Stereo Human Keypoint Estimation

Kyle Brown
Stanford University
Stanford Intelligent Systems Laboratory
kjbrown7@stanford.edu

Abstract

The goal of this paper is to accurately estimate human keypoints in 3D. This is achieved by deploying twin instances of a deep network in a stereo configuration, and then combining their respective 2D estimations to yield a 3D estimate for each keypoint. As of this writing (project milestone), results have not been obtained. Due to hardware constraints, the system will not actually be tested in real time. However, we hope to demonstrate a framework capable of accurate 3D keypoint estimation in near-real time if implemented with the appropriate hardware.

1. Introduction

Body language is an important mode of human-to-human communication. The way we move says a great deal about our intentions. An artificial agent that can accurately estimate human pose (especially for an arbitrary number of humans simultaneously) in real time is well on its way to effective, safe, and complex interaction with humans. Consider the case of an autonomous vehicle. At a bare minimum, the vehicle must be able to detect and roughly localize pedestrians. Obviously this is prerequisite to avoiding fatal accidents. But what if a police officer standing at an intersection uses hand signals to direct traffic? Will the car be able to follow the officer's commands? Or will the car freeze, unable to comprehend anything more about the situation than the fact that a pedestrian is standing in the road?

This paper considers the problem of human pose estimation within the context of autonomous driving. Specifically, we exam a front-facing stereo camera configuration with one camera placed at the front left corner and another at the front right corner of the windshield.

2. Related Work

One of the most promising (and very recent) achievement in Human Pose Recognition is Mask R-CNN [1]. They demonstrate an extremely versatile and simple network architecture that surpasses the state-of-the-art in all the Microsoft COCO challenges, including Human Keypoint Estimation. Other approaches to human pose estimation involve iterative refinement of a location estimate for a given body part based on the estimated for other body parts. This is the case in Pose Machines [3] and Convolution Pose Machines [5]. Stacked Hourglass [2]. Tompson et al approach the task via joint training of a CNN and graphical model [4].

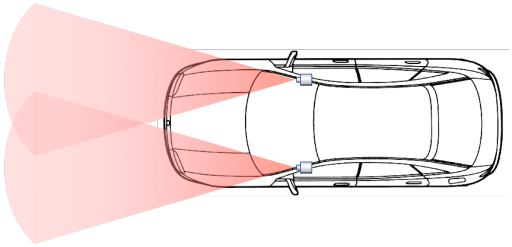

Figure 1. Diagram of the stereo camera configuration

3. Methods

The Deep Network will consist of a pretrained "backbone", the exact architecture of which I have not yet decided. Candidates include DenseNet and various incarnation of ResNet and Inception. I will freeze the backbone layers and then build several layers on top to allow an approach similar to that described in the Mask-RCNN paper. They train a network with three losses simultaneously: A bounding box regression loss, a mask softmax loss, and a keypoint regression loss. He et al report improved keypoint regression when training with all three losses at once (although the converse is not true). So far I have been experimenting with TensorFlow Slim to get the hang of freezing layers and adding custom layers on top. Like He et al, I plan to train the top layers of my network on the Coco dataset, which has very good keypoint, segmentation and bounding box labels.

Once the Deep Network has been trained, I will deploy

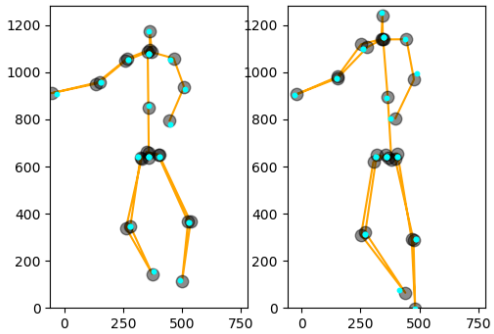

Figure 2. Plot showing two views of a 3D human stick figure. The cyan dots represent ground truth, whereas the black dots represent a simulated prediction attained by perturbing with a displacement drawn from a gaussian distribution

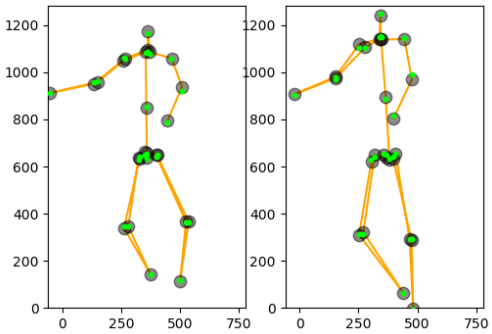

Figure 3. Plot showing two views of a 3D human stick figure. The cyan dots represent ground truth, whereas the black dots represent a simulated prediction attained by perturbing with a displacement drawn from a gaussian distribution

it on pre-recorded videos from two iPhones in a stereo configuration. The videos will be preprocessed to match the input size expected by the network. The keypoint predictions from each network will be passed to a 3D point estimator which will use Weighted Least Squares regression to estimate the 3D location of each keypoint. Over successive frames, the algorithm will calculate a running average of the "limb-length" distances between connected keypoints (i.e. ankle-to-knee, knee-to-hip). As the number of frames seen by the algorithm increases, the running average limb lengths will be used to imposed stricter and stricter constraints on the 3D point estimation. Thus, over time, the algorithm will converge to a 3D skeleton with nearly constant joint lengths.

4. Dataset and Features

As mentioned above, I plan to use a pre-trained architecture (probably trained on imagenet), freeze the first layers, and train the rest on Coco. The relevant labels include human keypoints, masks and bounding boxes.

5. Experiments/Results/Discussion

As mentioned above, I have yet to decide which network backbone to use (yeah, I know–time to get cracking). However, I have been setting up the 3D estimation architecture with simulated data.

Figures 2 and 3 show some initial results for simulated recovery of 3D key point locations from noisy inputs. Note that more key points appear than one would expect. Don't be confused: this is just an artifact of a trick I used to easily plot lines between the joints. Such artifacts will *not* exist in the final results.

I have yet to implement the running average constraints on joint lengths, but things are coming together.

6. Conclusion/Future Work

There's not much to conclude yet, but I will point out that leveraging a stereo camera configuration might be a useful tool for quickly generating high fidelity training labels for new human keypoint recognition. Just sayin.

7. Future Work

8. Appendices

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. 3 2017. 1

[2] A. Newell, K. Yang, and J. Deng. Stacked Hourglass Networks for Human Pose Estimation. 3 2016. 1

[3] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh. Pose Machines: Articulated Pose Estimation via Inference Machines. 1

[4] J. Tompson, A. Jain, Y. Lecun, and C. Bregler. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. 1

[5] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional Pose Machines. 1 2016. 1