3D pose estimation using CNN and 3D depth map probabilistic model.

Prof. Nayana Mahajan,
Assistant Professor Department of Electronics,
Vidyalankar Institute of Technology, Mumbai,
Maharashtra, India

Mayur Bharankar, Rishikesh Mudaliyar,
Sourabh Kothari, RoshanKumar Poojari
Students,
Department of Electronics,
Vidyalankar Institute of Technology, Mumbai,
Maharashtra, India

Abstract:

We have integrated the techniques of 3D pose estimation and action recognition from 2D images. By using Convolutional Neural network and training the network by providing a set of images from the pretrained data for detection of pose from the images. This estimation of the pose will further be used for the human computer interface for other projects. With this method we intend to reduce the challenges between 2D joint detection and 3D pose/model reconstruction.

Keywords: Convolutional Neural network (CNN), Pose estimation, Action recognition , Kinect

1. Introduction:

3D pose estimation from 2D raw images is one of the most challenging tasks in computer vision. It involves tackling two inherently ambiguous tasks.[1] One of the biggest reasons of these challenges and problems is the largen number of ambiguity and variations in the actions and poses of various humans, skin tone, colour , lighting conditions etc .Even achieving this issue another milestone is to project these 2D points or landmarks into 3D space which is almost infinite is very challenging as the possibilities for the same and creating the probabilistic model is infinite too. Finding the correct 3D pose that matches the image requires injecting additional information usually in the form of 3D geometric pose priors and temporal or structural constraints. A new joint approach to 2D landmark detection and full 3D pose estimation from a single RGB image that takes advantage of reasoning jointly about the estimation of 2D and 3D landmark locations to improve both tasks. We propose a novel CNN architecture that learns to combine the image appearance-based predictions provided by convolutional-pose-machine style 2D landmark detectors , with the geometric 3D skeletal information from Kinect encoded in a novel pretrained model of 3D human pose. Information captured by the 3D human pose model is embedded in the [2][3][4] CNN architecture as an additional layer that lifts 2D landmark coordinates into 3D while imposing that they lie on the space of physically plausible poses. The advantage of integrating the output proposed by the 2D landmark location predictors – based purely on image appearance – with the 3D pose predicted by a probabilistic model, is that the 2D landmark location estimates are improved by guaranteeing that they satisfy the anatomical 3D constraints encapsulated in the human 3D pose model. In this way, both tasks clearly benefit from each other. A further advantage of this approach is that the 2D and 3D training data sources may be completely independent. The deep architecture only needs that images are annotated with 2D poses, not 3D poses. The human pose model is trained independently and exclusively from 3D Kinect dataset. This decoupling between 2D and 3D training data presents a huge advantage since we can augment the training sets completely independently. For instance we can take advantage of extra 2D pose annotations without the need for 3D ground truth or extend the 3D training data to further mocap datasets without the need for synchronized 2D images

2. Overview:

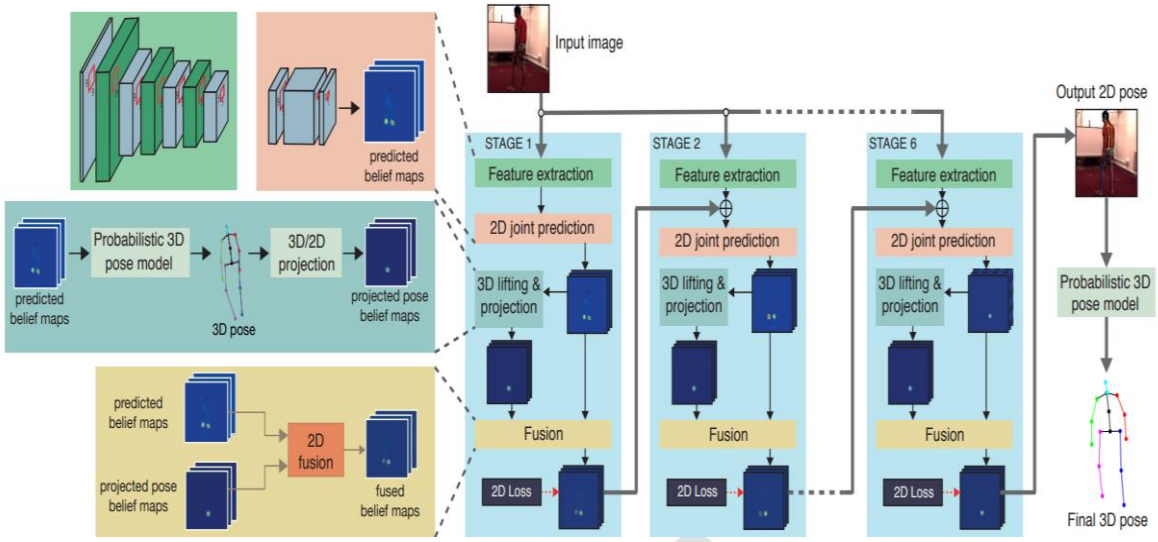

Figure 1: The multistage deep architecture for 2D/3D human pose estimation

Each stage produces as output a set of belief maps for the location of the 2D landmarks (one per landmark). The belief maps from each stage, as well as the image, are used as input to the next stage. Internally, each stage learns to combine: (a) belief maps provided by convolutional 2D joint predictors, with (b) projected pose belief maps, proposed by the probabilistic 3D pose model. Figure 1 The 3D pose layer is responsible for lifting 2D landmark coordinates into 3D and projecting them onto the space of valid 3D poses. These two belief maps are then fused into a single set of output proposals for the 2D landmark locations per stage. The accuracy of the 2D and 3D landmark locations increases progressively through the stages. [7]The loss used at each stage requires only 2D pose annotations, not 3D. The overall architecture is fully differentiable – including the new projected-pose belief maps and 2D-fusion layers – and can be trained end-to-end using back-propagation.

A huge amount of Work has been pursued in recovering 3D Pose estimates from the 2D images[20][23][21][14][16][17. Before the advancements in the statistical perspective for the modelling, anatomical data was relied for the human pose estimation .More recent methods have focused on learning a prior statistical model of the human body directly from 3D mocap data.

3. Human Pose Estimation

3.1 2D pose Estimation

Human 2D pose estimation—the problem of classifying the figurative key points or "parts"—has largely focused on identifying human body. Referring the pose of the people in the gatherings and the social events, which are a great source of the variations in human body movements. Firstly, each image may contain an unknown number of people that can occur at any position or scale. Secondly, interactions between people induce complex spatial interference, due to contact, ambient occlusion, and limb articulations, making association of parts difficult and complex when it comes to detection. Thirdly, runtime complexity tends to grow with the number of people in the image, making Realtime performance a challenge. Real time data is completely random and dynamic hence any sort of data can be fed in the system. A common approach is to employ a person detector and perform single-person pose estimation for each detection. [27][29][30][17] These top-down approaches directly leverage existing techniques for single-person pose estimation, but suffer from early commitment: if the person detector fails–as it is prone to do when people are in close proximity–there is no recourse to recovery. Furthermore, the runtime of these top-down approaches is proportional to the number of people: for each detection, a single-person pose estimator is run, and the more people there are, the greater the computational cost. In contrast, bottom-up approaches are attractive as they offer robustness to early commitment and have the potential to decouple runtime complexity from the number of people in the image. Yet, bottom-up approaches do not directly use global contextual cues from other body parts and other people. In practice, previous bottom-up methods do not retain the gains in efficiency as the final parse requires costly global inference. Figure 2 But here we fetch the data from pretrained dataset viz Human 3.6M, COCO data sets. These data sets have a lot of images of different variations in 2D poses and other variations to serve as a super set for all the motions to be detected. This is then provided to a CNN layer as a reference which is later combined with the 3D Kinect data set[33][8].

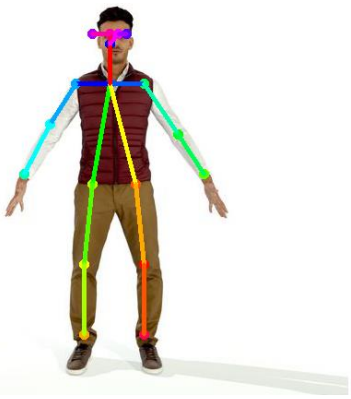

Figure 2: 2D Pose Estimation

3.2 3D Pose estimation

The problem of 3D pose estimation is divided into two parts. The camera coordinates are fetched by the 2D pose estimation and these new coordinates now stand as the basis for the 3D pose estimation. A Skeletal-bone representation (similar to the rigs) Figure 3 of the human pose was proposed to reduce the data variance, however, such a structural transformation might effect negatively tasks that depend on the extremities of the human body, since the error is accumulated as we go away from the root joint. Moreover, quick actions lead to slower response from the system as it depends upon the system of calculation. volumetric stacked hourglass architecture. However, the method suffers from the significant increase in the number of parameters and in the required memory to store all the gradients.

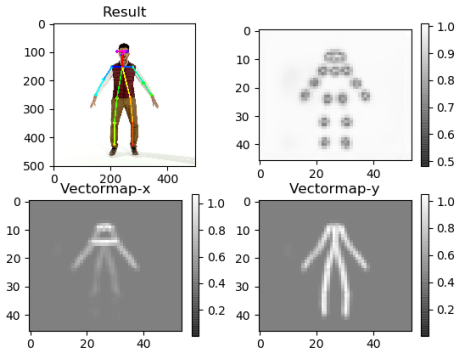

Figure 3: Output with Vector

4. Architecture:

A) Software Architecture:

1. Convolutional Neural Networks

The CNN used in this project consists of five convolutional layers, three pooling layers, two parallel sets of two fully connected layers, and loss layers for 2D and 3D pose estimation tasks. The CNN accepts a 225 × 225px sized image as an input. The filter sizes of convolutional and pooling layers are the same as those of ZFnet, but we reduced the number of feature maps to make the network smaller.

1) ***Predicting CNN-based belief-maps***: We use a set of convolutional and pooling layers, equivalent to those used in the original CPM architecture, that combine evidence obtained from image learned features with the belief maps obtained from the previous stage ($t-1$) to predict an updated set of belief maps for the 2D human joint positions.

2) ***Lifting 2D belief-maps into 3D***: the output of the CNN based belief maps is taken as input to a new layer that uses new pretrained probabilistic 3D human pose model to lift the proposed 2D poses into 3D.

3) ***Projected 2D pose belief maps***: The 3D pose estimated by the previous layer is projected back onto the image plane to produce a new set of projected pose belief maps. These maps encapsulate 3D dependencies between the body parts.

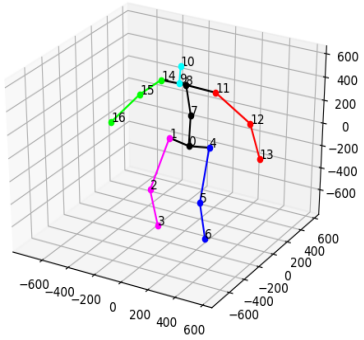

Figure 4: 2D Pose Estimation

In our approach, we also propose an intermediate volumetric representation for 3D poses, but we use a much lower resolution than in and still are able to increase significantly the state-of-the-art results, since our method is based on a continuous regression function. We track the body movements and gestures of the user without the need of them holding any device. This is done only by using the RGB camera and an infrared depth detection sensor. These Sensory interfaces are expected to allow more intuitive experience of interaction with computers and provide a greater level of control than that offered by the traditional keyboard and mouse and other input devices. In the depth Sensor the three main components are the infrared projector, infrared camera and the RGB camera. The RGB camera captures the image whilst the IR projector projects thousands of the IR dots on the subject. These Dots are invisible to the naked eye but are captured by the IR camera which helps in creating a depth map. Now in the CNN network A 3D pretrained model Figure 4 will be added as the last layer of the convolutional neural network which will gratify the 2D estimated

4) ***2D Fusion layer***: The final layer in each stage learns the weights to fuse the two sets of belief maps into a single estimate passed to the next stage.

5) ***Final lifting***: The belief maps produced as the output of the final stage ($t = 6$) are then lifted into 3D to give the final estimate for the pose using our algorithm to lift 2D poses into 3D.

B) Softwares used:

1. ***Anaconda***: Anaconda Figure 5 is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. Package versions are managed by the package management system conda. The Anaconda distribution is used by over 12 million users and includes more than 1400 popular data-science packages suitable for Windows, Linux, and MacOS

2. ***OpenCV***: OpenCV [Figure 6] (Open source computer vision) is a library of programming functions mainly aimed at real-time computer vision. OpenCV supports the deep learning frameworks TensorFlow, Torch/PyTorch and Caffe.

3. ***Tensor Flow***: TensorFlow is an open-source software library for dataflow and differentiable programming across a range of tasks. Figure 7 It is a symbolic math library, and is also used for machine learning applications such as neural networks.

C) Hardware:

1. ***Kinect***: Kinect (codenamed Project Natal during development) is a line of motion sensing input devices produced by Microsoft. Initially, the Kinect was developed as a gaming accessory for Xbox 360 and Xbox One video game consoles and Microsoft Windows PCs. Based around a webcam-style add-on peripheral, it enabled users to control and interact with their console/computer without the need for a game controller, through a natural user interface using gestures and spoken commands

5. Results:

The following results are obtained depicting the 2D pose estimation and then providing it as the input to the 3D probabilistic model.

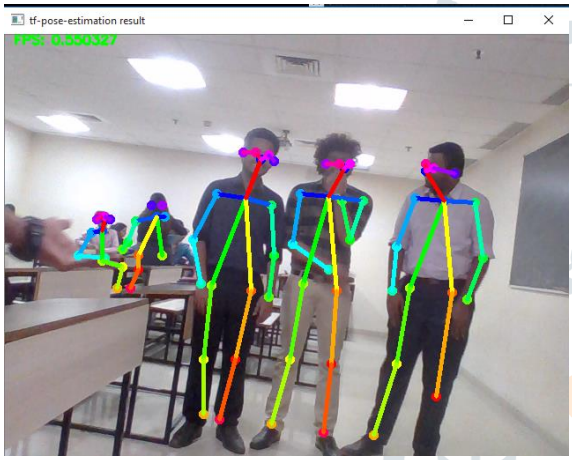

Figure 5: Multiperson 2D pose estimation.

This is the 3D skeletal structure fetched and produced from the 2D coordinates

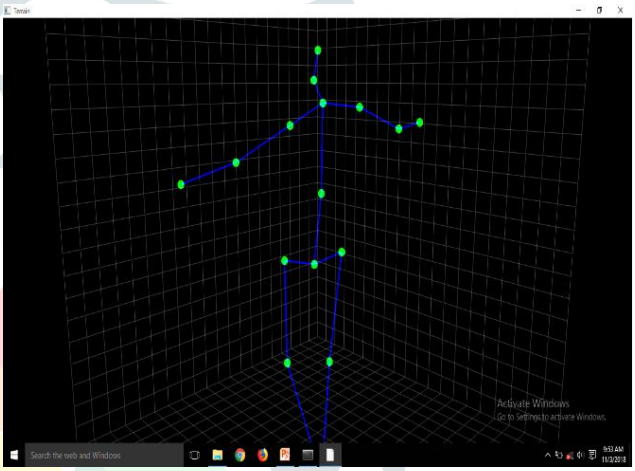

Figure 6: Output in 3D Plot.

6. Conclusion:

According to the results inferred , the 2D estimated poses of many people at once which is then converted to a 3D plot. CNN and posing the 3D Kinect data set as an extended model of the network, we were able to achieve greater accuracy when compared to normal detection. Also the boundaries to the 3D depth maps which restrained it to suffice the 2D modelled data, giving us apparently accurate 2D to 3D conversion. By using the depth maps for the estimation accurate anatomical constrains were added to the final model.

7. Reference:

1. U. Iqbal, M. Garbade, and J. Gall. "Pose for action - action for pose." FG-2017, 2017
2. Kokkinos. "Ubernet: Training a 'universal' convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. Computer Vision and Pattern Recognition". (CVPR), 2017
3. X. Zhou, M. Zhu, G. Pavlakos, S. Leonardos, K. G. Derpanis, and K. Daniilidis. "Monocap: Monocular human motion capture using a CNN coupled with a geometric prior." CoRR, abs/1701.02354, 201
4. Bulat and G. Tzimiropoulos. "Human pose estimation via convolutional part heatmap regression." In ECCV, 2016.x
5. S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. "Convolutional pose machines.'' In CVPR, 2016
6. K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. "LIFT: Learned InvariantFeatureTransform." European Conference on Computer Vision (ECCV), 2016
7. S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. "Convolutional pose machines." arXiv preprint arXiv:1602.00134, 2016
8. G. Ch'eron, I. Laptev, and C. Schmid. "P-CNN: Pose-based CNN Features for Action Recognition." In ICCV, 2015
9. I. Akhter and M. J. Black. "Pose-conditioned joint angle limits for 3D human pose reconstruction." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1446–1455, 2015
10. X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis. "Sparseness meets deepness: 3D human pose estimation from monocular video." arXiv preprint arXiv:1511.09439, 2015.
11. V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh. "Pose machines: Articulated pose estimation via inference machines." In ECCV, 2014
12. Zeiler, M.D., Fergus, "R.: Visualizing and understanding convolutional networks." In: European Conference on Computer Vision, Springer (2014) 818–833
13. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: "Caffe: Convolutional architecture for fast feature embedding." arXiv preprint arXiv:1408.5093 (2014)
14. X. Fan, K. Zheng, Y. Zhou, and S. Wang. "Pose locality constrained representation for 3D human pose reconstruction." In European Conference on Computer Vision, pages 174–188. Springer, 2014
15. L.Pishchulin,M.Andriluka,P.Gehler,andB.Schiele. "Poselet conditioned pictorial structures." In CVPR, 2013
16. Y. Yang and D. Ramanan. "Articulated human detection with flexible mixtures of parts." In TPAMI, 2013
17. V. Ramakrishna, T. Kanade, and Y. Sheikh. "Reconstructing 3D human pose from 2D image landmarks." In European Conference on Computer Vision, pages 573–586. Springer, 2012.
18. M. Andriluka, S. Roth, and B. Schiele. "Monocular 3D pose estimation and tracking by detection." In CVPR, 2010
19. S. Johnson and M. Everingham. "Clustered pose and non linear appearance models for human pose estimation." InBMVC, 2010
20. M. Andriluka, S. Roth, and B. Schiele. "Pictorial structures revisited: people detection and articulated pose estimation". In CVPR, 2009.
21. C. H. Ek, P. H. S. Torr, and N. D. Lawrence. "Gaussian process latent variable models for human pose estimation." In A. Popescu-Belis, S. Renals, and H. Bourlard, editors, MLMI, volume 4892 of Lecture Notes in Computer Science, pages 132–143. Springer, 2007
22. P. F. Felzenszwalb and D. P. Huttenlocher. "Pictorial structures for object recognition." In IJCV, 2005.
23. D.Ramanan,D.A.Forsyth,andA.Zisserman. "StrikeaPose: Tracking people by finding stylized poses." In CVPR, 2005
24. V. Parameswaran and R. Chellappa. "View independent human body pose estimation from a single perspective image." In Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, volume 2, pages II–16. IEEE, 2004
25. C. Barron and I. A. Kakadiaris. "Estimating anthropometry and pose from a single uncalibrated image." Computer Vision and Image Understanding, 81(3):269–284, 2001
26. C. J. Taylor. "Reconstruction of articulated objects from point correspondences in a single uncalibrated image." In Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on, volume 1, pages 677–684. IEEE, 2000
27. Camillo J. Taylor, "Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image", . December 2000
28. Christopher Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. "P nder: Real-time tracking of the human body." IEEE Trans. Pattern Anal. Machine Intel l., 19(7):780{785, July 1997

29. Alex Pentland and Bradley Horowitz. "Recovery of nonrigid motion and structure." IEEE Trans. Pattern Anal. Machine Intel l., 13(7):730{742, July 1991.
30. Zhe Cao Tomas Simon Shih-En Wei Yaser Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields" The Robotics Institute, Carnegie Mellon University {zhecao,shihenw}@cmu.edu {tsimon,yaser}@cs.cmu.edu
31. Sungheon Park, Jihye Hwang, and Nojun Kwak "3D Human Pose Estimation Using Convolutional Neural Networks with 2D Pose Information." Korea sungheonpark@snu.ac.kr, hjh881120@gmail.com, nojunk@snu.ac.kr
32. H.-J. Lee and Z. Chen. "Determination of 3D human body postures from a single view." Computer Vision, Graphics, and Image Processing, 30(2):148–168, 1985.
33. "Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image" Denis Tome University College London D.Tome@cs.ucl.ac.uk Chris Russell The Turing Institute and The University of Edinburgh crussell@turing.ac.uk Lourdes Agapito University College London l.agapito@cs.ucl.ac.uk.