

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

186,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



# Articulated Human Pose Estimation Using Greedy Approach

*Pooja Kherwa, Sonali Singh, Saheel Ahmed, Pranay Berry and Sahil Khurana*

## Abstract

The goal of this Chapter is to introduce an efficient and standard approach for human pose estimation. This approach is based on a bottom up parsing technique which uses a non-parametric representation known as Greedy Part Association Vector (GPAVs), generates features for localizing anatomical key points for individuals. Taking leaf out of existing state of the art algorithm, this proposed algorithm aims to estimate human pose in real time and optimize its results. This approach simultaneously detects the key points on human body and associates them by learning the global context. However, In order to operate this in real environment where noise is prevalent, systematic sensors error and temporarily crowded public could pose a challenge, an efficient and robust recognition would be crucial. The proposed architecture involves a greedy bottom up parsing that maintains high accuracy while achieving real time performance irrespective of the number of people in the image.

**Keywords:** Neural networks, Pose- estimation, Greedy Search, Neural Network, Heat-maps

## 1. Introduction

Human pose estimation is a complex field of study in artificial intelligence, which requires a depth knowledge of computer vision, calculus, graph theory and biology. Initially this work start by introducing an image to a computer through camera and detect humans in the image known as object detection, as one of computer vision problem. In real world detecting an object from an image [1] and estimating its posture [2, 3] is two different aspects of objects. The latter is a very challenging and complex task. Images are filled with occluded objects, humans in close proximity, occlusions or spatial interference makes the task even more strenuous. One way of solving this problem is to use single person detector for estimation known as top down parsing [4–9]. This approach suffers from preconceived assumptions and lacks robustness. The approach is biased towards early decisions which makes it hard to recover if failed. Besides this, the computational time complexity is commensurate with the number of people in the image which makes it not an ideal approach for practical purpose. On a contrary the bottom up approach seems to perform well as compare to its counterpart. However earlier bottom up versions could not able to reduce the computational

complexity as it is unable to sustain the benefits of being consistent. For instance, the pioneering work E. Insafutdinov et al. Proposed a bottom up approach that simultaneously detects joints and label them as part candidates [10]. Later it associates them to individual person. Even solving the combinatorial optimization problem over a complete graph is itself NP hard. Another approach built on with stronger joint detectors based on ResNet [11] and provides ranking based on images, significantly improved its runtime but still performs in the order of minutes per image. The approach also requires a separate logistic regression for precise regression. After studying sufficient approaches and their shortcomings in the literature of image processing and object detection, this chapter introduces a efficient approach for human pose estimation.

## **1.1 Contribution of the work**

Optimizing the current state of the art results and introducing a new approach to solving this problem is the highlight of this chapter. In this chapter, we presented a bottom up parsing technique which uses a non-parametric representation, features for localizing anatomical key points for individuals. We further introduced a multistage architecture with two parallel branches one of the branches estimates the body joints via hotspots while the other branch captures the orientations of the joints through vectors This proposed approach is based on bottom up parsing, localizes the anatomical key points and associates them using greedy parsing technique known as greedy part association vectors. These 2D vectors aims to provide not only the encoded translator position but also the respective directional orientations of body parts. This approach also able to decouple the dependency of number of persons with running time complexity. Our approach has resulted in competitive performance on some of the best public benchmarks. The model maintains its accuracy while providing real time performance.

This chapter comprises of 6 sections: Section 2 discussed related work, in Section 3, proposed methodology is explained, in details with algorithms, in Section 4 results are discussed, and finally the chapter is concluded with future work in Section 5.

## **2. Related work**

The research trend that was primarily focused on detection of objects, visual object tracking and human body part detection, has advanced to pose estimation recently. Various visual tracking architectures have been proposed such as those based on convolutional neural networks and particle filtering and colored area tracking using mean shift tracking through temporal image sequence [12]. A survey of approaches for intruder detection systems in a camera monitored frame for surveillance was explained by C. Savitha and D. Ramesh [13]. A. Shahbaz and K. Jo also proposed a human verifier which is a SVM classifier based on histogram of oriented gradients along with an algorithm for change detection based on Gaussian mixture model [14]. But still there was a need of more precise detection algorithm that would accurately predict minor features as well. Human and Object detection evolved to detection of human body parts. L. Kong, X. Yuan and A.M.Maharajan introduced framework for automated joint detection using depth frames [15]. A cascade of Deep neural networks was used for Pose Estimation formulated as a joint regression problem and cast in DNN [16]. A full image and 7-layered generic convolutional DNN is taken as input to regress the location of each body joint. In [17], long-term temporal coherence was propagated to each stage of the overall video and data of joint position of initial posture was generated. A multi-

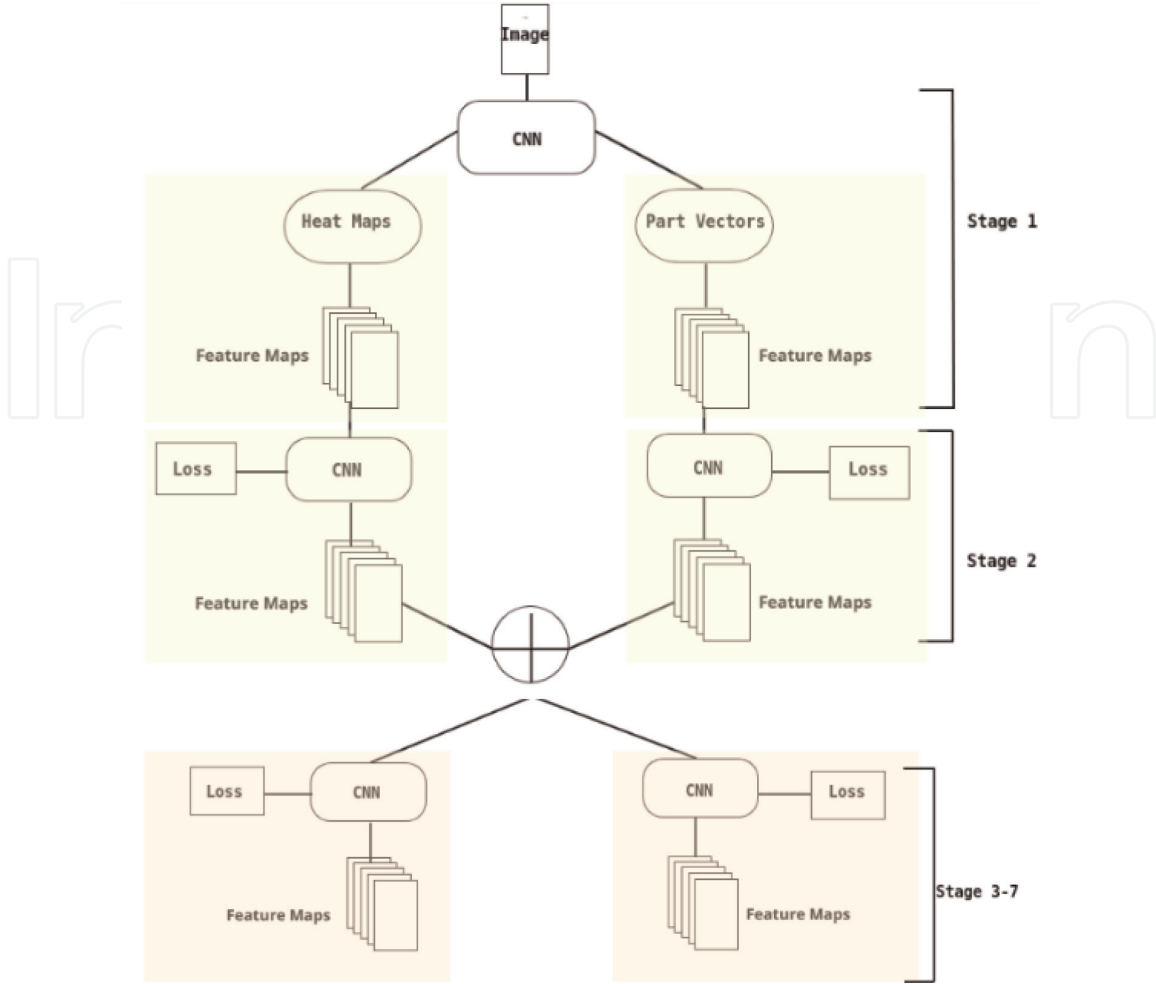
feature, three-stage deep CNN was adopted to maintain temporal consistency of video by halfway temporal evaluation method and structured space learning. Speeded up Robust features (SURF) and Scale Invariant Feature Transform (SIFT) was proposed by A. Agarwal, D. Samaiya and K. K. Gupta to deal with blur and illumination changes for different background conditions [18]. Paper [19] aims to improve human ergonomics using Wireless vibrotactile displays in the execution of repetitive or heavy industrial tasks. Different approach was presented to detect human pose. Coarse-Fine Network for Key point Localization (CFN) [20], G-RMI [21] and Regional Multi-person Pose Estimation (RMPE) [22] techniques have been used to implement top-down approach of pose detection (i.e. the person is identified first and then the body parts). An alternate bottom-up approach was proposed by Z. Cao, T. Simon, S. Wei and Y. Sheikh based on Partial Affinity Fields to efficiently detect the 2D pose [23]. X. Chen and G. Yang also presented a generic multi-person bottom-up approach for pose estimation formulated as a set of bipartite graph matching by introducing limb detection heatmaps. These heatmaps represent association of body joint pairs, that are simultaneously learned with joint detection [24]. L. Ke, H. Qi, M. Chang and S. Lyu proposed a deep conv-deconv modules-based pose estimation method via keypoint association using a regression network [25]. K. Akila and S. Chitrakala introduced a highly discriminating HOI descriptor to recognize human action in a video. The focus is to discriminate identical spatio-temporal relations actions by human-object interaction analysis and with similar motion pattern [26]. Y. Yang and D. Ramanan proposed methods for pose detection and estimation for static images based on deformable part models with augmentation of standard pictorial structure model by co-occurrence relations between spatial relations of part location and part mixtures [27]. A Three-dimensional (3D) human pose estimation methods are explored and reviewed in a paper, it involves estimating the articulated 3D joint locations of a human body from an image or video [28]. One more study includes a 2-D technique which localize dense landmark on the entire body like face, hands and even on skin [29].

### 3. Proposed approach for human pose detection

#### 3.1 Methodology

**Figure 1** depicts the methodology of our proposed approach, our approach works as black box which receive an image of a fixed size and produces a 2D anatomical key point of every person in the image. After performing the needed preprocessing, the image is passed through a feed forward convolutional neural network. The architecture has two separate branches that runs simultaneously

- i. On one branch it predicts an approximations represented by a set of hotspots  $H$  for each body joint locations while the
- ii. Other branch predicts a set of 2D vectors representing joints associations  $P$  for each pair of different joints. Each set  $H$  is a collection of  $\{H_1, H_2, H_3, \dots, H_j\}$   $j$  hotspots one for each joint and  $P$  is a collection of  $\{L_1, L_2, L_3, \dots, L_k\}$   $k$  part association vector field for each pair or limb. The output of these two branches will be summed up using parsing algorithm and feed forward to multiple layers of convolutional net ultimately giving 2D anatomical key points for every person in the image.



**Figure 1.** Schematic diagram of a multistage architecture. Two parallel branches feeds forward the network. Heat maps predicts the approximation while part association vectors predict association and orientations.

### 3.2 Part detection using heat-maps

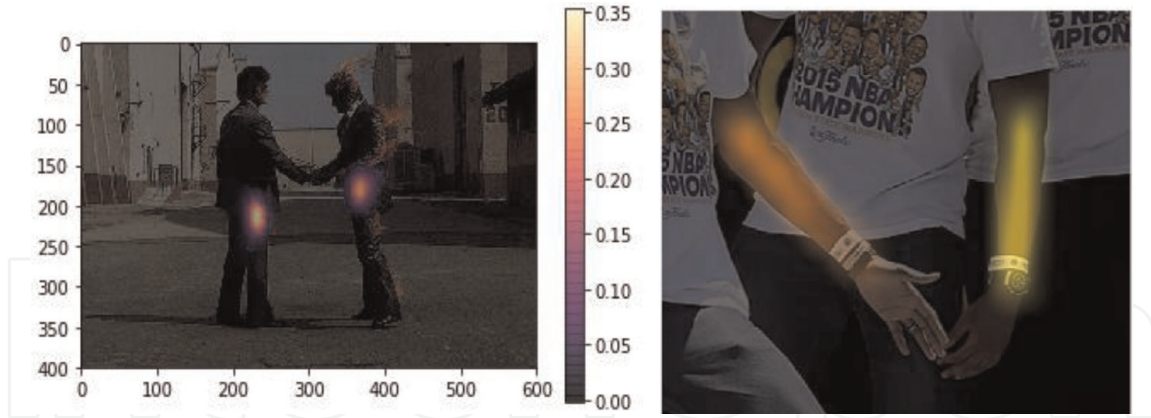
The heat maps produced by convolutional neural net are highly reliable supporting features. The heat maps are set of matrices that stores the confidence that the network has that a pixel contains a body joint. As many as 16 matrices for each of the true body joints. The heat map specifies the probability that a particular joint exist within a particular pixel location. The very idea of having heat maps provide support in predicting the joint location. The visual representation of heat maps could give an intuition of a presence of body joint. The darker the shade or sharper the peak represents a high probability of a joint. Several peaks represent a crowded image representing one peak for one person (**Figure 2**).

Calculating the confidence map or heat maps  $C_{jk}^*$  for each joint requires some prior information for comparison. Let  $x_{jk}$  be the empirical position of a body joint  $j$  of the person  $k$ . These confidence maps at any position  $m$  can be created by using the empirical position  $x_{jk}$ . The value of confidence map at location  $p$  in  $C_{jk}^*$  is given by

$$C_{jk}^*(m) = \exp\left(\frac{-\Delta^2}{\sigma^2}\right) \quad (1)$$

where  $\sigma$  is spread from the mean and  $\Delta$  is the absolute difference of  $x_{jk}$  and  $m$ .





**Figure 2.**  
 Illustration of one segment of the pipeline i.e. predicting heat maps through neural network. It gives the confidence metric with regards to the presence of the particular body part in the given pixel.

All the confidence maps get aggregated by the network to produce the final confidence map. The final confidence map is generated by the network obtained from the aggregation of the individual maps.

$$C_{jk}^*(m) = \max \left( C_{jk}^*(m) \right) \quad (2)$$

These confidence maps are rough approximations, but we need the value for that joint. We need to extract value from the hot spot. For the final aggregated confidence map we take the max of the peak value while suppressing the rest.

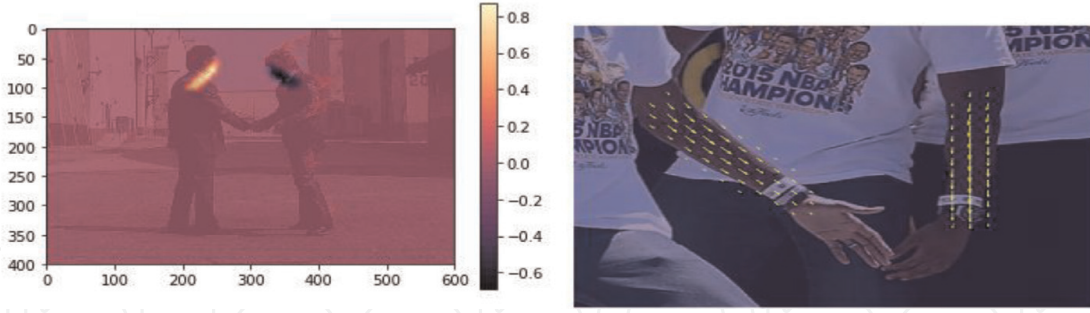
### 3.3 Greedy part association vector

The problem that comes while detecting the pose is that even if we have all the anatomical key points how we are going to associate them. The hotspot or the key points itself have no idea of the context on how they are connected. One way to approach this problem is to use a geometrical line midpoint formula. But the given approach would suffer when the image is crowded as it would tend to give false association. The reason behind the false association is the limitation of the approach as it tend to encode only the position of the pair and not the orientations and also it reduces the base support to a single point. In order to address this issue, we want to implement a greedy approach known as greedy part association vector which will preserve the position along with the orientation across the entire area of pair support. Greedy part association vectors are the 2D vector fields that provides information regarding the position and the orientation of the pairs. These are a set of coupled pair with one representing x axis and the other representing the y axis. There are around 38 GPAVs per pair and numerically index as well (Figure 3).

Consider a limb  $j$  with 2 points at  $x_1$  and  $x_2$  for  $k^{th}$  person in the image. The limb will have many points between  $x_1$  and  $x_2$ . The greedy part association vector at any point  $c$  between  $x_1$  and  $x_2$  for  $k^{th}$  person in the image represented by  $G_{j,k}^*$  can be calculated as.

$$G_{j,k}^* = \hat{c} \text{ if } c \text{ is on limb } j \text{ and person } k. \text{ Or } 0 \text{ otherwise.} \quad (3)$$

where  $\hat{c}$  a unit vector along the direction of limb equivalent to



**Figure 3.** Illustration of the other segment greedy part vectors, preserving the position along with the orientations and finally associates the joints through greedy parsing.

$$\frac{x_2 - x_1}{\sqrt{x_2^2 - x_1^2}} \quad (4)$$

The empirical value of final greedy part association vector will be the average of GPAVs of all the person in the image.

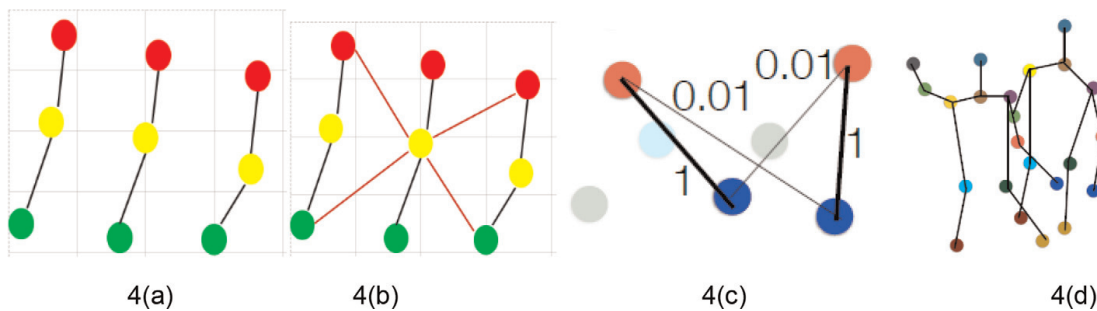
$$G_j^* = \frac{\sum_k G_{j,k}^*}{n_j(c)} \quad (5)$$

where  $G_{j,k}^*$  is the greedy part association vector at any point and  $n_j(c)$  is the total number of vectors at the same point  $c$  among all people.

### 3.4 Multi person pose estimation

After getting the part candidates using non-maximum suppression, we need to associate those body parts to forms pairs. For each body part there are  $n$  numbers of part candidates for association. On an abstract level one-part can form association with every possible part candidate forming a complete graph (**Figure 4**).

For example, we have detected a set of plausible neck candidates and a set of hip candidates. For each neck candidates there is a possible connection with the right hip candidates giving a complete bipartite graph having the nodes as part candidates and the edges as possible connections. We need to associate only the optimal part giving rise to a problem of  $N$  dimensional matching problem which itself a NP hard problem. In order to solve this optimal matching problem, we need to assign weights to each of possible connection. This is where the greedy part association vectors come into the



**Figure 4.** (a–d) Solving the assignment problem for associating body joints to form the right pair. Assigning weights to each possible connection with the help of greedy association vectors.

pipeline. These weights are assigned using the aggregated greedy part association vector.

In order to measure the association between two detected part candidates. We need to integrate over the predicted greedy part association vector found in previous section, along these two detected part candidates. This integral will give assign a score to each of the possible connections and store the scores in a complete bipartite graph. We need the find the directional orientation of the limb with respect to these detected part candidates. Empirically we have two detected part candidates namely  $t_1$  and  $t_2$  and the predicted part association vector  $G_j$ . An integral over the curve will give a measure of confidence in their association.

$$E = \int_{i=0}^{i=1} G_j(c(m)) \cdot \hat{d} \cdot dm \quad (6)$$

where  $G_j(c(m))$  greedy part association vector and  $\hat{d}$  is a unit vector along the direction two non-zero vectors  $t_1$  and  $t_2$ .

After assigning weights to the edges our aim is to find the edge for a pair of joints with the maximum weight. For this we choose the most intuitive approach. We started with sorting the scores in descending manner followed by selecting the connection with the max score. We then move to the next possible connection if none of the parts have been assigned a score, this is a final connection. Repeat the third step until done.

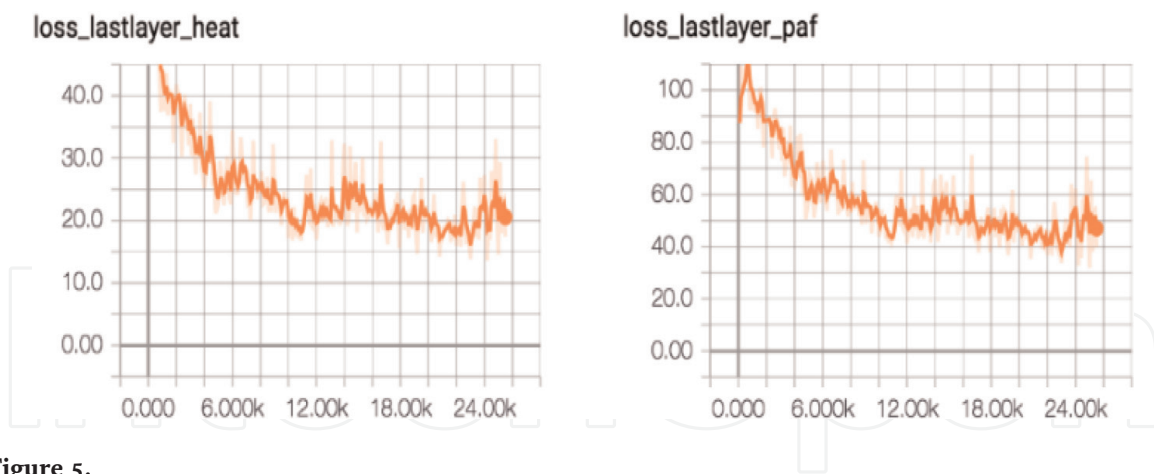
The final step involves merging the whole detected part candidates with optimal scores and forming a complete 2D stick figure of human structure. One way to approach this problem is that let us say each pair of part candidates belong a unique person in the image that way we have a set of humans i.e.  $\{H_1, H_2, H_3, \dots, H_k\}$  where  $k$  is the total number of final connection. Each human in the set contain a pair i.e. pair of body parts. Let represent the pairs as a tuple of indices one in x direction and one in y direction.  $H_i = \{(m_{idx}, m_x, m_y), (n_{idx}, n_x, n_y)\}$ . Now comes the merging we conclude that if two human set shares any index coordinates with other set means that they share a body part. We merge the two sets and delete the other. We perform the same steps for all of the sets until no two human share a part ultimately giving a human structure.

#### 4. Results

For the training and evaluating the final build we used a subset of a state-of-the-art public dataset, the COCO dataset. COCO dataset is collection of 100 K images with diverse instances. We have used a subset of those person instances with annotated key points. We have trained our model on 3 K images, cross validated on 1100 images and tested on 568 images. The metric used for evaluation is OKS stands for Object key point similarity. The COCO evaluation is based on mean average precision calculates over different OKS threshold. The minimum OKS value that can have is 0.5. We are only interested in key points that lie within 2.77 of the standard deviation (**Figure 5**).

Above table compares the performance of our model with the other state of the art model. **Table 1** shows the mAP performance comparison of our model with others on a testing dataset of 568 images. We can see clearly our novel approach outperforms the previous key point benchmarks. We can also see our model achieved a significant





**Figure 5.**  
Convergence of training losses for both the heat maps (L) and greedy part vectors (R).

Method	Head	Shoulder	Elbow	Hip	Knee	Ankle	Wrist	mAp
Deep cut	73.4	71.8	57.9	56.7	44.0	32.0	39.9	54.1
Iqbal et al	70.0	65.2	56.4	52.7	47.9	44.5	46.1	54.7
Deeper cut	87.9	84.0	71.9	68.8	63.8	58.1	63.9	71.2
Proposed Approach	90.7	90.9	79.8	76.1	70.2	66.3	70.5	77.7

**Table 1.**  
mAP performance comparison of our model with others on a testing dataset of 568 images.

Method	Head	Shoulder	Elbow	Hip	Knee	Ankle	Wrist	mAp
Deep cut	78.4	72.5	60.2	57.2	52.0	45.0	51.0	54.1
Iqbal et al	58.4	53.9	44.5	42.2	36.7	31.1	35.0	54.7
Deeper cut	87.9	84.0	71.9	68.8	63.8	58.1	63.9	71.2
Proposed Approach	90.1	87.9	75.8	73.1	65.2	60.3	66.5	73.7

**Table 2.**  
Performance comparison on a complete testing dataset of 1000 images.

rise in mean average precision of 6.5%. Our inference time is 3 order less. **Table 2** presents the performance comparison on a complete testing dataset of 1000 images. Here again we can see our model outperforming the rest. Our model achieved a rise of almost 2.5% in mean average precision as compare to other models. The above comparison of our model with earlier state of the art bottom up approaches presents the significance of our model.

### 5. Conclusion and future work

Solving one of the complex problems in computer vision was a huge challenge. Optimizing the current state of the art results and introducing a new approach to solving this problem is the highlight of this chapter. In this chapter, we presented a bottom up parsing technique which uses a non-parametric representation, features for localizing anatomical key points for individuals. We further introduced a multistage

architecture with two parallel branches one of the branches estimates the body joints via hotspots while the other branch captures the orientations of the joints through vectors. We ran our model on a publicly available COCO dataset for training, cross validation and testing. Finally, we evaluated the results and achieved a mean average precision of 77.7. We compare our results with existing models and achieved a significant rise of 2.5% in mAP with less inference time. We have showed the results in **Tables 1 and 2**. We aim to expand our project in future by proposing a framework for human pose comparator based on the underlying technology used in single person pose estimation to compare the detected pose with that of the target in real-time. This would be done by developing a model to act as an activity evaluator by learning physical moves using key points detection performed by the source and compare the results with the moves performed by the target along with a scoring mechanism that would decide how well the two sequence of poses match. In a nutshell, we aim to build an efficient comparison mechanism that would accurately generate the similarity scores based on the series of poses between the source and the target as the future scope of this project.

## Author details

Pooja Kherwa\*, Sonali Singh, Saheel Ahmed, Pranay Berry and Sahil Khurana  
Maharaja Surajmal Institute of Technology, New Delhi, India

\*Address all correspondence to: [poona281280@gmail.com](mailto:poona281280@gmail.com)

## IntechOpen

© 2021 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [2] A. Toshev and C. Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 1653-1660.
- [3] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [4] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional Pose Machines," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [5] W. Ouyang, X. Chu, and X. Wang, "Multi-source Deep Learning for Human Pose Estimation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.
- [6] J. Tompson, R. Goroshin, A. Jain, Y. Lecun, and C. Bregler, "Efficient object localization using Convolutional Networks," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
- [7] V. Belagiannis and A. Zisserman, "Recurrent Human Pose Estimation," 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 2017.
- [8] A. Bulat and G. Tzimiropoulos, "Human Pose Estimation via Convolutional Part Heatmap Regression," *Computer Vision – ECCV 2016 Lecture Notes in Computer Science*, pp. 717–732, 2016.
- [9] T. Pfister, J. Charles, and A. Zisserman, "Flowing ConvNets for Human Pose Estimation in Videos," 2015 IEEE International Conference on Computer Vision (ICCV), 2015.
- [10] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model," *Computer Vision – ECCV 2016 Lecture Notes in Computer Science*, pp. 34–50, 2016.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [12] R. J. Mozhdehi and H. Medeiros, "Deep convolutional particle filter for visual tracking," 2017 IEEE International Conference on Image Processing (ICIP), Beijing, 2017, pp. 3650-3654.
- [13] C. Savitha and D. Ramesh, "Motion detection in video surveillance: A systematic survey," 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2018, pp. 51-54.
- [14] A. Shahbaz and K. Jo, "Probabilistic Change Detector with Human Verifier for Intelligent Sterile Zone Monitoring," 2018 IEEE 27th International Symposium on Industrial Electronics (ISIE), Cairns, Australia, 2018, pp. 777-781.
- [15] L. Kong, X. Yuan, and A. M. Maharjan, "A hybrid framework for automatic joint detection of human

poses in depth frames,” *Pattern Recognition*, vol. 77, pp. 216–225, 2018.

[16] A. Toshev and C. Szegedy, “DeepPose: Human Pose Estimation via Deep Neural Networks,” *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 1653-1660..

[17] S. Liu, Y. Li and G. Hua, “Human Pose Estimation in Video via Structured Space Learning and Halfway Temporal Evaluation,” in *IEEE Transactions on Circuits and Systems for Video Technology*

[18] A. Agarwal, D. Samaiya and K. K. Gupta, “A Comparative Study of SIFT and SURF Algorithms under Different Object and Background Conditions,” *2017 International Conference on Information Technology (ICIT)*, BHUBANESWAR, India, 2017, pp. 42-45

[19] W. Kim, M. Lorenzini, K. Kapicioglu and A. Ajoudani, “ErgoTac: A Tactile Feedback Interface for Improving Human Ergonomics in Workplaces,” in *IEEE Robotics and Automation Letters*.

[20] S. Huang, M. Gong, and D. Tao, “A Coarse-Fine Network for Keypoint Localization,” *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[21] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, “Towards Accurate Multi-person Pose Estimation in the Wild,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, “RMPE: Regional Multi-person Pose Estimation,” *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[23] Z. Cao, T. Simon, S. Wei and Y. Sheikh, “Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 1302-1310

[24] X. Chen and G. Yang, “Multi-Person Pose Estimation with LIMB Detection Heatmaps,” *2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018, pp. 4078-4082

[25] L. Ke, H. Qi, M. Chang and S. Lyu, “Multi-Scale Supervised Network for Human Pose Estimation,” *2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018, pp. 564-568

[26] K. Akila and S. Chitrakala, “Discriminative human action recognition using -HOI descriptor and key poses,” *2014 International Conference on Science Engineering and Management Research (ICSEMR)*, Chennai, 2014, pp. 1-6

[27] Y. Yang and D. Ramanan, “Articulated Human Detection with Flexible Mixtures of Parts,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, p

[28] Wang, Jinbao, et al. “Deep 3D human pose estimation: A review.” *Computer Vision and Image Understanding* (2021): 103225.

[29] Jin, Sheng, et al. “Whole-body human pose estimation in the wild.” *European Conference on Computer Vision*. Springer, Cham, 2020.