Pose Estimation for Humanoid Robots

Ofek Peres
Department of Mechanical Engineering
Princeton University
operes@princeton.edu

Divyanshu Pachisia
Department of Mechanical Engineering
Princeton University
divyanshupachisia@princeton.edu

Abstract

Human-Robot interaction is predicted to be an integral part of the future and it will become essential to perform pose estimation on both humans and robots. Given the sparsity of labeled humanoid robot data, this paper investigates whether the state-of-the-art Stacked Hourglass Network [1] trained on the MPII human pose dataset [2] generalizes to predict poses of Atlas, a popular humanoid robot developed by Boston Dynamics. From publicly available videos [3, 4], we extract and label poses of 101 images of Atlas consistent with the format in the MPII Human Pose Dataset. We find that the network generalizes poorly to estimating the pose of humanoid robots, with a Percent of Correct Keypoints (PCKh) metric of 31.5% compared to the 90.9% PCKh metric achieved on the MPII human dataset. Investigating further, we find that appending human features such as faces and clothes (pants) improves the PCKh metric on the humanoid robots to above 45%. This result provides greater interpretability to the Stacked Hourglass Neural Network by showing that it uses distinctly human features, such as faces or clothes, to perform pose estimation.

1 Introduction

Humanoid Robots are autonomous systems whose physical structure resembles that of a human, either fully or in part [5]. Their human-like structure makes humanoid robots well suited to traverse environments in the real world that are traditionally designed for humans [5] and for applications that involve human-robot interaction, such as rehabilitation [6] and service [7]. Given the utility of humanoid robots, a world where humans and humanoid robots coexist is likely to exist in the near future. To enable autonomous systems to work in a such a world, they need to not only predict the behavior of humans but also that of robots. For example, an autonomous car would not only need to predict human pedestrian behavior but also humanoid robot pedestrian behavior.

A key tool in behavior prediction for autonomous systems is *pose estimation*, i.e., locating the position of joints in a body. There has been a lot of recent work on pose estimations *for humans*, with deep learning based methods performing best (see [8] and [9] for a review). This has been fueled by the creation of datasets for human pose such as MPII [2] and FLIC [10]. While there has been some recent work to create datasets for humanoid robots [11], labeled data for humanoid robots is sparse. It would therefore be particularly useful if networks trained on human images would generalize well to humanoid robot data. Given the similarity in the structure of humanoid robots to humans (Figure 1) one may expect this generalization to work well. This paper examines whether this is indeed the case and through this exercise provides interpretability of what features current pose estimation networks are sensitive to.

2 Related Work

The Stacked Hourglass Network [1] is a state-of-the-art deep neural network commonly used for pose estimation achieving above 90% accuracy on the MPII dataset [2]. It's basis lies in an architecture that contains successive steps of up-sampling and pooling which is the origin of the "stacked

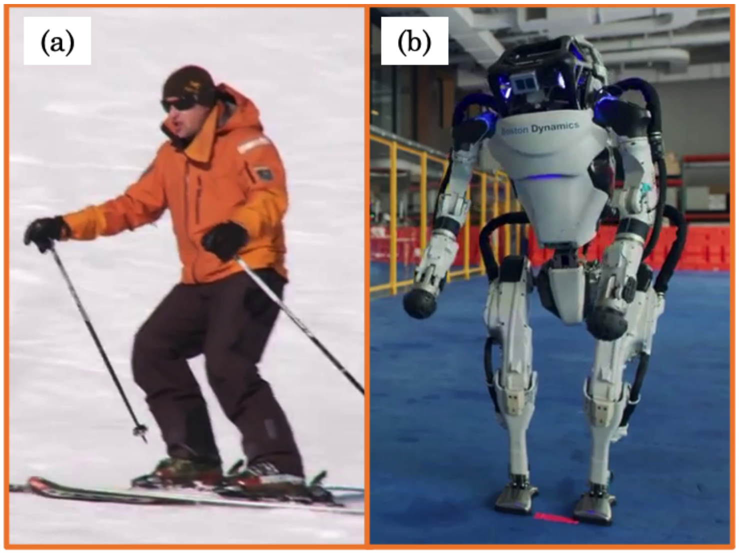

Figure 1: The similarity in structure between a human and humanoid robot. (a) sample image from the MPII dataset and (b) the ATLAS humanoid robot. Given this similarity in structure, we examine whether pose estimation methods trained on humans generalize to humanoid robots.

hourglass" name. While there have been other networks that further improve this work [12, 13], in this paper we utilize the Stacked Hourglass Network due to the availability of code and a pre-trained model at `https://github.com/princeton-vl/pytorch_stacked_hourglass`. We use their pre-trained model that took 3 days on a 12GB NVIDIA TitanX GPU [1] to train on the MPII training dataset (25k images).

There has also been some interesting recent work on pose estimation for humanoid robots [11] that involved training a network on a custom "HumanoidRobotPose" dataset, that specifically focuses on soccer playing robots. While creating a humanoid robot dataset is certainly a promising approach, in this work we take a step back and instead evaluate the generalization of networks trained on human images which are much more readily available. This is a useful exercise not only for the task of humanoid robot pose estimation, but also to help provide interpretability of current state-of-the-art pose estimation networks.

3 Method

We extract and label images of Atlas and then run it through the pre-trained (on the MPII dataset) Stacked Hourglass Network [1], to evaluate its generalizability to humanoid robots. We also augment images of Atlas with a human face and pants to investigate the sensitivity of the network to human features. The code used for this is available at `https://github.com/OfekPeres/COS429_FinalProject`.

3.1 Creating the Datasets

In order to analyze performance of the network, 101 images from 2 Boston Dynamics videos ([3] & [4]) were utilized. The videos were read in programmatically from YouTube via the python pytube package and every thirty frames of the video were written to disk as an image for analysis.

Three additional datasets were formed from these preliminary Atlas images to explore how the network would respond to different human features. A human face was added to each image programatically by placing the face between the "head top" and "upper neck" labels. Similarly, pants were added to Atlas by placing them between the "hip" and "ankle" labels. Finally, both a face and pants were added to each image. This produced a total of 404 images which were analyzed to explore how the Stacked Hourglass Network responds to different human features, helping provide interpretability to the network.

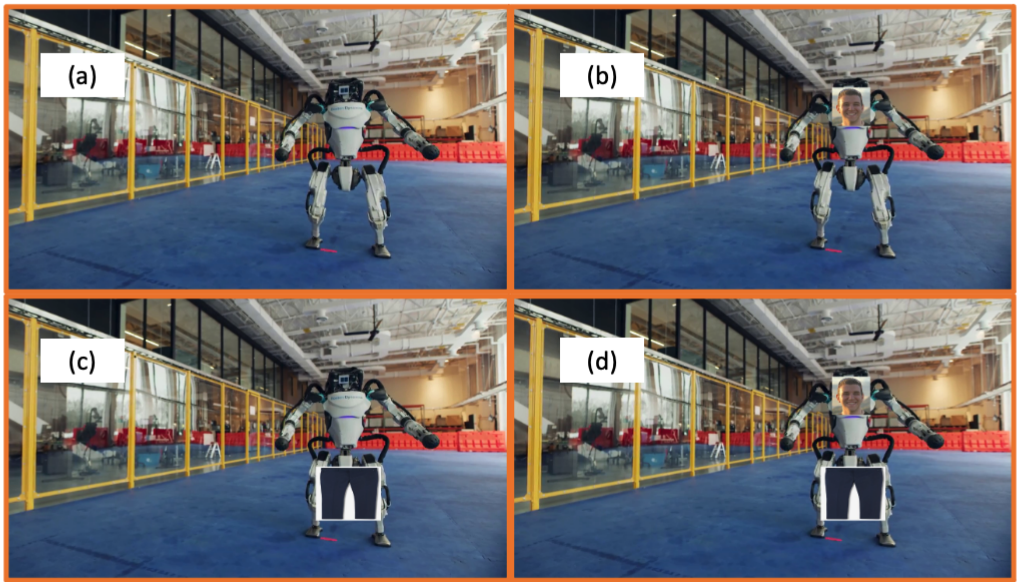

Figure 2: Visualizing the datasets: (a) Original data of Atlas only, (b) Atlas with human face, (c) Atlas with pants, (d) Atlas with human face and pants. 101 images from each of these variations were run through the Stacked Hourglass Network [1].

3.2 Annotating the Dataset

Once the images for the dataset were obtained, a python script was created to automate the process of annotating each of the images. Each image required 17 labels total; 1 for the center and 16 for the joints. Furthermore, the annotations needed to convey if the joint was visually occluded or not. This was handled by utilizing mouse events and click handlers with OpenCV's `imshow` function. Figure 3 visualizes what an annotated image looks like.

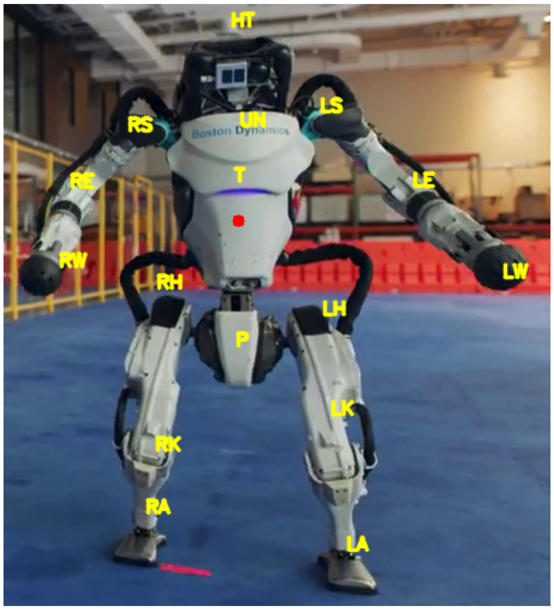

Figure 3: Annotated Atlas. (RA: Right Ankle, RK: Right Knee, RH: Right Hip, P: Pelvis, LH: Left Hip, LK: Left Knee, LA: Left Ankle, RW: Right Wrist, RE: Right Elbow, RS: Right Shoulder, LS: Left Shoulder, LE: Left Elbow, LW: Left Wrist, T:Thorax, UN: Upper Neck, HT: Head Top

3.3 Pose Estimation using the Stacked Hourglass Network [1]

Once the datasets were labeled, they were run through the Stacked Hourglass Network pre-trained on the MPII dataset and results were obtained. The code and network to run this was adapted from the code available at `https://github.com/princeton-vl/pytorch_stacked_hourglass`. It is important to note that the Stacked Hourglass Network deals with ambiguity arising from multiple people (or robots) in the same image by cropping each input image around the "center" coordinates obtained from its annotation data. This means that in order to utilize the network for a new image, the image *must* be correctly annotated.

3.4 Evaluation

The Percent Correct Keypoints (PCK) Metric measures the accuracy of localization of the joints to the ground truth annotated image joints and is a commonly used metric for the task of pose estimation. When the threshold for correctness is defined as 50% of the head segment length, the metric is known as the PCKh Metric; this was the metric used to evaluate performance of the network on each joint of Atlas. If the predicted joint fell within a euclidean distance of 50% of the head segment length it was considered to be correct and otherwise was incorrect. The PCKh measures the % success rate of each joint location according to this distance metric.

4 Results

Figure 4 and Table 1 below demonstrate the Stacked Hourglass Network's results on all 4 datasets as well as the original MPII dataset as a reference. The network did not perform well on unaugmented Atlas images, with an overall score of 31.5% compared to the 90.9% of the MPII human images, indicated that a network trained on humans does not generalize to humanoid robots.

Augmenting the Atlas images with a human face markedly increased not only the head accuracy, but also increased the hip and elbow accuracy, resulting in an overall accuracy of 45.5%. Augmenting the Atlas images with pants had a similar effect, with increased accuracy not only on knees and hips, but also on the head, shoulder, and elbow, resulting in an overall accuracy of 44.8%. Augmenting the Atlas images with both a human face and pants resulted in an overall accuracy of 50.1%.

These results on the augmented datasets seem to imply that the network relies on human features, such as faces or clothes, to perform pose estimation instead of relative positioning between limbs. This is markedly different from a more geometrical approach that allows humans to easily identify joint locations on a humanoid robot.

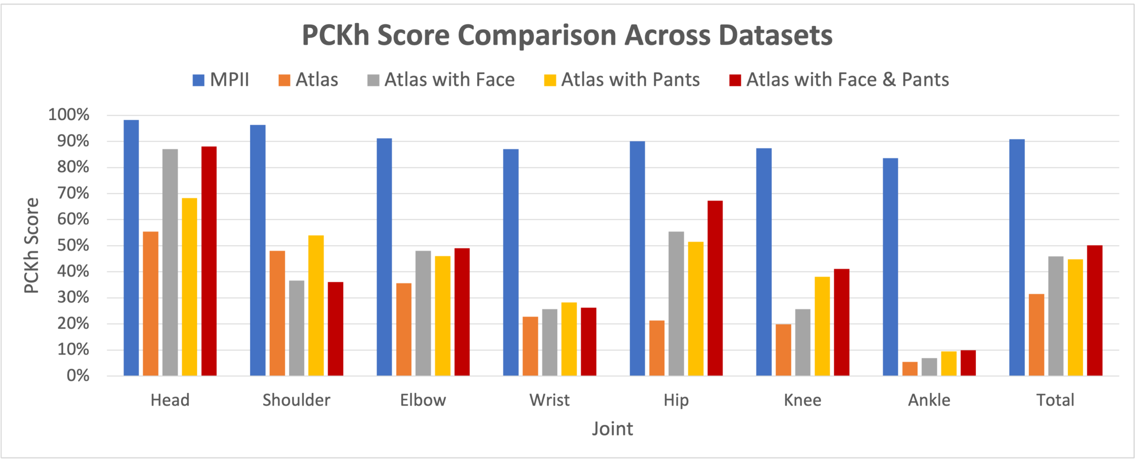

Figure 4: Bar Chart of Stacked Hourglass Network Results on Different Datasets

	Head	**Shoulder**	**Elbow**	**Wrist**	**Hip**	**Knee**	**Ankle**	**Total**
MPII	98.2%	96.3%	91.2%	87.1%	90.1%	87.4%	83.6%	90.9%
Atlas	55.4%	48.0%	35.6%	22.8%	21.3%	19.8%	5.4%	31.5%
Atlas with Face	87.1%	36.6%	48.0%	25.7%	55.4%	25.7%	6.9%	45.9%
Atlas with Pants	68.3%	54.0%	46.0%	28.2%	51.5%	38.1%	9.4%	44.8%
Atlas with Face & Pants	88.1%	36.1%	49.0%	26.2%	67.3%	41.1%	9.9%	50.1%

Table 1: Table of Stacked Hourglass Network [1] Results on Different Datasets

5 Discussion

The lack of generalizability of the Stacked Hourglass Network on both augmented (with faces and pants) and unaugmented humanoid images provides valuable information that could help us understand the inner workings of the network. In this section, we first evaluate the results of each dataset separately and then qualitatively analyze the results on a few representative images shown in Figure 5 and Figure 6.

5.1 Unaugmented Atlas Images

On unaugmented Atlas images we see that the total correct predictions (as defined by the PCKh metric described in Section 3.4) drastically drops from the 90.9% on human images to 31.5% on Atlas images (see Table 1). The most significant decrease was seen on the ankle joints which fell from 83.6% to 5.4%. Overall, the network did worse on the lower half of Atlas (hip - 21.3%, knee - 19.8% and ankle - 5.4%) versus the upper half (head - 55.4%, shoulder - 48.0%, elbow - 35.6% and wrist - 22.8%). This is not surprising since the lower half of Atlas looks less human-like (Figure 1) with gaps in between the pelvis and legs.

5.2 Atlas augmented with a human face

When augmenting Atlas with a human face (see Figure 2 (b)) we see that the accuracy of head detections increases from 55.4% to 87.1% which is not surprising since the humanoid robot head is replaced with a human head. Along with the increase in the accuracy for the head region, the elbow, wrist, hip, knee and ankle prediction accuracy also increases (See row 3 in Table 1). This indicates that the pose estimation network uses the head as a reference to locate the other joints. This is an interesting finding and could lead to explorations such as whether the Stacked Hourglass Network would work well on people with masks or where peoples heads are occluded. It is important to note that the only decrease in accuracy is the shoulder joint which reduces from 48.0% to 36.6%. We hypothesize that this is because Atlas has a wider shoulder compared to its head size and so augmenting a human face results in narrower shoulder detections. Overall augmenting Atlas with a human face increases the total accuracy from 31.5% to 45.9%, demonstrating the use of the head as a reference point for pose estimation.

5.3 Atlas augmented with pants

When augmenting Atlas with pants (Figure 2 (c)) we see an increase in accuracy for hip and knee detection from 21.3% to 51.5% and from 19.8% to 38.1%, respectively. This follows due to the familiarity of the network with people wearing pants and associating them with hips and knees. Across every joint, adding pants increased the network's accuracy. It would be very interesting to see how the network would perform on data of Atlas actually wearing jeans and sneakers as these findings suggest that this simple addition would result in an increase in the usability of human trained networks on humanoid robots.

5.4 Atlas augmented with a human face and pants

Augmenting Atlas with both a face and pants resulted in overall best performance (on humanoid robots) of 50.1%, this is still significantly worse than performance on the MPII dataset of 90.9%. It is interesting to note that despite the overall increase, the shoulder joint accuracy still decreased indicating that the face is more important to shoulder detection than pants are. This follows from the fact that adding only a face reduced shoulder joint accuracy but adding pants increased shoulder joint accuracy. Apart from the shoulder, all other joint accuracies increased and the trend of the upper body performing better than the lower body held, with the exception of hips being located more accurately than wrists.

5.5 Representative Images

In Figure 5 we visualize Atlas in a common running pose mid-step and in Figure 6 we visualize Atlas standing upright with extended arms. For both images, sub-figure (a) contains the reference annotations created for evaluation (see Section 3.2 for details) and in the subsequent sub-figures (b)-(e) we visualize the output predictions for the 4 different datasets of Atlas we created (Figure 2.

5.5.1 Representative Image 1: Atlas Mid-Step

On the unaugmented image (Figure 5 (b)) the predicted joint locations were all incorrect with the exception of the elbow joints. However, when we add a face to the image of Atlas (Figure 5 (c)) it helps reorient all the joints so that they better overlay on the body. In particular, the head-top (HT), upper-neck (UN), wrists (RW and LW), shoulders (RS and LS) and right knee (RK) are all predicted correctly when a face is added. When we only add pants to Atlas (Figure 5 (d)) we find that the knee and ankle predictions are perfectly overlaid on the pants, indicating that the network has learnt to identify clothing as a means to estimate pose. However, in this image, the upper half of the body is not located well and it still offset as in the unaugmented image. Finally, when we add both a face and pants to Atlas (Figure 5 (e)) we find that most predictions are correct, with the pants helping identify the left leg accurately and the face helping orient the joints in the upper half of the humanoid robot body.

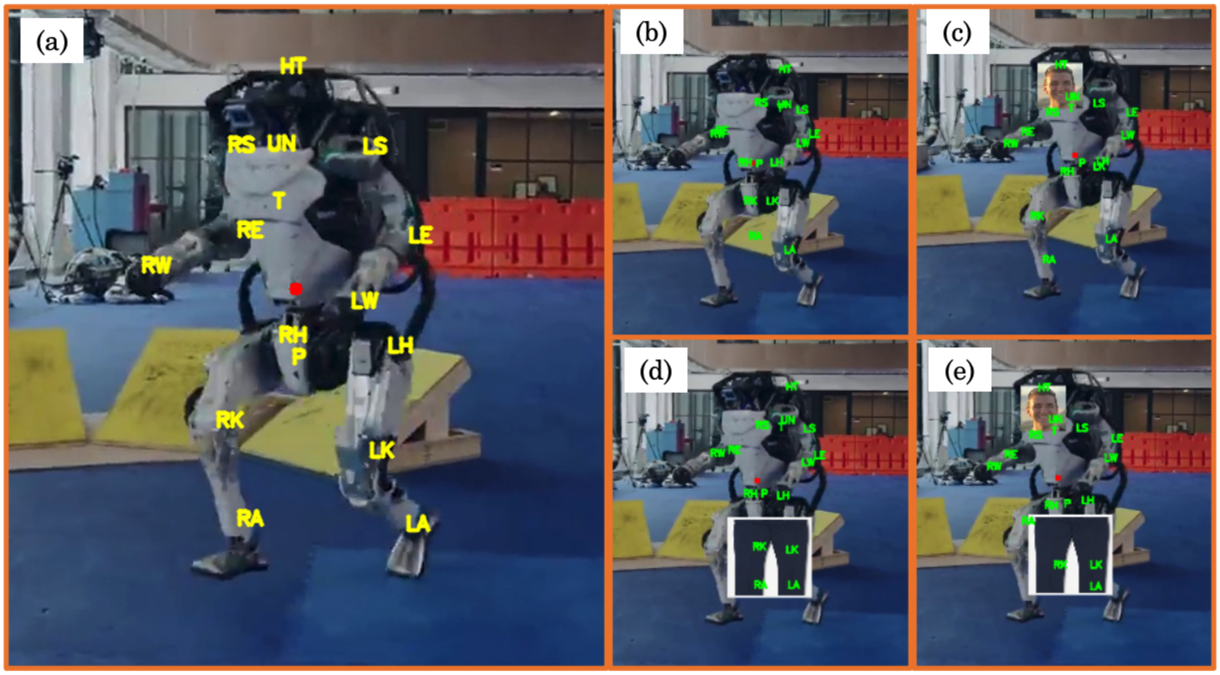

Figure 5: Visualizing Joint Predictions on Atlas on a Representative Image (1 of 2). (a) Manually Labeled Joints (b) Predicted Joint Locations on unaugmented Atlas image (c) Predicted Joint Locations on Atlas with a face (d) Predicted Joint Locations on Atlas with pants (e) Predicted Joint Locations on Atlas with a face and pants

5.5.2 Representative Image 2: Upright Atlas with extended hands

On the augmented Atlas image (Figure 6 (b)), the network did not accurately predict any of the joints. When augmented with a face (Figure 6 (c)), the network performance improved and was able to accurately find the head top (HT), upper neck (UN), and pelvis (P). Augmentation with pants (Figure 6 (d)) resulted in improved lower body predictions, correctly labeling the hips (LH and RH) and pelvis (P) as well as the upper neck (UN), head top (HT), and right arm (RA). Both the head and pants together (Figure 6 (e)) helped the network align many more of the joints and resulted in correct predictions for the head top (HT), upper neck (UN), pelvis (P), hips (LH and RH) and thorax (T).

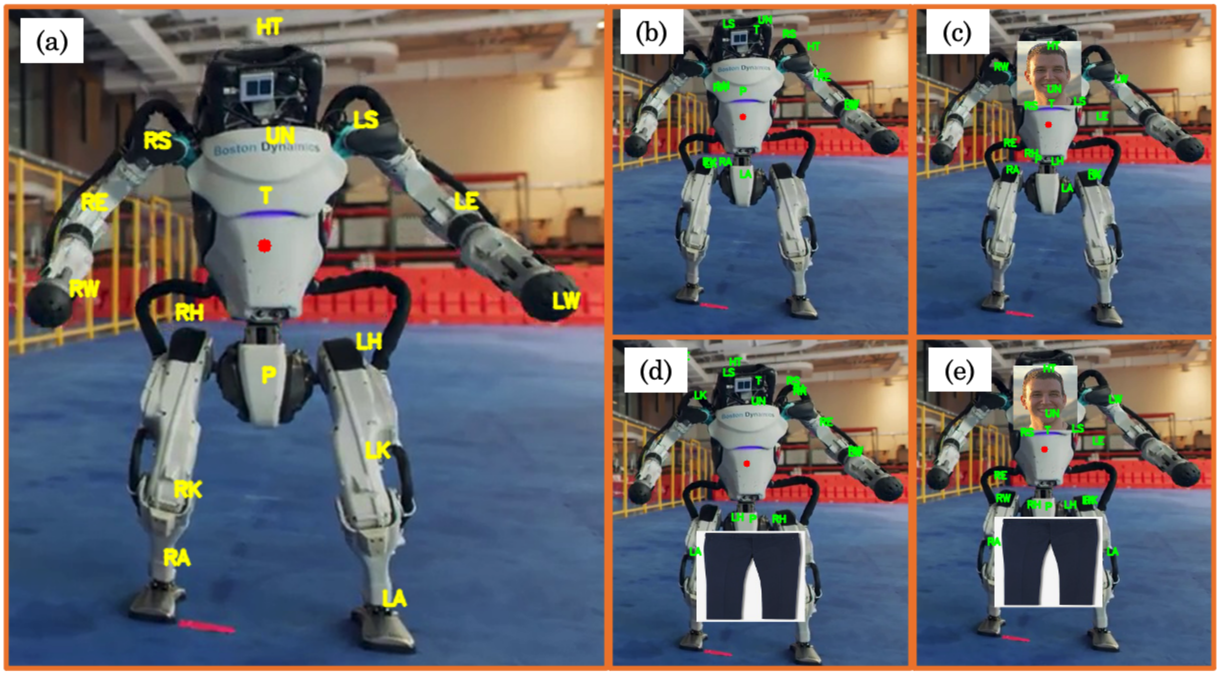

Figure 6: Visualizing Joint Predictions on Atlas on a Representative Image (2 of 2). (a) Manually Labeled Joints (b) Predicted Joint Locations on unaugmented Atlas image (c) Predicted Joint Locations on Atlas with a face (d) Predicted Joint Locations on Atlas with pants (e) Predicted Joint Locations on Atlas with a face and pants

6 Conclusion and Future Work

Through this project we have shown that the Stacked Hourglass Network [1] that was trained on the MPII human pose estimation dataset [2] does not generalize to predict the pose of the humanoid robot, Atlas. Through examining different augmentations to the images of Atlas, namely adding a face and pants to the images in roughly the correct locations, we find that network performance increases significantly. This shows that the current stacked hourglass network [1] relies on distinctly human features (e.g. faces) and clothing (e.g. pants) to locate joints. This is in contrast to a more geometric relationship between relative locations of joints that allows humans to easily identify joints on Atlas. This means that to perform pose estimation in a world with increased human-robot interaction, we would need to augment our datasets with robot images so that autonomous systems, such as self driving cars, can identify robots.

There are several directions of research that this project lays the foundation for. First, it would be interesting to evaluate other networks on these images and compare how they perform to the Stacked Hourglass Network [1]; this would also provide a means to interpret what features these networks are sensitive to. Additionally, it would be very interesting to evaluate the network on images of Atlas in a variety of different outfits ranging from formal tuxedo wear, to winter gear, etc. The goal would be to make atlas look as human as possible via the simple addition of clothing and shoes.

Acknowledgments

Professor Olga Russakovsky
Sunnie Kim
Professor Jia Deng
Alejandro Newell

References

[1] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*, pp. 483–499, Springer, 2016.

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[3] B. Dynamics, "Do you love me?." `https://www.youtube.com/watch?v=fn3KWM1kuAw`, 2020.

[4] B. Dynamics, "Atlas — partners in parkour." `https://www.youtube.com/watch?v=tF4DML7FIWk`, 2021.

[5] R. Bogue, "Humanoid robots from the past to the present," *Industrial Robot: the international journal of robotics research and application*, 2020.

[6] A. Mohebbi, "Human-robot interaction in rehabilitation and assistance: a review," *Current Robotics Reports*, pp. 1–14, 2020.

[7] J. Berg and S. Lu, "Review of interfaces for industrial human-robot interaction," *Current Robotics Reports*, vol. 1, no. 2, pp. 27–34, 2020.

[8] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, and L. Shao, "Deep 3d human pose estimation: A review," *Computer Vision and Image Understanding*, p. 103225, 2021.

[9] R. Josyula and S. Ostadabbas, "A review on human pose estimation," *arXiv preprint arXiv:2110.06877*, 2021.

[10] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *In Proc. CVPR*, 2013.

[11] A. Amini, H. Farazi, and S. Behnke, "Real-time pose estimation from images for multiple humanoid robots," *arXiv preprint arXiv:2107.02675*, 2021.

[12] A. Bulat, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "Toward fast and accurate human pose estimation via soft-gated skip connections," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 8–15, IEEE, 2020.

[13] Z. Su, M. Ye, G. Zhang, L. Dai, and J. Sheng, "Cascade feature aggregation for human pose estimation," *arXiv preprint arXiv:1902.07837*, 2019.