DeepPoseNet: A Comprehensive Study on Human Pose Estimation with Deep Learning Technique

Syed Mujtaba Haider[1], Abdul Basit Mughal[2], Rafi Ullah Khan[3]
[1]Department of Computational Mathematics, University of Pavia, Pavia, Italy, [2]Department of Computer Science, Bahria University, Islamabad, Pakistan, [3]Department of Electrical & Computer Engineering, Pak Austria Fachhochschule: Institute of Applied Sciences and Technology, Haripur, Pakistan

Email: rafi.ullah@paf-iast.edu.pk, syedmujtaba.haider01@universitadipavia.it, sardarbasitmughal344@gmail.com

Corresponding Author: Rafi Ullah Khan (rafi.ullah@paf-iast.edu.pk)

Received: 19-12-2023; **Accepted**: 14-04-2024; **Published**: 10-05-2024

***Abstract*:** The article is a thorough analysis on the deep learning techniques used in image processing to estimate human poses. This entails examining several essential architectures such as CNNs and why traditional methods are unfit. It clarifies attentive mechanisms and transfer learning parts. This approach uses a two stage CNN model, whereby first network identifies some body parts, while the other focuses on these identified body bits. We use an intricate VGG16 to pinpoint body parts with accuracy. These models are compared using benchmark data sets and performance measures of special interest in the application of the MPII dataset for model training as well as verification. Deep pose estimation has huge social and economic consequences. They include human-computer interaction, sports analysis, healthcare, and many more. Conclusion gives an outline of important insights made above, highlighting positive aspects identified as well as gaps that require additional research including call towards cooperation between disciplines for enhanced growth in this field.

Keywords: Pose Estimation, Deep Learning, Convolutional Neural Networks, Transfer Learning, Spatial Configuration, Computer Vision.

1. INTRODUCTION

Human pose estimation is a key component of the broad efforts of trying to deconstruct the complex relationships between spatial arrangements in pictures and films. This complicated procedure specifically pinpoints the vital skeletal joints in the human anatomy, giving insights to the human movement and object interactions. Pose estimation reaches beyond just its direct application in robotics, augmented reality, sports analytics, healthcare, etc. to the very heart of understanding human behavior and communication with a machine. In this context, it is important to note that in the field of conventional methods for pose estimations depended on handcrafted feature and heuristic algorithm usually proved to be insufficient to overcome the difficulties that were connected with a complicated environment. There were occlusions, diversity in views, and complexities associated with articulated structures that required a new way of thinking. The appearance of the transformative deep learning techniques of CNN's means going beyond the limits of traditional approaches.

In this respect, Deep Learning is introducing a completely new approach to solving problems based on an algorithmic approach whereby the neural network can automatically discover high-level abstract properties in raw data. Such nuance is highly sophisticated for recognition and has its place, among others, for CNNs having convolutional layers and feature hierarchy construction. However, this shift is not just about increased precision but rather strengthening the stability and capacity of pose estimation models, initiating a new wave of opportunity. Deep Learning Methods for Pose Estimation in current landscape. Its path traverses elaborate architectures, creative techniques, core elements that have taken the domain to unexplored avenues of precision as well as relevance. In our effort to provide penetrating observations which could

function like the guidance of light for further progress, we will examine the advantages and disadvantages associated with the existing strategies.

Pose estimation is now the way forward in improving computational vision techniques and our comprehensive exploration of the new approaches is set to make a difference across a range of domains. The exploration of image search in new directions extends well beyond just the immediate impact on the technology domains involved. Instead, it enhances our ability to better understand human engagement with images as well as how we see and relate in ways that have never before been seen or attempted. This leads us into an entirely.

2. LITERATURE REVIEW

Computer vision fundamentally comprises pose estimation which entails accurate specification of object arrangement in images or video in relation to a particular field. The objective of this review is to examine a number of poses from the conventional methodologies to the most cutting-edge deep learning techniques and to evaluate both two dimensional and three-dimensional approaches.

Initially, methods were heavily dependent on manually created features as well as intuitive algorithms. In 2004, [1] presented a game-changing technique that was based on pictorial structures with graphs modeling dependencies of parts involved. This work served as a foundation building block for future research, providing a rigorous framework of pose estimation within the computer vision domain. Deep learning has led to a total transformation in how poses are estimated. In 2004 [2] presented a custom deep neural network architecture designed specifically for pose estimation that outperformed conventional approaches. CNNs have been playing an important role since then, as they are able to obtain hierarchical features as well as spatial relationships [3]. Notable is an innovation that led to posing machines. In 2016 [4] presented a Convolutional Pose Machine. This model can achieve extremely high accuracy by iterating it repeatedly through different rounds of refining poses predictions. The iterative process of iterative refinement works for almost every architecture; therefore, it is an essential part of powerful and resilient pose estimation approaches. Transfer learning has proven to be one of the most essential methods in dealing with scarcity of datasets. In 2021 [5] proved that it is possible to use the previously trained networks to solve large-set images and to transfer this knowledge in a poses estimation domain. This strategy has been instrumental in enhancing the generalizability skills of pose estimation models.

Pose estimation has recently come into the domain of three-dimensional (3D).In 2017 [6] proposed a volumetric method of estimating 3D poses based on the regression of volumetric heat maps. This is a major achievement for augmented reality, robotics, and human-computer interaction. Finally, pose estimation research has undergone a revolution from traditional methodologies to advanced multi-layered neural networks. Integration of CNNs, Pose Machine, and Transfer Learning have improved accuracy and practicality. The development of pose estimation research has led to a number of ongoing investigations which would in turn offer improved scope into diverse field applications. In the realm of human gait identification, the initial phase involves pose estimation, a process facilitated by our proposed system founded on deep neural networks. Human pose estimation is approached as a deep neural network-based regression problem, focusing on the joints of the body. The cascade regression model, employed in this system, boasts high accuracy. One of its key advantages is its ability to provide comprehensive reasoning about the human pose. The primary objective of human pose estimation is to identify and estimate the structure of the human body. Currently, it stands as a predominant and extensively researched topic.

The methodologies for human pose estimation encompass a wide array of approaches [7]. Diverging from marker-based methods, this system employs tracking-based approaches. Although marker-based methods, utilizing specific markers attached to the human body, remain popular, this paper concentrates on monocular vision-based techniques. In these methods, markers, foreground-background segmentation, or temporal tracking mechanisms like extra means are not utilized. The hierarchical division of the human body is a fundamental aspect of this approach. Our proposed model, rooted in convolutional neural networks (CNNs), predicts 2D human body poses from images. This model generates a heat map for each body key point, allowing it to learn and characterize both the part appearance and the context of the part configuration. The model comprises three integral parts. In the first part, a feed-forward architecture is combined with a recurrent module, enhancing the overall system's performance. The second part facilitates end-to-end training from scratch, and auxiliary losses are computed post-training to enhance accuracy and performance. The prediction of key point visibility is the third step. Running the feedforward network several times, a heat map is produced. While the feedforward module plays an important role as part detector, it provides no contextual information from other parts because its receptive field is smaller. A regression network regulates the feedforward network [7].

The model proposed in [8] is a CNN-based system for prediction of 2D human body poses from images. Apparently, not only the appearance of a part arrangement but also its context is learned and described by this model. Body key points

are efficiently generated to produce a heat map. The model is split into three major components. Feed-forward architecture is seamlessly combined with another recurrent module in an early section. The latter definitely enhances the system as a whole. This model can then be trained from scratch end to end in the second step. After training, auxiliary losses are calculated to fine-tune the accuracy and overall performance of the model. In the final segment, the model calculates the prediction of key point visibility.

The feedforward network is executed multiple times, producing a heat map as its output. The feedforward module functions as a part detector, generating a key point heat map. However, it operates without awareness of context from other parts due to smaller receptive fields. The feedforward network, based on a regression network, incorporates modifications for the initial convolutional layers, utilizing smaller filters (3x3) combined with nonlinear activation functions. Pooling layers are applied twice to achieve an output heat map resolution that is appropriately large. The Rectified Linear Unit (ReLU) activation function is applied after every convolution and on the prediction layers, with the output taking the form of a heat map. The recurrent module in our network specifically applies to layers 6 and 7. At each stage, the input of the recurrent module is fused with the outputs of layers 3 and 7. It's important to note that the input of layer 3 remains fixed throughout, while the input of layer 7 is dynamically updated as needed. In the concluding stages of our network, comprehensive end-to-end training becomes possible. Moreover, body parts' heat maps are constructed from sets of key points. A body part heat map is defined by establishing the midpoint between two key points as the center of the Gaussian distribution. The variance is determined based on the Euclidean distance between these two key points. Heat maps are then generated for both body limbs and key points. Typically, heat maps represent body joints, with body part heat maps predominantly capturing limbs.

The network takes an RGB image as input, with a resolution of 248x248, and produces output heat maps at a resolution of 62x62. Prior to input, the image undergoes normalization through mean subtraction across each channel. Following image normalization, data augmentation is performed, encompassing rotations, scaling, flipping, and cropping of the input image. For training and testing, the MPII Human Pose dataset is utilized, providing a robust foundation for evaluating the performance and generalization capabilities of the proposed model. In [9], the author presents an effective supervised mechanism designed to comprehensively capture the structure of human poses from a sequence of images. The proposed method comprises two dual learning components, each focusing on a distinct aspect: 2D to 3D pose transformation and 3D to 2D pose projection. This dual-part approach serves as a crucial bridge between 3D and 2D human poses, enabling accurate prediction of 3D human pose estimation.

The model's primary objective is to predict 3D human poses based on input images. The 2D to 3D pose module is responsible for predicting human joints in the form of 3D coordinates, while the 3D to 2D pose projector retrieves the 2D pose through regression operations. Parameters for pose estimation from 2D to 3D are applied consistently across all frames, preserving temporal motion coherence. The 2D pose subnetwork encodes each frame within a monocular sequence, encompassing comprehensive information about pose estimation, such as human body shape. The shallow convolution layers are used to extract low-level information, basic image representations of humans. The architecture of the 2D pose includes convolutional pose machines, whereby when given an image as input, the network produces feature maps and then out comes a 2-dimensionalpose vector. The aim of this integrated method is to increase the efficiency and stability with which human poses are estimated by linking together 2D-to-3D transformation and 3D-to -2 D projection. The 2D-to-3D Pose Transformer Module is intended to develop one's abilities at predicting a human in three dimensions using information extracted via the 2D pose subnetwork. These features are first passed through convolutional and then fully connected layers. Each layer of the former uses 128 different kernels, with a max-pooling layer used between them. Features from these layers are then fed into fully connected layers, generating a feature vector. Features extracted from the 2D pose architecture are converted into a vector of length 1024. In succession, these feature vectors are input to LSTM for 3D pose sequence prediction. Separately, the 3D-to-2D Projector Module is made up of a series of connected layers with ReLU as the activation function and batch normalization operations. In the first few fully connected layers, a regression function is defined between middle 3D pose predictions and final pose prediction. The rest of the fully connected layers are for the projection function. The Human3.6M dataset and the Human Eva-I are used for training and testing respectively. The Human3.6M dataset provides 3D human pose images totaling more than 3.6 million pictures, which depict the same set of correlations between various viewpoints and the front (or back) views across different scenarios for each actor from a group of professional actors who all performed similar actions in order to show how emotions affect body poses during tense emotional states. The dataset has been divided into three sets, with five subjects each for training and two for testing.

At the same time, the Human Eva-I dataset is made up of video sequences from four different subjects depicting humans performing simple actions like walking, jogging and boxing. It also provides annotations of 3D poses for each frame in the video sequences, making it easier to train a model with greater generalization ability across different actions and subjects. The paper in [10] deals with 3D human pose prediction, which already faces obstacles due to joint self-occlusion

and poor generalization. The network under consideration has a novel scheme. It expresses 3D poses as directed graphs and uses graph convolution to boost the accuracy of prediction. Training and testing is done on the Human 3.6m and MPI-INF-3DHP datasets respectively. Establishing relations among joints is the objective of four components to overall network architecture. Human poses are conceived as a directed graph, with the initial module being pose regression. The stacked graph convolution layers employed in this module overcome the challenges of joint self-occlusion and produce accurate 3D estimates.

Second, 2D and 3D pose features are concatenated. This fusion process helps more precise and consistent estimation of 3D poses. The third module is for predicting the projection matrix when there is no ground truth, giving flexibility to the model. Last, the final module serves as a discriminator that learns to discern between 2D ground truths and contribute towards refining of predictions. The Human 3.6M dataset is used for training and assessment, which has one of the largest datasets in human pose estimation collection field. The dataset is made up of 3.6 million RGB images, divided into actions belonging to one of 15 different scenarios. The introduced dataset contains 11 subjects performing activities such as walking, eating and sitting (sitting is half upper body with both hands free). Some of the images are taken within a motion capture system for better training and evaluation of the proposed model.

3. METHODOLOGY

Human pose estimation from images and videos is one of the most challenging tasks in the field of Computer vision. The complication is in dealing with an extremely wide range of human poses, and allowing for changes, Variations in clothing and foreshortening related to human form, or the handling of situations where many people take part in proximity. To deal with these problems, CNNs are used. These networks provide robust low and mid-level appearance features which capture the relevant information crucial for accurate pose estimation. In this context, the use of a CNN-based cascade architecture is especially effective because, being designed for the study of fine-grained part relations and solid inference of pose, it is robust even when there are drastic change's part occlusions. The cascade architecture itself is divided into two parts. The first portion is for creating a heat map as output, representing the key points of a pose in heat maps. Afterwards, the second part regressions refine the pose estimation on these heat maps. This architecture has the advantage that it can focus the network on important points within an image, refining pose estimation Overall, the use of CNN-based cascade architecture shows a direct and effective way to deal with the complexities of human pose estimation from visual data.

In body parts detection for pose estimation, a fine-tuned VGG16 model is used to improve overall performance of the system. The main output of this body part detection process is a heat map. Thus, this is the input to the next stage of the model which refines with precision both location and size as shown in figure 1. For training and testing purposes, the MPII dataset is used. This set of data then becomes a firm basis on which to train the model to accurately identify and point to different body parts. The dataset is divided into 70 %, which forms the training set and the rest 30 % is assigned to testing. Partitioning in this way ensures that the model is trained on a broad range of examples. It is evaluated on a different set of data, which helps to explain its good generalization ability for unseen instances. The fine-tuned VGG16 and MPII dataset not only reflect the increasingly popular trend of using pre-trained models, but they also show how important it is to choose them wisely large-scale datasets and models to achieve state-of the art performance in human pose estimation, an extremely difficult task. The splitting of the data set into training and testing sets makes it possible to give a comprehensive assessment as to how well the model is doing and generalization capabilities.

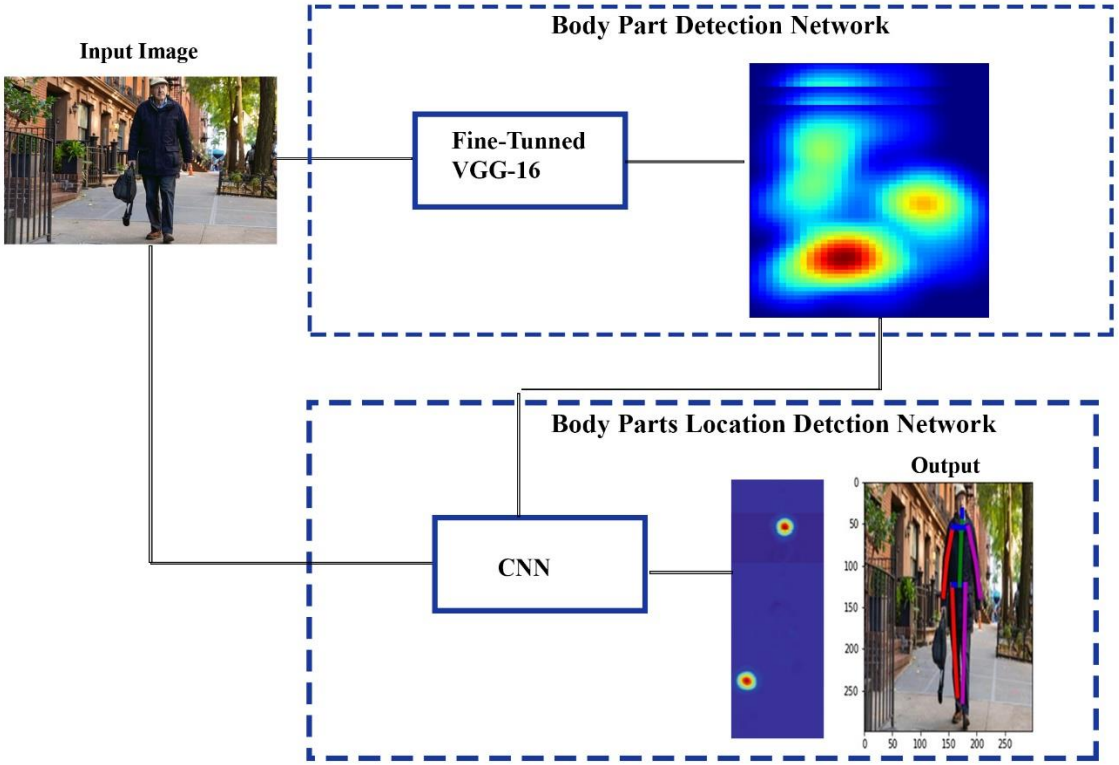

Figure 1. Workflow of Pose Estimation Network.

3.1 Body Parts Detection Network

In the pose estimation network, we used a VGG-16 fine-tuned for recognition of body parts. The Fully connected layers of VGG-16 were replaced with totally convolutional layers whose kernel size was 1. In human body parts detection, the output is a heat map. Because localization accuracy is insufficient, with a stride of 32 pixels, choosing to reduce it to 8 pixels.

Figure 2 shows the fine-tuned VGG-16 network used for part detection of human body parts. Initially, an image is passed through the finely tuned VGG-16 network. This process generates heat maps that accented parts of the human body. Next, these heat maps become inputs for another model responsible for locating body parts.VGG-16, is a very influential convolutional neural network architecture concerning its depth and simplicity. VGG-16, whose image classification performance has a structure with several layers of small convolutional filters. This leads to an efficient and accurate model. Human body parts detection with VGG-16. The fine-tuning process was one in which the network's parameters were refined to increase its powers of discrimination and accurate localization within images. To make full use of the network's power, we customized VGG-16 for our purposes and improved accuracy at identifying or delineating human body parts. It turns out that VGG-16 is flexible and adaptable. Proof of its value comes in tackling difficult visual recognition problems.

.

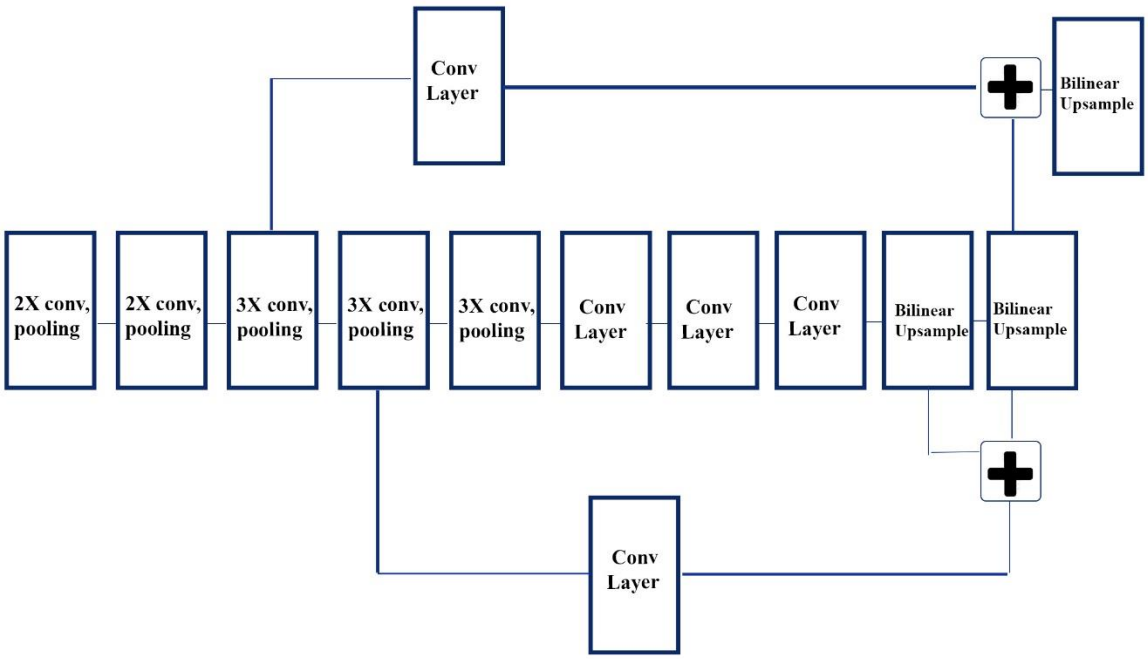

Figure 2. Fine-tuned VGG-16 Network for Body Part Detection.

3.2 Body Parts Location Detection Network

Our method is based on a Convolutional Neural Network (CNN) model. The input to the model includes both the heat map produced through body part detection, and also the original input image. This is the combined input that is fed through the CNN Network, this network made up of a number of convolutional layers and additional layers for feature extraction. In the refinement of fine features during the last three layers, the kernel size is fixed at 1, strategically enhancing the network's capability. This configuration is deliberate, ensuring a delicate equilibrium between breadth and detail in the network's representation. When delving into body part detection, the importance of attending to numerous details becomes evident.

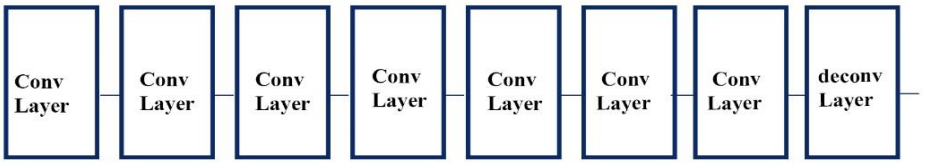

Figure 3.CNN Network for Body Parts Location Detection.

Figure 3 illustrates the architecture of the pose estimation network and the human body parts location detection model. This model is designed to highlight specific human body parts in the form of heat maps. Each heat map corresponds to a distinct body part, and these individual heat maps are seamlessly joined. These interconnected landmarks are then picked out and identified by being connected with different colors. The use of this visualization technique helps enhance the interpretability of the model's output, displaying a complete and colorful image representing detected human body parts.

3.3 Dataset

In our pose estimation network, we used MPII dataset for training and testing. The dataset is divided into 70 percent for training and 30 percent for testing. The first part in our pose estimation model is resizing the image to 300 pixels. After this every image is rescaled to 380×380 pixels resolution. Since these preprocessing steps provide a tiny amount

of standardization, we can at least say that our tests on whether the model works are slightly more consistent and reliable.

The MPII Human Pose Dataset as shown in figure 4 is considered one of the largest and most used datasets for human pose estimation models. Datasets consist of more than 25,000 images which are collected from YouTube videos covering a broad field of real-world situations. The dataset includes a lot of human motions and gestures, so it's excellent for training pose estimation models across different environments. Since every image in the MPII dataset is accompanied by exact information on 16 different joints, it becomes a rich source of real data for training models. Joints are annotated on the major parts of the anatomy, including head and neck; shoulders, elbows, and wrists; hips, knees. This is a very broad-ranging dataset, from different ages of people to all styles for clothes and backgrounds--we can only look deeply under such varied conditions. Its comprehensive, realistic image contents make it especially appropriate as a basis for evaluating the value of models in real-world applications. This is the time when researchers can use this dataset to compare and judge how well different pose estimation algorithms work. We can progress in computer vision [11].

Figure 4. Images from MPII Dataset [11].

3.4 Pose Estimation Model Algorithm

---

1. **Input Image:**
   - I_{input} Is the input image or video frame.
2. **Preprocessing:**
   - Crop the input image to a resolution of 380×380 pixels.
   - Apply any necessary color normalization or enhancement techniques.
3. **Human Body Parts Detection using VGG-16:**
   - Let I be the input image.
   - $H = \text{VGG-16}(I)$ Produces heat maps where H represents the heatmap tensor.
4. **Human Body Parts Location Detection using CNN:**
   - Let I' be the resized input image, and H' be the heatmap tensor.
   - $L = \text{CNN}(I', H')$ Outputs the locations of body parts, where L is the location tensor.
5. **Draw Lines on Landmarks:**
   - Let $\text{HeatmapDrawn} = \text{DrawHeatmap}(I, H)$ overlay heatmaps on the original image.
   - Let $\text{LinesDrawn} = \text{DrawLines}(I, L)$ connect landmarks with lines on the original image.

---

4. RESULTS

In the pose estimation network, the initial image is provided to the system, and subsequently, the image undergoes pre-processing. The input is cropped to a resolution of 380×380 pixels, followed by the application of necessary color normalization or enhancement techniques. The first network in the pose estimation model focuses on human body part detection. For this purpose, a fine-tuned VGG-16 network is employed, wherein the fully connected layers of VGG-16 are replaced with fully convolutional layers having a kernel size of 1. In the realm of human pose estimation, the output for body parts detection manifests as a heat map. Recognizing the inadequacy in localization accuracy with a stride of 32 pixels, a decision was made to adjust it to 8 pixels. The second network within the pose estimation model is dedicated to locating human body parts. The input to this model includes both the heat map generated from body part detection and the original input image. This composite input is then fed into the CNN Network, structured with a series of convolutional layers and additional layers to enhance feature extraction. Throughout the training and testing phases, the MPII dataset was utilized, with a partition of 70% for training and 30% for testing. This partitioning facilitated an evaluation of the model's accuracy in detecting human body parts and their respective locations.

Table 1 Hyper parameter tuning for Pose Estimation Network.

Learning Rate	0.00001
Epochs	30
Batch Size	08

In our pose estimation network, figure 5 presents the results; it shows how many joints and which ones were selected by us. The visualization depicts how accurately these joints have indeed been identified, representing in clear terms the model's ability to locate anatomical landmarks through pose estimation.

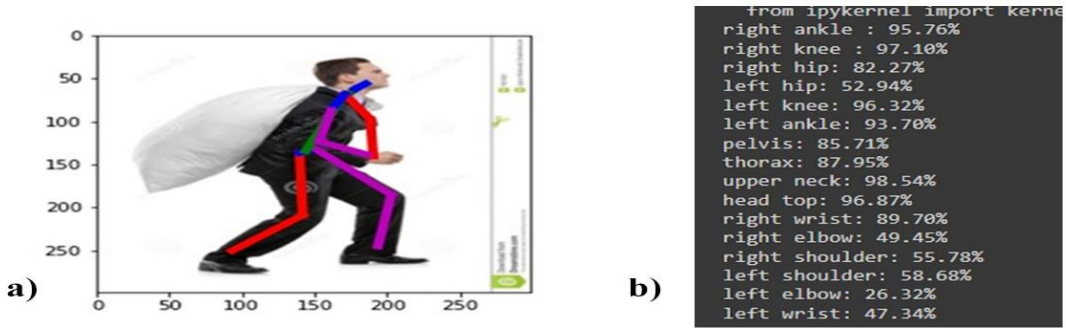

Figure 5. Pose Estimation results.

The varying accuracy of the different joints illustration in figure 6, demonstrating how accurately each of the 16 joints has been identified. Lines are formed on the input image based on the accuracy of joint identification, offering a visual representation of the model's precision in locating specific anatomical landmarks. This graphical presentation allows for a nuanced understanding of the model's performance, showcasing the effectiveness of joint identification through the formation of lines on the input image.

Figure 6. Result of Pose Estimation Network.

An evaluation using PCKh is shown in figure 7, for each joint on MPII shows a consistently high accuracy Level, implying that the majority of predicted key points are within an acceptable threshold distance from Their true locations.

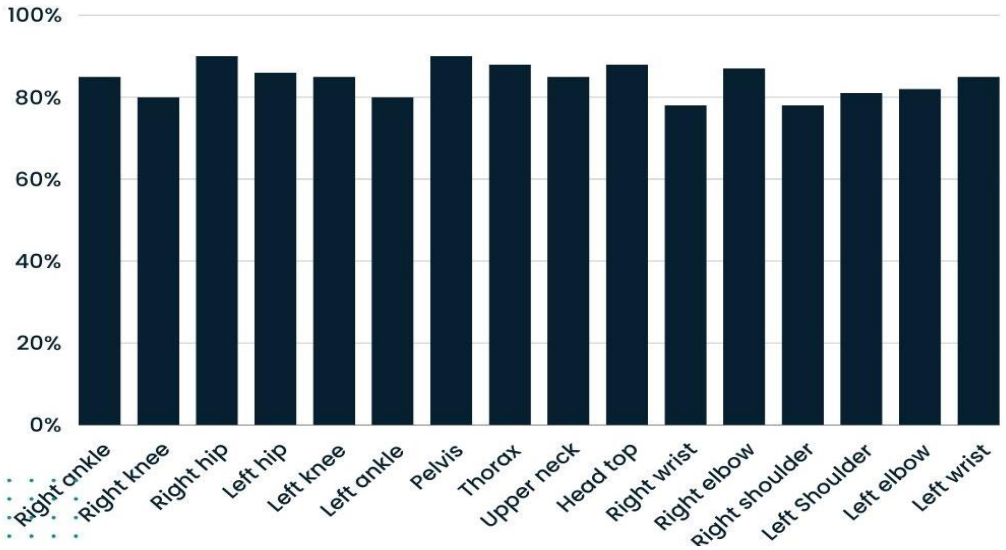

Figure 7. PCKh for each joint for our method on the MP II dataset.

In our model, the cascade CNN architecture provides excellent results compared to previous models. The CNN model generates a percentage of 16 joints, and the MPII dataset is used for training and testing. A fine-tuned VGG16 model is employed for body part detection, which detects the body parts and generates the heatmap. Other CNN networks detect the location and body parts, and the marker method is used to join the different joints with different colors. A comparison based on PCK on the MPII dataset demonstrates that the total accuracy of each joint exceeds 90 percent.

5. CONCLUSION

Human pose estimation is an important step on the way to understanding difficult relationships within images and videos. The complicated process involves pinpointing the important skeletal joints in human anatomy, AND ferreting out these clues about how we move our bodies through space gives us some kind of understanding into how people interact with objects. The MPII dataset is used for training and testing in our pose estimation model. We use a VGG-16 fine-tuned model, Using VGG-16 we can differentiate between human body parts with great precision and locate anatomical

landmarks The Fully connected layers of VGG-16 were replaced with totally convolutional layers The learning rate of our human pose estimation model is 0.00001 and batch size is 08. Our pose estimation model also uses a CNN network to precisely locate these body parts. The success of the model can be seen from its excellent accuracy in detecting 16 joints, shown especially on test data. Our model performs robustly in the human pose estimation task, as this strong performance testifies.

REFERENCES

[1] Agarwal, A., & Triggs, B. (2004). Tracking articulated motion using a mixture of autoregressive models. In Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part III 8 (pp. 54-65). Springer Berlin Heidelberg

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

[3] Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14 (pp. 483-499). Springer International Publishing.

[4] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., & Schiele, B. (2016). Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14 (pp. 34-50). Springer International Publishing.

[5] Lin, J., Wei, Z., Li, Z., Xu, S., Jia, K., & Li, Y. (2021). Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3560-3569).

[6] Pavlakos, G., Zhou, X., Derpanis, K. G., & Daniilidis, K. (2017). Coarse-to-fine volumetric prediction for single-image 3D human pose. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7025-7034).

[7] Singh, A., Agarwal, S., Nagrath, P., Saxena, A., & Thakur, N. (2019, February). Human pose estimation using convolutional neural networks. In 2019 amity international conference on artificial intelligence (AICAI) (pp. 946-952). IEEE.

[8] Belagiannis, V., & Zisserman, A. (2017, May). Recurrent human pose estimation. In 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017) (pp. 468-475). IEEE.

[9] Wang, K., Lin, L., Jiang, C., Qian, C., & Wei, P. (2019). 3D human pose machines with self-supervised learning. IEEE transactions on pattern analysis and machine intelligence, 42(5), 1069-1082.

[10] Rogez, G., Weinzaepfel, P., & Schmid, C. (2019). Lcr-net++: Multi-person 2d and 3d pose detection in natural images. IEEE transactions on pattern analysis and machine intelligence, 42(5), 1146-1161.

[11] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., & Schiele, B. (2016). Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14 (pp. 34-50). Springer International Publishing.