for 2D human pose estimation

S. No.	Year	Author	Work	Datasets	Evaluation Metric
1.	2019	Nie et al. [144]	Proposed a novel technique called hier--archical contextual refinement networks (HCRNs) for HPE by dividing the task into many layers on the basis of the complexity of the joints.	LSP, FLIC, MPII Human pose single-person and MSCOCO dataset	PCK and PCP
2.	2018	Kawana et al. [80]	The ensemble of models using CNN, each optimized for the certain number of poses, and can model a variety of configurations for the human body.	FLIC, BBC , LSP, MPII dataset	PCP, PCK and PDJ
3.	2017	Witoonchart et al. [101]	Structured support vector machine is formulated for the layers of CNN. The first layer is augmented inference layer and the second is the normal convolutional layer.	PARSE dataset	PCP
4.	2017	Jammalamadaka et al. [145]	Proposed two technique for HPE, the first one is deep poselet, in which CNN is utilized to make the pose representation appropriate for the pose search and the is deep pose embedding technique where an optimized neural network is used to map the image to some dimensional space contains the like poses close and unlike poses with more distant.	H3D, PASCAL, Movie dataset, FLIC, MPII human pose, pose in wild	Precision-recall curve,pose search performance (MoP)
5.	2017	Ai et al. [146]	Propose a new technique by including the structure pose prior to learn the CNN for HPE rather than using features from CNN.	FLIC and LSP	PCP and PDJ
6.	2017	Marras et al. [147]	Presented a new method to estimate the accurate heat-map based joint location using cascaded scenario, which contains a combination of coarse, part-based spatial model and fine-scale CNN.	Fashion pose, MPII, LSP	PCK
7.	2016	Yang et al. [148]	Introduce a new HPE framework which has the integration of expressive mixture of parts model with DCNN. Then include the domain prior knowledge in the module.	LSP, FLIC, PARSE dataset	PCP, PDJ
8.	2016	Bulat et al. [149]	Give a new cascaded architecture called detection-followed-by-regression CNN for HPE. Where the detection part gives the part heatmaps and the other regresses these heatmaps.	LSP and MPII	PCKh
9.	2015	Zhao et al. [150]	Proposed a deep structure for HPE, that emphasis on to produce an efficient body part detection and make spatial constraint in between the parts.	LSP, FLIC and PARSE dataset	PCP
10.	2015	Duan et al. [151]	Give a hierarchic multi-layer, tree oriented part based technique for HPE, where at the top layer, the model was used for body detection and at below layer the body is fragmented into many parts for estimation.	UIUC sport, Leeds sports pose, and FLIC datasets	PCP
11.	2014	Cherian et al. [152]	Introduces a two-step method for HPE, the first step considers the temporal-links of the consequent frame for body parts like elbows and wrists. Another one, mix the body-part sequence for the estimation purpose.	VideoPose and MPII	Key point localization error
12.	2014	Toshev et al. [108]	Give a Deep neural network (DNN) oriented regression-based technique to regress body joints correctly. Also presented the cascading module of the given regressor.	FLIC, LSP	PCP, PDJ

TABLE 2.4: Comprative analysis for 3D human pose estimation

S. No.	Year	Author	Work	Dataset	Evaluation metric	
1.	2020	Zhang et al. [153]	Proposed an novel adversail learning scheme, that learn invariant HP latent using 3D annotations to estimate the HP from monocular images along with 2D annotated dataset.	Human3.6M, MPII	MPJPE	Viewpoint invariant.
2.	2018	Kanzawa et al. [154]	Proposed method recovers the 3D joint angles and shape of the human body using RGB image using convolutional encoder, iterative 3D regression technique.	MPI-INF-3DHP	MPJPE/ reconstruction error	Minimize the reprojection loss of key points, which allow training the model with in-the-wild data which have only 2D annotation ground truth data. Does not use any paired 3D data.
3.	2018	Pavlakoset [50] al.	Presented a technique using different settings of 3D HPE. First, they predict only the depth of the human joints. Then, combine the ordinal relation with 2D keypoint annotations to predict the 3D pose coordinates. Another part explains the interpretation of these relations within a volumetric representation of 3D HP.	Human 3.6 M, HumanEva-I, MPI-INF-3DHP	MPJPE	Provide an Automatic way to gather 3D annotations.
4.	2018	Huang et al. [155]	They extend the SMPLify to multi-view: Estimate 2D joints using Deepcut and segment the body out from the background using CNN. Apply DCT oriented prior to solving the leg swap issue in HPE.	HumanEva	3D joint error	By extending SMPLify and making it multi-view results in having more knowledge regarding human body. Incorporating silhoulttes effectively enhance the estimated shape and 3D pose accuracy.
5.	2017	Jahangiri et al. [156]	The CNN-oriented detector is used to provide the heatmaps of 2D joint locations, 3D torso and projection matrix. Using these many 3d poses hypotheses was generated.	Human3.6M	MPJPE	Doesn't trouble with the dataset bias. Due to the compositional property, handles the different body shape, clothing and viewpoint problem very well.
6.	2017	Tekin et al. [105]	Proposes two stream architecture, the first stream give the heatmap for 2D joint locations and second is image stream extract features and fuse all these for 3D pose vector.	Human3.6M	The 3D joint position error	The framework is common and easily utilize to extend it for other modalities.
7.	2017	Mehta et al. [83]	Over a full-size image, the person-centered crop is efficiently extracted by bounding box tracking. Using the crop, and CNN based model, they predict the heatmaps for 2D joints. The heatmaps are utilized for reading off the 3D poses.	MPI-INF-3DHP	PCK, AUC, MPJPE	Does not require 3D annotation for learning.
8.	2017	Rogez et al. [157]	Extracted the object region by RPN network and give pose proposal for the certain number of the anchor-pose proposal in these regions. Then proposals are scored through classifiers and regress by a regression for PE.	Human3.6M	Average 3D pose error	The method restores complete body pose, despite the fact that the person is partially occluded.
9.	2017	Martinez et al. [114]	The proposed system has a linear combination of batch normalization, dropout, and RELU activation. They give a score of joint location position in 3d.	Human3.6M	3D error	
10.	2016	Yasin et al. [12]	The method utilizes two training sources. The first one is from motion capture dataset, having 3D pose. The second is 2D annotated image dataset. The method projects the 3D pose to 2D and regression technique is utilized from 2D annotation.	HumanEva-I, Human3.6M	3D error	Collecting large amounts of training data that contain unconstrained images and are annotated with accurate 3d pose is infeasible. Therefore, they propose to use two independent training sources.
11.	2015	Li et al. [102]	Takes input as image and 3d pose. The output is score matching value. ConvNet for image feature extraction with two sub-networks for transforming features and pose into joint embedding.	Human3.6M	MPJPE	It insures a large margin between the score values for correct input pairs and for incorrect input pairs.

2.7.2 3D Human Pose Estimation

- 3D HPE is an ill-posed problem, particularly by using the single monocular image. Due to 3D information loss, same image projection is derived from different 3D objects. In these situations, self-occlusion is common to occur, which lead towards the ambiguity and make the system less efficient.

- Variation of human pose in images or video, when human performs a complex task like gymnastic actions.

- Multiple 3D HPE, where a person is interacted with another person and to the environment. It becomes difficult because of self-occlusion and limb occlusion using self and other objects.

- We are not able to incorporate the 3D HPE technique in a real-time application, if it is not performing efficiently for outdoor environment, with different lighting condition and varying background, along with unconstrained behavior of the subject.

2.7.3 Human Activity Recognition

- The extraction of key feature-based information from the sensor data is challenging.

- Most studies have utilized optical flow frames, RGB frames, skeleton images, dynamic images, depth maps separately, whereas some methods combine them to improve the performance.

- Many state-of-the-art techniques on HAR are view-dependent and handle the recognition problem from one determined view. To do HAR from various views has been considered a challenging job for researchers and requires much improvement.

- The trajectory of the actions from diverse viewing directions are distinct, and few of the body segments (part of lower part of leg, hand, part of the body, etc.) are occluded because of the change of view.

Despite much progress in the field, pose estimation and activity recognition remains a challenging and largely unsolved problem. After doing the review of all the challenges,

we observe that the progress of most research work has been made towards the 2D and 3D HPE from a single image and HAR from video, which is a severely under-constrained problem.

2.8 Conclusion

This chapter studied the state-of-the-art approaches in the disciplines of HPE and HAR. It also examined the challenges and issues of these domains. This chapter also discussed the benchmark databases utilized to perform the performance evaluation of the proposed methods. Finally, the evaluation metrics were also explained, which are used to do the performance evaluation.