Estimation Human Pose in 3D from 2D image

Shivam Thukral
Department of Computer Science
University of British Columbia
tshivam2@cs.ubc.ca

Kishan Sarpangala
Department of ECE
University of British Columbia
kishans1@ece.ubc.ca

ABSTRACT

In this study, we are trying to solve the 3D pose estimation problem from a single-shot captured from a monocular RGB camera. Our approach is in real-time and robust to (a) Various poses in the wild (b) Multi-Person (c) Can handle upto 15 FPS for video speed (d) Illumination invariant. Our method works as follows; In the first stage we have a Convolution Neural Network (CNN) which jointly estimates 2D and 3D joint locations in real time together without requiring cropped images. In the second stage, a Fully-CNN (Convolutional Neural Network) turns the 2D pose and 3D features for individuals into a complete temporally stable 3D global pose estimate. This method returns full skeleton pose in terms of joint angles for an individual pose. We demonstrated the capability of this system by testing in a range of challenging environments; the results are provided below. This approach is highly effective in 3D character control where cheap RGB camera needs to be deployed.

KEYWORDS

Machine Learning, Computer Vision, Neural Networks, Pose Estimation, Kinematic Skeleton, Monocular Camera, 3D to 2D

1 INTRODUCTION:

Majority of human depictions are in the form of 2D data such as paintings, images and videos. Humans are uniquely capable to identify and understand the facts, ideas and feelings depicted in this 2D data space without a prior information of spatial relationships present in such a complex environment. Machines on the other hand are not innately capable and need to be trained to understand the spatial arrangement of information. In this project, we will focus on a particular instance of this spatial reasoning problem; 3D human pose estimation from a single image. This approach is designed to work for a low cost RGB camera unlike the marker-less 3D motion capture methods that track articulated human poses using active RGB-D(D-Depth) camera or from multi-view cameras. The parameters of the model are learned from data including pose-dependent blend shapes and a regressor which operates on vertices to joint locations[16].

Natural human activities takes place with multiple people in a cluttered scene and hence, such images exhibit self occlusions with body and also occlusions with the environment. The latest approaches leverage the power of deep neural network to capture 3D human pose from a single color image, opening the door to many exciting applications in virtual and augmented reality. Our system produces the skeleton joint angles of an individual in the scene, along with estimates of 3D localization of the subject in the scene relative to the camera. For video, this method works at 15 FPS and delivers close to state-of-art accuracy and temporal stability. The results is comparable with off the self depth sensing based MoCap systems (Motion Capture).

1.1 Challenges:

Convolutional Neural Networks (CNN) have shown impressive results for 3D pose estimation from single RGB images. However, this problem remains an overall challenging task:

- Reconstruction of 3D pose from 2D RGB images is generally considered more difficult than reconstruction of pose in 2D due to larger 3D pose space or depth smoothing issues (Jitter problems) and other ambiguities.
- Estimating human 3D posture from a single image is a challenging task, as 3D reconstruction should be invariant to complex backgrounds (indoor and outdoor), occlusions (full body visibility), illumination (brightness and contrast of the RGB image) and image imperfections such as blurred images due to weather conditions like fog or smoke or badly configured hardware.
- Apart from that, designing a 3D data set is a challenging task. A 3D posing dataset is designed using MOCAP systems that require a setup of multiple sensors and bodysuits that is impractical to use outdoors, making it even more difficult for outdoor datasets. These MOCAP systems has certain requirements like restriction on clothing type; this system prefers or works best in recordings with skin-tight clothing. Such approaches are very expensive as they involve a full studio setup. However, synthetic data can be generated by retargeting MOCAP sequences to 3D avatars, but such results lack realism.

1.2 Motivation:

- **Personal Motivation :** We wanted to have a better understanding of the current trends in Computer Graphics and Computer Vision,this motivated us to pick this topic. Currently both the fields are getting intertwined with the latest deep learning approaches. By replicating the main idea in the selected paper we are ensuring that we develop a deep understanding for the subject and appreciate the complexity in hand.

CPSC535P, 2019, UBC, CA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- **Project Motivation :** Solving this hard problem has various deployment venues like VR, augmented reality, HCI, real-time motion driven 3D game character control, apparel size estimation etc. Apart from this human motion capture has a wide range of applications in computer animations and also other areas such a Bio-Mechanics and Medicine.

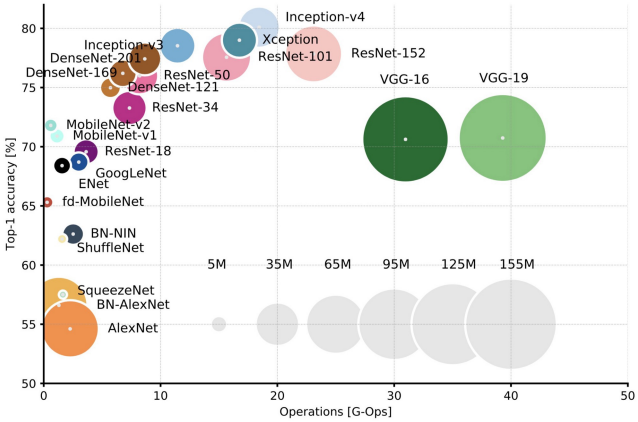


Figure 1: Comparison of different Deep Learning Models

1.3 Contributions:

- A CNN based single-shot, single-person pose estimation method. This approach jointly predicts 2D and 3D joint locations of individuals in the scene. Initially, we used bounding box estimation to compute for 2D pose estimation later we forgo the idea as it was expensive to do bounding box computation. We replaced it with CNN regression of 2D and 3D joints.
- Our system accurately predicts temporally stable joint angles of a 3D skeleton using model based kinematic skeleton fitting against 2D/3D pose predictions.
- A complete algorithm for individual 3D motion capture from a single camera that achieves **real-time performance** without sacrificing reliability or accuracy.
- We did several comparisons of different setups including
  - Experimented with different **Network structures: (VGG-16, ResNet-101, ResNet-50, Inception-V3)** and found which suits our application the best. We explored and experimented with different hyper-parameters. In the end we decided to utilize ResNet-101 and ResNet-50. Please check Figure 1 and Figure 2.
  - Experimented with different **2D and 3D Pose Datasets**
  - We also experimented with different **platforms:** (PyTorch, TensorFlow, Keras), different versions of standard libraries.
- **Model Generation:** Initially, we developed our own flavor of computer vision mode inspired flavor of VGG-16 but the results were not as good. Hence, we experimented with off the shelf CNN structures but despite this we had to tune hyper-parameters on several occasions. After several experiments and hyper-parameter tuning combinations we decided to go with ResNet-100.
- **Literature Survey:** We read several highly cited papers on (a) 3D human pose estimation, (b) different deep learning model architectures, (c) 2D pose to 3D pose estimation and (d) HeatMap generation algorithms. We went ahead and have cited several of the papers on these topics throughout the report.
- Combining different techniques to make the 3D pose estimation pipeline work.

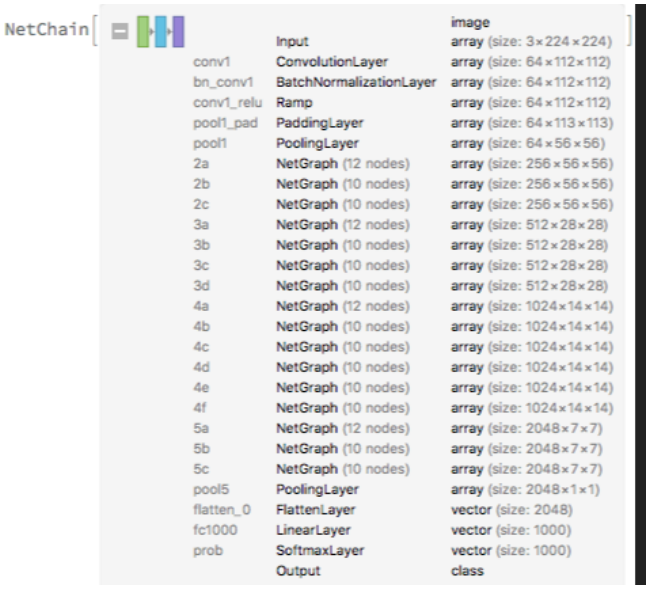


Figure 2: Deep Learning Models Comparison and Configuration of ResNet-50

2 STRETCH GOAL:

We were able to finish the 3D pose estimation task from single RGB image in time and we were able to successfully able to extend this work to real-time videos. The paper which we followed for this approach is Mehta et al[19].

3 RELATED WORK:

Approaches to human pose estimation can be grouped into three broad categories:

- ***Model-based generative methods:*** The Pictorial Structure Model (PSM) consists of two concepts, the first modeling the presence of each body part and the second modeling the spatial relationship between the adjacent parts. Here, PSM can be used to estimate 3D poses by discretizing space. This makes it more complicated, as posing space grows cubically with the resolution of choice.

- ***Discriminative Methods:*** Considers this as an issue of regression. First of all, we remove the features from the image space. Then mapping (using unstructured SVM) is performed from segmented function spaces to joint positions.

- ***Deep Learning Approaches:*** This approach avoids structural dependencies by developing a mapping feature which embeds the human body structure and then learns a template that discovers patterns of human pose from the dataset given (usually very large datasets).

Now we will discuss, some of the state-of-art work which has been done in this field.

- ***Single Person 2D Pose Estimation*** A common approach for 2D pose estimation for humans is first detect the person and then do 2D pose estimation[24][8]. Such methods fails when the detectors are not able to detect the person. This could be due to many reasons; one such reason is occlusion. In this case we first localise the joints of the given human with CNN-based detectors and later find the associations between the joints[23][9].

- ***Single-Person 3D Pose Estimation:*** Monocular single person 3D pose estimation methods train a discriminative predictor for 3D poses[4], and hence do not work well in varied poses, appearances, backgrounds and occlusions. However,they show promising results on standard datasets[10][25]. Most of the 3D datasets used in such training's are restricted to indoor environments. Annotating images for 3D poses is a much harder task. Annotations to their 2D images can be used to find relationships either through dense shape correspondences[1] or through body joint depth ordering constraints[22]. Other approaches divide the problem into two stages : first estimate 2D points and then elevate them to 3D space[27] by neural network regression, or fitting a SPML[16] body model. Some previous research work combine SMPL within the CNN to exploit 3D and 2D annotations. Some papers use the features learned by 2D pose estimation through CNN for 3D pose estimation/calculation[29].

- ***Multi-person 3D Pose Estimation***
  - **George Papandreou et al(2017):** They proposed a clear, but strong, top-down method made up of two phases. They predicted the location and size of boxes that are likely to contain people in the first level, for which they used the Faster RCNN. Then, they estimated the individual person's KeyPoints potentially found in each proposed bounding box; this was done in the second stage. The authors then predicted dense heatmaps and offsets for each joint from of keypoint using a full convolutionary ResNet. To integrate these inputs, a novel aggregation technique was implemented to produce highly localized predictions of the keypoints.[21]

  - **Dushyant Mehta et al(2017):** For general scenes from a monocular RGB camera, the authors proposed a new single-shot approach for multi-person 3D pose estimation. The authors approach used a new occlusion-robust pose-maps (ORPM) that allow other people and objects in the scene to infer the full body pose even under heavy partial occlusions. Though ORPM generates a fixed number of maps encoding all people in the scene's 3D common locations. Body part associations allows to infer 3D pose without explicit bounding box prediction for an arbitrary number of people.[18]

  - **Zhe Cao et al (2016):** The authors presented an approach to detect multiple people's 2D pose in an image efficiently. The approach used a non-parametric representation to learn how to associate body parts with individuals in the image, which they refered to as Part Affinity Fields (PAFs). The architecture encodes a global context, allowing for a greedy phase of bottom-up parsing that maintains high accuracy while achieving real-time performance, regardless of the number of people in the picture. The architecture was built up of two branches of the same sequential prediction system to jointly learn component locations and their association.

- ***Video Based Pose Estimation:***
  - **Rohit Girdhar et al (2017):** Their paper addresses the issue of estimating and monitoring in dynamic, multi-person video human body keypoints. They suggested an extremely lightweight but highly effective solution based on earlier developments in human identification and comprehension of images. Their method operates in two stages: in frames or short clips, keypoint estimation followed by lightweight tracking to produce related keypoint predictions throughout the video. They experimented with Mask R-CNN as well as 3D extension of their unique model for frame-level pose estimation, which utilized temporal information over small clips to produce more accurate frame predictions[7].

4 DATASET:

We have explored following datasets in order to train our CNN based model:

- **MPII:** Human Pose dataset is a state of the art benchmark dataset for evaluation of articulated human pose estimation.The dataset includes around 25K images containing over 40K people with annotated body parts. Overall, the dataset has 410 human activities and each image has been labelled with the corresponding activity.[2]

- **LSP:** This dataset has 2000 pose annotated images of mostly sports people. The images are scaled such that most prominent person in the image is 150 pixels in length. Each image has been annotated with 14 joint locations[13]

- **Human 3.6M**: is a 3D human pose dataset containing 3.6 million 3D human poses and corresponding images. This dataset has 11 professional actors of which 6 are male and 5 are female. This has scenarios involving discussion, smoking,

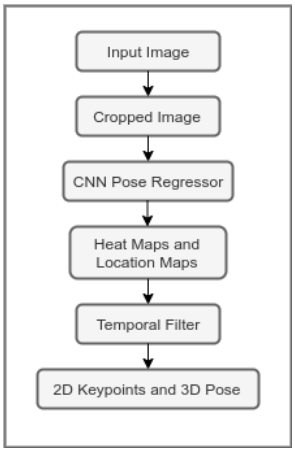


Figure 3: A short overview of the pipeline followed

taking photo, etc. We propose to use a small subset of this dataset for our training purposes[11]

- **MPI-INF-3DHP:** The dataset comprises of 8 subjects (4 male and 4 female), covering 8 activities with 2 sequences per subject. One clothing set is casual everyday apparel, and the other is plain-colored to allow data augmentation. Activities range from walking and sitting to complex exercises poses[17]

5 OVERVIEW:

Our system is capable of obtaining a temporally consistent, full 3D skeletal human pose from a single RGB camera. Our system has two main components. The first is a Convolutional Neural Network (CNN) that regresses 2D and 3D joint locations under the defined conditions of monocular image captured. It is trained on annotated 3D human pose datasets[12] using additionally annotated 2D human pose datasets[14][3] for improved performance in the wild. The second component blends regressed joint positions with a method of fitting kinematic skeletons to create a temporarily stable, camera-relative, complete 3D skeletal pose. The main idea of our method is a CNN that predicts the relative 3D joint positions of 2D and root (pelvis) in real time. Other note worthy things that our model is able to handle are (1) It is able to crop human pose from a single RGB image in real time. (2) Ensures temporarily smooth tracking over time.

6 3D POSE ESTIMATION:

The entire pipeline can be divided into four major steps: (Refer Figure 3 for quick summary)

(1) **Finding Bounding Box and Computing 2D Pose**
- A 2D bounding box is generated on the object of interest. The subject is localised in the frame with 2D bounding box, computed from 2D joint HeatMaps. The most likely joint locations which act as a bounding box detector are provided by the heat map maxima.
- We experimented with Resnet-101[28] but we got better results using VGG-19[26] initially for 2-D pose estimation. We have ReLu pooling and convolution stages within our architecture. In the end we went back to Resnet-101(a flavor of CNN structure) for 3-D pose estimation after further evaluation; the final dataset on which we trained our model was in MPII or LSP.

(2) **3D Pose Regression:**
The 3D pose, represented as a vector of 3D joint positions, relative to the root. The regression is performed on the cropped bounding box generated in step 1. This step also makes use of CNN technique. The CNN structure used in this step is explained below :

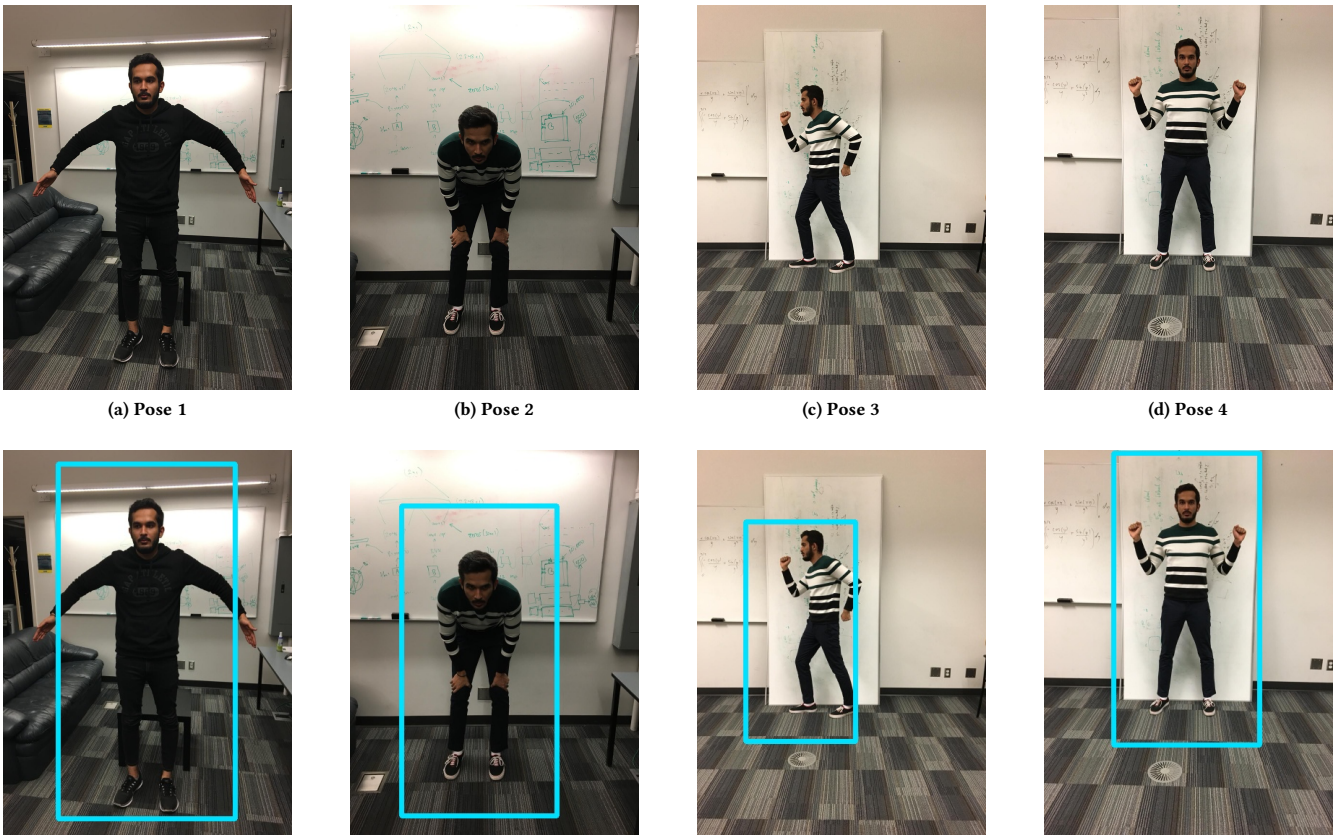
- CNN structure is based on Resnet-101. It is 101 layers deep. The network has an image input size of 224x224.
- The core idea of ResNet structure is to introduce "Identity Shortcut Connection" that skips one or more layers; which resolved the problem of vanishing gradients.
- The skip connection approach mitigates the problem of vanishing gradient by allowing this alternate shortcut for gradient to flow through. Also this approach allows the model to learn an identity function which ensures that the higher layer will perform at least as good as the lower layer, and not worse.
- ResNet-101 Network Converges faster compared to plain counterpart of it.

(3) **Temporal Filtering:** (Required for video processing)
- Casiez et al. (2012) [5] used the 1 filter as a lightweight filter and was specifically designed to reduce jitter and delay when monitoring human movement. This will be deployed in our approach.
- This uses a first-order low-pass filter with an adaptive cutoff frequency: at low speeds, by minimizing jitter, a small cutoff stabilizes the signal, but as speed increases, the cutoff increases to minimize the delay.

(4) **Kinematic Skeleton Fitting:**
The 2D predictions in terms of heat maps maximas from CNN and 3D predictions from CNN pose regressor is fed into temporal filtering and smoothing is performed which is used to find temporarily consistent 3D skeleton pose. This approach combines:

- Predicted 2D and 3D joint positions to fit a kinematic skeleton in a least squares fashion.
- Ensures temporally smooth tracking over time (for video based - extended goal)

In the first step, 2D predictions are temporally filtered and used to obtain 3D coordinates of each joint from the location-map predictions. The resulting 2D and 3D joints are used to



(a) Pose 1 (b) Pose 2 (c) Pose 3 (d) Pose 4

(e) Bounding-Box on Pose 1 (f) Bounding-Box on Pose 2 (g) Bounding-Box on Pose 3 (h) Bounding-Box on Pose 4

Figure 4: Core Dataset: Different Poses to test the algorithm and Bounding-Box

minimise the objective function using Levenberg-Marquardt Algorithm:

$$E_{Total} = E_{IK} + E_{Proj} + E_{smooth} + E_{Depth}$$

- *Inverse Kinematic Term* - E_{IK} : Determines the overall pose by similarity to the 3D CNN output.
- *Projective Term* - E_{Proj} : Determines the root joint location in camera frame and corrects 3D pose by re-projecting it onto the 2D detected KeyPoints.
- *Temporal Smoothing Term* - E_{Smooth} : Temporal stability is enforced with smoothness prior.
- *Depth Smoothing Term* - E_{Depth} : Overcomes the strong depth uncertainty in the monocular reconstruction by penalising large variations in depth.

7 RESULTS:

In this section, we will show the results of 3D pose estimation. In the first set of results, we show Shivam in different poses, these poses are varied to an extent; which captures most of the variations we want to test upon to check the robustness of our model and the overall system.

- **Bounding Box**: The second row displays the bounding box generated for each of human poses using the HOG descriptor. Refer the figure 4 for further details.
- **HeatMaps**: The first row in Figure 5 shows heatmap locations corresponding to human joints. Please note we have sliced the joint heat maps for representation purposes.
- **2D KeyPoints**: Each of the KeyPoint position K_j is simply given by maximum value in the heatmap. Refer Figure 5 for further details.
- **2D-Limb-Structure**: For each of the given poses, the 2d-limb-structure is formed by jointing all the key-points with ts parent joint. Please refer figure 6.
- **3D-Skeleton**: 3D pose estimation for each of the poses can be seen in figure 7. From the results, we can see that the results are temporally stable.

8 LIMITATIONS:

Following are the limitations of this work:

- **Self Occlusions:** When we have significant amount of self occlusions then this algorithm fails to predict the 3D pose.
- **Outlier Poses:** Poses which are away from the training dataset are difficult to handle. Including more data with different poses can potentially solve this problem.
- **Multiple People:** In some cases, the CNN architecture fails to detect or detects incorrectly multiple persons in the image. We feel that this is because of lack of training data. Our model can be made more robust to such cases when we introduce new data.
- **Occluded Faces:** This is a classic problem in such model based algorithms. When the face is occluded the 2D Keypoints are not generated properly.
- **Fast Motion:** Very fast motions can exceed the convergence radius of inverse kinematic optimisation, but such erroneous poses can be recovered by integrating 2D and 3D poses.
- **High-End Hardware:** To generate faster results, high computing power systems with GPU's can be very beneficial.
- **Mirror Image** : Our deep nets currently doesn't support or guarantee equivariant outputs.

9 DISCUSSION:

Annotated dataset is the main hindrance in improving the performance of such deep learning systems. Even the latest annotated 3D pose datasets combined with synthetic data is not enough[6]. Since this problem has a lot of variation ranging form different human poses, shape, appearance and variability in the backgrounds. One way to handle such issue is to train our neural network deeper but this can lead to over-fitting in many cases.

Current, implementation does not handle multiple persons with significant occlusion very well. This is also because of lack of training dataset. Furthermore, the 3D predictions can be improved through the improvements in 2D joint positions by optimising the HeatMaps and the location maps. This issue can be handled by including other approaches such as repeated bottom-up, top-down filtering techniques a crucial way to enhance network performance in combination with intermediate supervision. A recent paper was published in which authors suggested to an architecture which is a network of "stacked hourglass" based on the successive steps of pooling and upsampling that are made to generate a final set of predictions [20]. This approach is supposedly more robust. Following are other noteworthy tasks which needs further elaboration

- **Video Stream:** Initially, our system couldn't handle video stream; the plan was to slice the video into frames. To handle this issue we slowed down the video considerably (FPS needed to be reduced) because our model can handle only 7 images per second. To explain further a video is compromised of several actions; which is than decomposed into poses. To effectively identify a pose and than move to next window to estimate the next human pose was hard. We averaged two poses to generate single pose. We successfully accomplished the stretch goal.
- **HeatMap:** We are reading and looking for better HeatMap algorithm than the one we implemented. Weare trying to implement the Part Affinity Fields corresponding to the limb connecting right elbow and right wrists.(This wasn't a feature we proposed earlier).This is something we want to experiment with. If we have a better heatmap algorithm then we can avoid implementing a robust Bounding Box approach.

10 CONCLUSIONS:

Human 3D pose estimation from a monocular RGB camera is a challenging problem. We presented a real-time approach for a person's 3D motion capture using a single RGB camera. It operates in generic scenes and is robust to occlusions and other objects to some extent. It provides joint angle estimates and localises the subject relative to the camera. Here we designed pose representation, network architecture and model based pose fitting solution. This combines a Fully-Convolutinal CNN that regresses 2D and 3D joint positions and kinematic skeleton fitting method, producing a stable, temporally consistent 3D reconstruction which can be used in embodied VR and interactive character control for computer games to name a few.

11 GITHUB:

Github link is available here[15]

REFERENCES

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7297–7306.

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter V. Gehler, and Bernt Schiele. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), 3686–3693.

[4] Liefeng Bo and Cristian Sminchisescu. 2010. Twin gaussian processes for structured prediction. *International Journal of Computer Vision* 87, 1-2 (2010), 28.

[5] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 2012. 1 Filter: A Simple Speed-based Low-pass Filter for Noisy Input in Interactive Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 2527–2530. https://doi.org/10.1145/2207676.2208639

[6] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. 2016. Synthesizing Training Images for Boosting Human 3D Pose Estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*. 479–488. https://doi.org/10.1109/3DV.2016.58

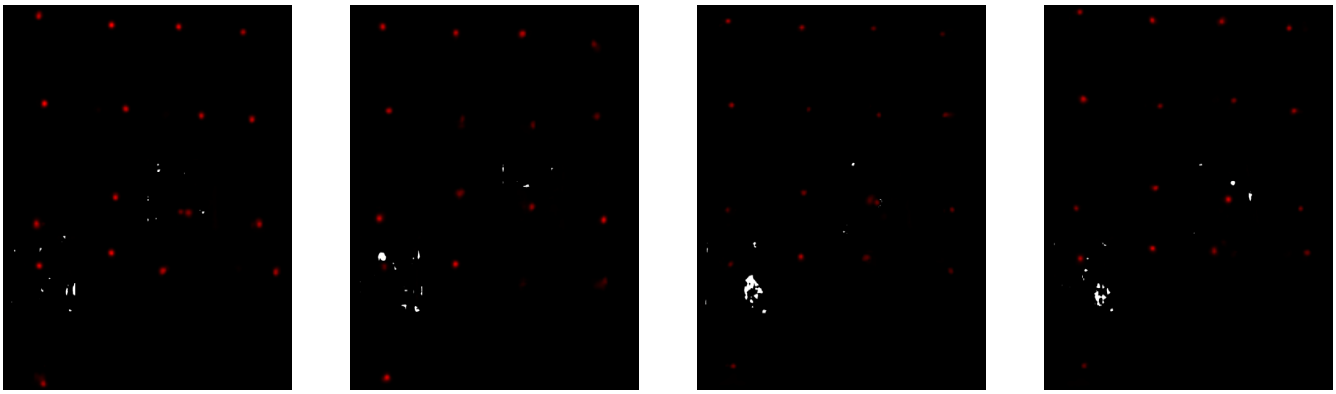
[7] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. 2017. Detect-and-Track: Efficient Pose Estimation in Videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), 350–359.

[8] Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. 2014. Using k-poselets for detecting people and localizing their keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3582–3589.

[9] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele. 2016. ArtTrack: Articulated Multi-Person Tracking in the Wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 1293–1301.

[10] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence* 36, 7 (2013), 1325–1339.

[11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 7 (July 2014), 1325–1339. https://doi.org/10.1109/TPAMI.2013.248



(a) HeatMap location for Pose 1 (b) HeatMap location for Pose 2 (c) HeatMap location for Pose 3 (d) HeatMap location for Pose 4

(e) KeyPoints on Pose 1 (f) KeyPoints on Pose 2 (g) KeyPoints on Pose 3 (h) KeyPoints on Pose 4

Figure 5: HeatMaps and KeyPoints: (Please Zoom in, the points are small in nature)

[12] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36 (2014), 1325–1339.

[13] Sam Johnson and Mark Everingham. 2010. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *Proceedings of the British Machine Vision Conference*. doi:10.5244/C.24.12.

[14] Sam Johnson and Mark Everingham. 2010. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *BMVC*.

[15] Shivam Thukral Kishan Sarpangala. 2019 (accessed Dec 17, 2019). *Estimation Human Pose in 3D from 2D image, GitHub*. https://github.com/ShivamThukral/CPSC535P/tree/master/CPSC535P_Project

[16] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.* 34 (2015), 248:1–248:16.

[17] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. 2017. Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE. https://doi.org/10.1109/3dv.2017.00064

[18] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. 2017. Single-Shot Multi-person 3D Pose Estimation from Monocular RGB. *2018 International Conference on 3D Vision (3DV)* (2017), 120–130.

[19] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. VNect: real-time 3D human pose estimation with a single RGB camera. *ArXiv* abs/1705.01583 (2017).

[20] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked Hourglass Networks for Human Pose Estimation. *CoRR* abs/1603.06937 (2016). arXiv:1603.06937 http://arxiv.org/abs/1603.06937

[21] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Christoph Bregler, and Kevin Murphy. 2017. Towards Accurate Multi-person Pose Estimation in the Wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 3711–3719.

[22] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. 2018. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 459–468.

[23] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. 2016. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4929–4937.

[24] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. 2012. Articulated people detection and pose estimation: Reshaping the future. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3178–3185.

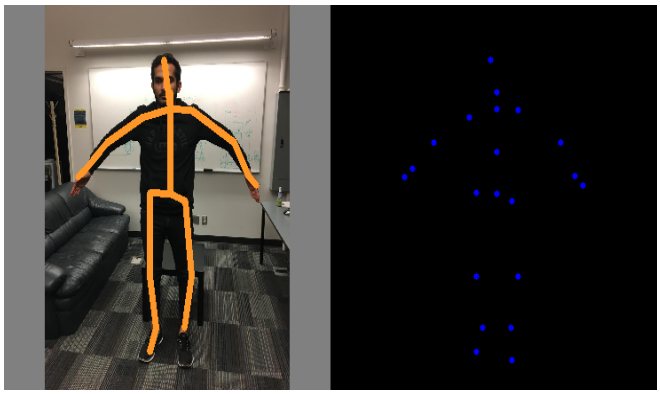
[25] L. Sigal, A. Balan, and M. J. Black. 2010. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision* 87, 1 (March 2010), 4–27.

[26] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

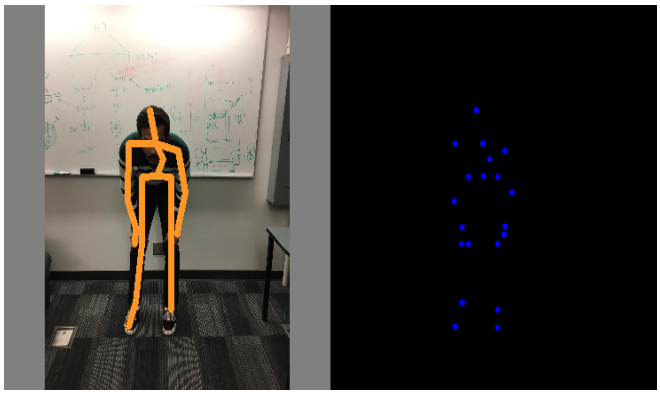
[27] Cristian Sminchisescu and Bill Triggs. 2003. Kinematic jump processes for monocular 3D human tracking. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, Vol. 1. IEEE, I–I.

[28] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *AAAI*.

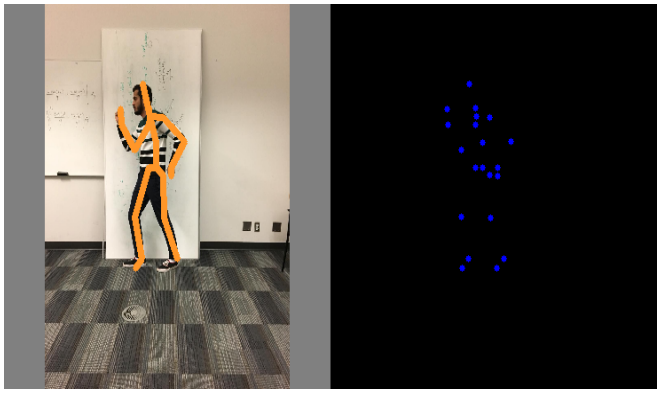
[29] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. 2017. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3941–3950.



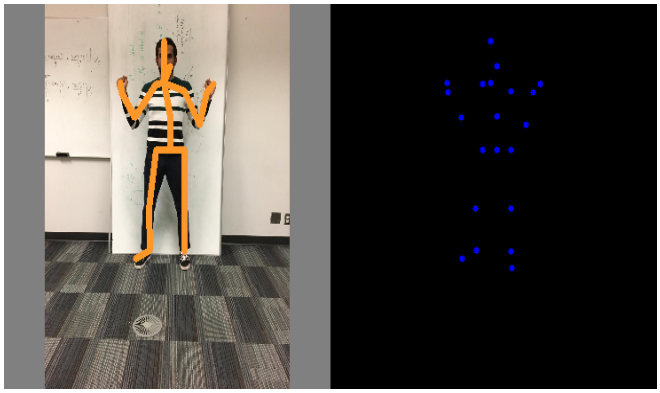
(a) 2D results for Pose 1



(b) 2D results for Pose 2

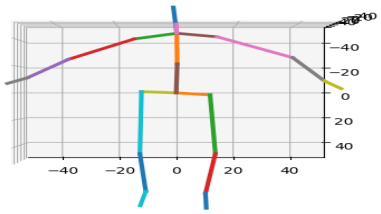


(c) 2D results for Pose 3

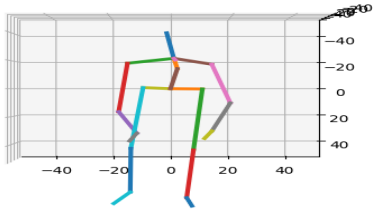


(d) 2D results for Pose 4

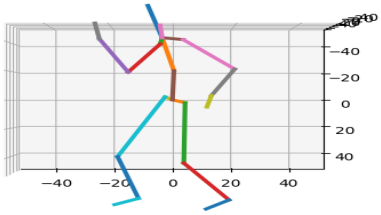
Figure 6: Results in 2D : Keypoints and limb structure drawn side-by-side



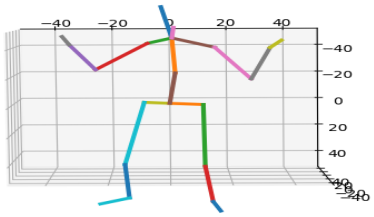
(a) 3D Estimation for Pose 1



(b) 3D Estimation for Pose 2



(c) 3D Estimation for Pose 3



(d) 3D Estimation for Pose 4

Figure 7: 3D Pose Estimation results for all the poses