A survey on Pose Estimation using Deep Convolutional Neural Networks.

Manisha Patel*, Dr Nilesh Kalani
Research scholar, Electronics and Communication Department, School of Engineering, RK University, Rajkot, India.

Director, IQAC, RK University, Rajkot, India.

*m71074@ldce.ac.in

Abstract. Human pose estimation is a technique, which identifies the human body's landmarks in images and videos. Human pose estimation can be divided into single person pose estimation and multi-person pose estimation, also an estimated human poses in crowded places as well as in videos. Depends upon the application such as activity recognition, Animation, Sports, Augmented reality, etc., Pose estimation output can be in 2D or 3D coordinate format. 3D pose estimation is estimated considering joint angles in 2D. Some challenges like small and barely visible joints, strong articulations, occlusions, clothing, and lighting changes increase difficulty in estimating pose. Remarkable progress has gained in the field of human pose estimation using Deep learning-based CNN models. In this paper, we compare and summarized various deep learning models for pose estimation of a single person and multi-person.

1. INTRODUCTION

In computer vision, the main goal is to model and replicate the things which see by humans. To do so, there are interactions between humans and objects for such a high level of reasoning, the machine first has to interpret the posture from the given visual data. The task of identifying and determining the posture of a person in videos or images is reflected as human pose estimation [1]. Human pose estimation is an important and complex problem in a real-time application, such as sports, assisted living, assisted driver systems, games, etc. It is used to track a person's small body movement and do a biomechanical analysis in real-time.

Single Person Pose Estimation is easier to locate than Multi-Person Pose Estimation as an additional problem of inter-person occlusion occurs in it. As per the requirement of output dimensions, pose estimation can be classified in 2D or 3D.

Despite many years of research, articulated pose estimation of human is a challenging problem. Among the most significant challenges are strong articulations, small and barely visible joints, cluttered background, the variability of clothing in images, lighting conditions(variations), occlusion in the scene, motion blur, etc. To identify human pose estimation using an artificial intelligence technique, a convolutional neural network draws
maximum attention nowadays. This review paper is organized as follows. Section 1 gives the introduction of human pose estimation and provides the purpose of the study. Section 2 describes the basic structure of the Convolutional neural network (CNN) with all elementary concepts and LeNet architecture. Section 3 describes various deep learning-based approaches for human pose estimation and Section 4 illustrates the Conclusion.

2. CONVOLUTIONAL NEURAL NETWORK

A convolutional neural network is a deep learning approach that will take an input image or video, will assign weights and biases to extract and classify the given image. Compare to ANN (Artificial Neural Network), CNN can learn highly abstract features to identify objects very efficiently. CNN has certain beneficial advantages like sparse connection, parameter sharing and equivariance representation due to which training parameters reduced significantly, resulting in improved generalization. The architecture of the Convolutional neural network was inspired by the human nervous system and the organization of the visual cortex. To extract the features from images the network trains on thousands of images. This feature of CNN models makes extremely accurate for the computer vision tasks. In the deep convolutional neural network, there are tens or hundreds of hidden layers, so the network becomes more complex which make it able to learn to extract the feature from an image.

![Convolutional Neural Network](image)

**Figure 1. Convolutional Neural Network**

In CNN fewer parameters that need to train, so the training is smooth and not suffer from overfitting. Further, the classification stage is incorporated with the stage of feature extraction as both use learning process. Implementation of CNN is compared to easy that large ANN. Due to all above remarkable performance CNN are widely used in image classification [8], object detection [9], face detection [10], vehicle recognition [11], speech recognition [12], diabetic retinopathy [13], facial expression recognition [14] and many more. Pose estimation incorporate with object detection, body key points location etc. so CNN is a highly advisable network and thus a majority of pose estimation model is comprised of CNN as a basic
building block. A basic model of CNN consists of four components: Convolutional layer, Pooling layer, Activation function and fully connected layer.

2.1 Convolutional Layer [7]
Raw input data (images) is load and store at the input layer whose final classification will be done at the output layer. The input data - images can be black & white or colour. In black & white images two parameters - height and weight and in colour image three parameter - height, width and number of channels are specified. In RGB image number of channels will be 3.

Convolution is a mathematical operation takes place between input image matrix and a filter or kernel. In convolutional layer, height and width of filter is smaller compare to input image matrix. The filter is slide across the width and height of input image matrix and compute dot products at every spatial position.

After applying convolution operation, the output $a_{ij}$ for next layer can be

$$a_{ij} = \sigma ((w * x)_{ij}) + b$$

Where $x$ is the input layer, $w$ is a weight vector or filter, $b$ is the bias and “ $*$ “ represents convolution operation and $\sigma$ is non-linearity introduced in the network.

2.2 Pooling Layer
When the features are detected its extracted location becomes less significant so the convolutional layer is followed by pooling or subsampling layer to reduces the number of learning parameters and introduce translation invariance [15]. As it reduces data representation progressively over the network, which directly helps to control over fitting. To resize the data, pooling layer uses certain operation like max, min and average. Pooling layer uses filters to perform a down sampling process on the input. The most common set up for a pooling layer is to apply 2×2 filter with a stride of 2, so will down sample each depth slice in the input volume by a factor of 2 on the spatial dimensions (width and height). Pooling layer has hyper parameters, which are not learnable.

2.3 Activation Function
To add nonlinearity, nowadays Rectified Linear Unit (ReLU) proved better than the former sigmoid function due to several advantages like – the calculation of partial derivatives of ReLU is easy [15], training time is lesser [8], and ReLU do not allow gradients to disappear. Dying ReLU problem can be resolved by using Leaky ReLU.

2.4 Fully Connected Layers
Fully connected layers have normal parameters and hyper parameters. In this layer feature map matrix will be flattened in to vector and feed to a fully connected layers. They perform a transformation on input data as a function of activations (weights and biases of neurons).
2.5 LeNet Architecture

To understand CNN in a better way, LeNet architecture is mentioned above, which have a convolutional layer, 5x5 kernel and processes every output with a sigmoid activation function. The output channel of first convolutional layer is 6, which increase up to 16 in second layer. Height and weight are shrunk while the number of channels is increased considerably. So it makes the parameter sizes of two convolutional layer similar. To overcome the over fitting problem pooling layer down sample the size of matrix up to one quarter Three fully connected layers of 12, 84 and 10 outputs, added with bias is passed through softmax activation function.

3. DEEP LEARNING-BASED APPROACHES

Deep Convolutional Neural Network techniques have shown outstanding performance in many fields of computer vision like visual classification tasks, object detection etc. In this section, some of the novel architecture is mentioned for pose estimation.

3.1 Deep pose[2]

Deep pose was the first paper that applied deep learning-based approach to the pose estimation of human. In this proposed network, pose estimation is formulated as a DNN based regression problem towards body joints. They define pose vector, which consists of all k body parts joints. To get better results they used a cascade of such regression refinements. They propose a simple but powerful approach in a holistic fashion, means even if certain joints are hidden, they can be estimated if the pose is reasoned about holistically.
This paper argues that DNN naturally provides this sort of reasoning. The input is a full image and location of each body joint can find out without the use of graphical models. They used AlexNet as their CNN architecture with an extra final layer. In this network seven layers are made up of convolutional layer, pooling layers and fully connected layer, in which only convolutional and fully connected layers have learnable parameters. They both contain linear transformation followed by a non-linear rectified linear unit (ReLU). The size of convolutional layers is defined as Width × Height × Depth, where the depth defines the number of filters. The model is trained using L2 loss for regression. The model takes an input image of size 220×220 via stride of 4 and the learning rate is set to 0.0005. The drop out regularization for the fully connected layers are set to 0.6.

The data sets used in this model are Frames Labeled in Cinema(FLIC) which has 4000 training and 1000 test images from popular Hollywood movies and second is LSD(Leeds Sports Dataset) having 11000 training and 1000 testing images from sports activities, which are quite challenging in terms of articulations.

3.2 Stacked Hour Glass [6]

In this landmark paper, a novel CNN architecture is presented which beat all previous method. The network called stacked hourglass- as it consists of successive steps of pooling and up sampling layers looks like as hourglass and these are stacked together to produce a final set of predictions. The motivation of this network is to capture information like person’s posture and limb articulation at every scale.

When designing a network for Local evidence, it is essential to identify features like eyes, face, hands, etc., but to identify full-body pose estimation- need to require Global context. Hourglass capture all of these features and present pixel-wise predictions as an output. The Hourglass model performs repeated bottom-up (From high resolution to low resolution) and top-down (from low resolution to high resolution) processing with intermediate supervision.

Figure- 4 Hourglass network for pose estimation consists of multiple stacked hourglass modules which allow for repeated bottom-up, top-down inference.

The hourglass model consists of convolutional and max-pooling layers to process features down to a very low resolution. At each max-pooling layers, the network branches off and applies more convolutions at original pre pooled resolution. When it reached the lowest resolution, the model begins the top-down sequence of up-sampling and combination of
features across scales. They do nearest neighbour up-sampling of the lower resolution followed by an element-wise addition of the two sets of features. Hourglass has a symmetric topology, means for every layer present on the way down there is a corresponding layer going up.

The output of the model is a set of a heat map, for every given heat map the network predicts the probability of a joint presence at each and every pixel. The network operates at full input resolution of 256×256, so it requires a significant amount of GPU memory, so the final output resolution (highest) is 64×64 but it is not affected the network’s ability to produce precise joint predictions. The full network consists of 7×7 convolutional layer with stride 2 which followed by residual module and max-pooling which down resolution from 256 to 64. The hourglass network was tested on FLIC and MPII benchmark datasets. It achieves more than 2% improved accuracy on MPII across all joints and almost 4 to 5% on difficult joints such as knees and ankles.

3.3 Convolutional Pose Machine [4]

Convolutional pose machine [5] learns implicit spatial models via a sequential composition of convolutional architecture, such that it learns both image features and spatial models for prediction tasks. The architecture of this model is similar to the recurrent network as it generates the end result in multiple steps. It receives the images of resolution 368×368 as input, then applied few convolutions to the image in order to predict for each pixel map with a key point (head, neck, right elbow etc.)

![Convolutional Pose Machine](image)

Figure 5. Convolutional Pose Machine

The loss function is a mean squared error. It gives the error between the expected output per keypoint and predicted one. A small Gaussian function applied to expected output maps to find the correct position of the keypoints. To overcome the problem of vanishing gradient, apply losses of after each step in the architecture. Benchmark datasets MPII, LSP and FLIC used in this model.
3.4 Open Pose [3]

Open pose, an efficient method for multi-person 2D pose detection is open-source. This paper proposes a real-time approach for detecting 2D human poses in images and videos. Bottom-up representation using Part Affinity Fields (PAFs) is presented in this paper. Most of the model uses a top-down approach for multi-person pose estimation. In top-down strategy, first, a person is detected and then have estimated its pose independently on each detailed region. This technique is directly applicable for single person pose estimation but in multi-person, it fails to capture the spatial dependencies across a different person that require global inference.

Figure 6. Two branch multilevel CNN [3]

The network first extracts features from an image using a few layers. The extracted features are then fed into two parallel branches of convolutional layers. Out of two, first predicts a set of 18 confidence map with each map represents a specific part of a human skeleton. The second branch predicts a set of 38 PAF which represents the degree of association between parts.

To understand and interpret multiperson 2D pose estimation is critical aspects in computer vision. In this paper several landmark contributions they present like, representation of keypoint includes both position and orientation of human limbs, second, they propose an architecture that jointly learns part detection and association. Regardless of the number of people in the image, high-quality parses of body pose can be demonstrated with Greedy parsing algorithm. They compare both PAF refinement and combined PAF and body part location refinement and conclude that only PAF refinement is more important as runtime and accuracy point of view. Finally, they estimate a single model which have combine body and foot estimation which reduces the inference time of training.

4. Conclusion

Deep convolutional neural networks are successfully applied to learn various computer vision tasks. Here we, thrown light upon the basic understanding and elementary concept of CNN to solve many complex problems and review deep learning approaches for human pose estimation. Deep neural network with a graphical model is introduced by Thomson. In this model the parameters are learned jointly with the given network. With the introduction of "Deep pose" [2] research on human pose estimation shift towards deep learning. A. Toshev
proposed CNN based regression problem towards body joints in a holistic fashion. To resolve the problem of vanishing gradient Intermediate supervision was applied at every stage in stacked hourglass [6] architecture. Convolutional pose machine [4] is a sequential architecture for tackling structured predictions in the computer vision field without the need of a graphical model. Intermediate supervision enforced at every stage to resolve the problem of vanishing gradient. Open pose[3] is a real-time multi-person system to detect the human body key points on a single image. Its library is widely used for human re-identification, retargeting and human-computer interaction.

REFERENCES

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In CVPR, 2014.
[2] A. Toshev and C. Szegedy, “Deeppose: Human pose estimation via deep neural networks,” in CVPR, 2014.
[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in CVPR, 2017.
[4] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” in CVPR, 2016
[5] V. Ramakrishna, D. Munoz, M. Hebert, J. Bagnell, and Y. Sheikh. Pose Machines: Articulated Pose Estimation via Inference Machines. In ECCV, 2014.
[6] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In ECCV, 2016.
[7] Sakshi Indolia, Anilkumar Goswami, S.p. Mishra, Pooja Asopa: Conceptual understanding of convolutional neural network- A deep learning approach. In ICCIDS 2018 science direct procedia computer science 132 (2018) 679-688.
[8] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). “Imagenet classification with deep convolutional neural networks.” In Advances in neural information processing systems (pp. 1097-1105).
[9] Szegedy, C., Toshev, A., and Erhan, D. (2013) “Deep neural networks for object detection.” In Advances in neural information processing systems (pp. 2553-2561).
[10] Timoshenko, D., and Grishkin, V. (2013) “Composite face detection method for automatic moderation of user avatars.” Computer Science and Information Technologies (CSIT'13).
[11] Luo, X., Shen, R., Hu, J., Deng, J., Hu, L., and Guan, Q. (2017) “A Deep Convolution Neural Network Model for Vehicle Recognition and Face Recognition.” Procedia Computer Science 107: 715-720.
[12] Sainath, T. N., Kingsbury, B., Mohamed, A. R., Dahl, G. E., Saon, G., Soltan, H., Ramabhadran, B. (2013) “Improvements to deep convolutional neural networks for LVCSR.” In Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on (pp. 315-320).
[13] Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P., and Zheng, Y. (2016) “Convolutional neural networks for diabetic retinopathy.” Procedia Computer Science 90: 200-205.
[14] Uçar, A. (2017, July) “Deep Convolutional Neural Networks for facial expression recognition.” In Innovations in Intelligent Systems and Applications (INISTA), 2017 IEEE International Conference on (pp. 371-375).
[15] Zhou, Y., Wang, H., Xu, F., and Jin, Y. Q. (2016) “Polarimetric SAR image classification using deep convolutional neural networks.” *IEEE Geoscience and Remote Sensing Letters* **13** (12): 1935-19.

[16] Lee, K. B., Cheon, S., & Kim, C. O. (2017) “A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes.” *IEEE Transactions on Semiconductor Manufacturing* **30** (2): 135-142.

[17] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, “Joint training of a convolutional network and a graphical model for human pose estimation,” in NIPS, 2014.

[18] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998) “Gradient-based learning applied to document recognition.” *Proceedings of the IEEE* **86** (11): 2278-2324.