Estimating Articulated Human Pose in 3D with Twin Hourglass Networks

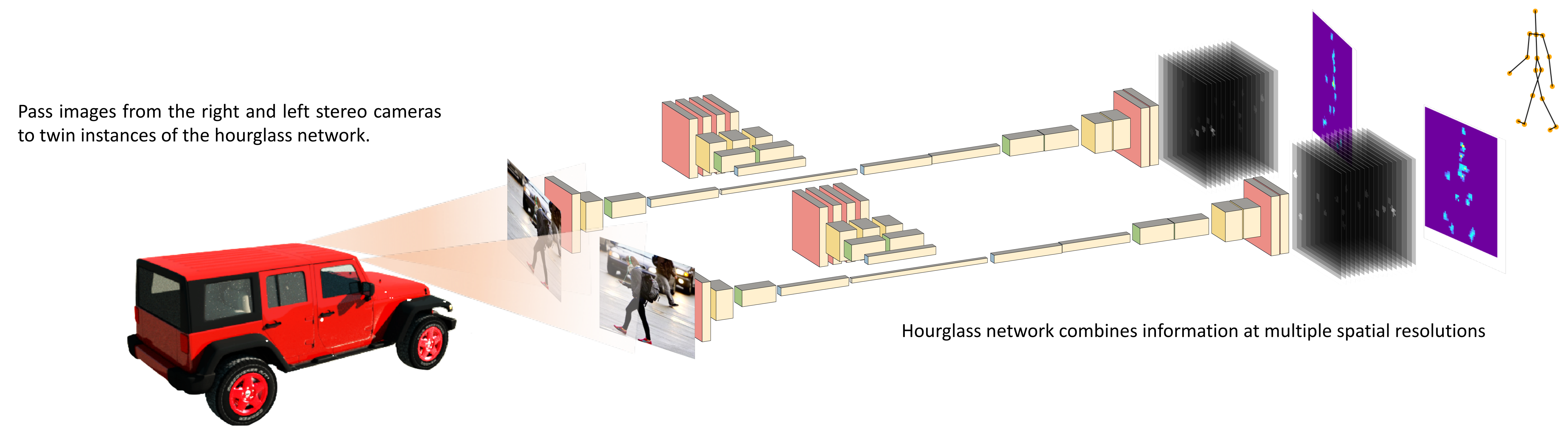
Abstract

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.

Current Status

To be clear, the entire pipeline as shown is not yet implemented end-to-end. The 3D pose estimator is partially implemented in simulation. The 2D key point extractor (i.e. the hourglass network) has only reached 20% training accuracy and is not yet robust enough to enable implementation of the full pipeline. Most of the time up to this point has been spent testing different architectures. The remaining time will be spent trying to improve test accuracy enough to the point where the full 3D pipeline is feasible

Proposed Pipeline



We consider a stereo configuration with cameras mounted on the right and left top corners of an automobile windshield. For simplicity, we consider only a single pedestrian. It is assumed that a region proposal network (or some other mechanism for determining bounding boxes) can successfully draw square bounding region around the pedestrian. At every time step, the cameras each receive a 2D projection of the 3D scene. These two slightly different images are fed through twin instances of the hourglass key point prediction network, yielding two sets (one for each camera) of probability distributions (in the form of heatmaps) for each of the 17 key points under consideration. These initial estimates are used to reconstruct a noisy and not necessarily realistic representation of the target individual's 3D articulated pose. Repeated observation over consecutive frames allows for a running average estimate of limb lengths (i.e. distance between connected key points). By leveraging basic (and true) assumptions about human skeletal structure, the 3D pose estimate can be iteratively improved until converging to a fairly accurate representation.

Network Architecture

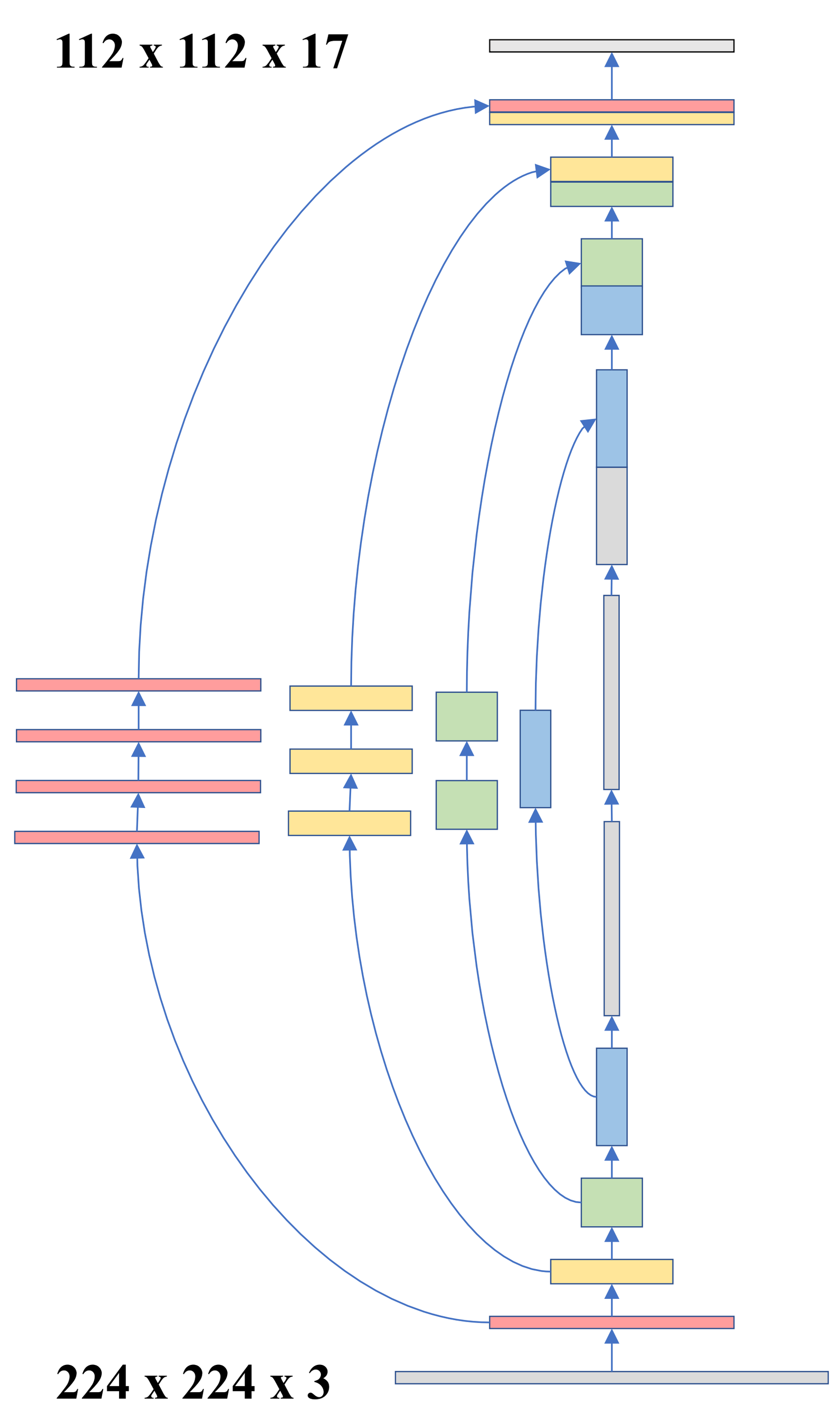
Fully Convolutional
All convolutions (forward and transpose) employ 3x3 filters except for the 1x1 convolutional layer at the prediction layer. Each conv2dl layer is followed by a batch normalization layer and a relu non-linearity.

Bridge Layers
Each layer of down-sampling branches off to a bridge layer that maintains the same spatial resolution until branching back into the network.

Bottleneck
Spatial down-sampling via strided convolution, followed by up-sampling via fractionally-strided transpose convolution.

Modifications to Consider:

- Stacking with residual connections
- Standard regularization techniques like dropout
- Varying depth and thickness of filter banks
- Add Mask branch (Mask RCNN)



Training

Dataset
Microsoft COCO annotated human keypoint dataset:
~80,000 training images
~40,000 validation images

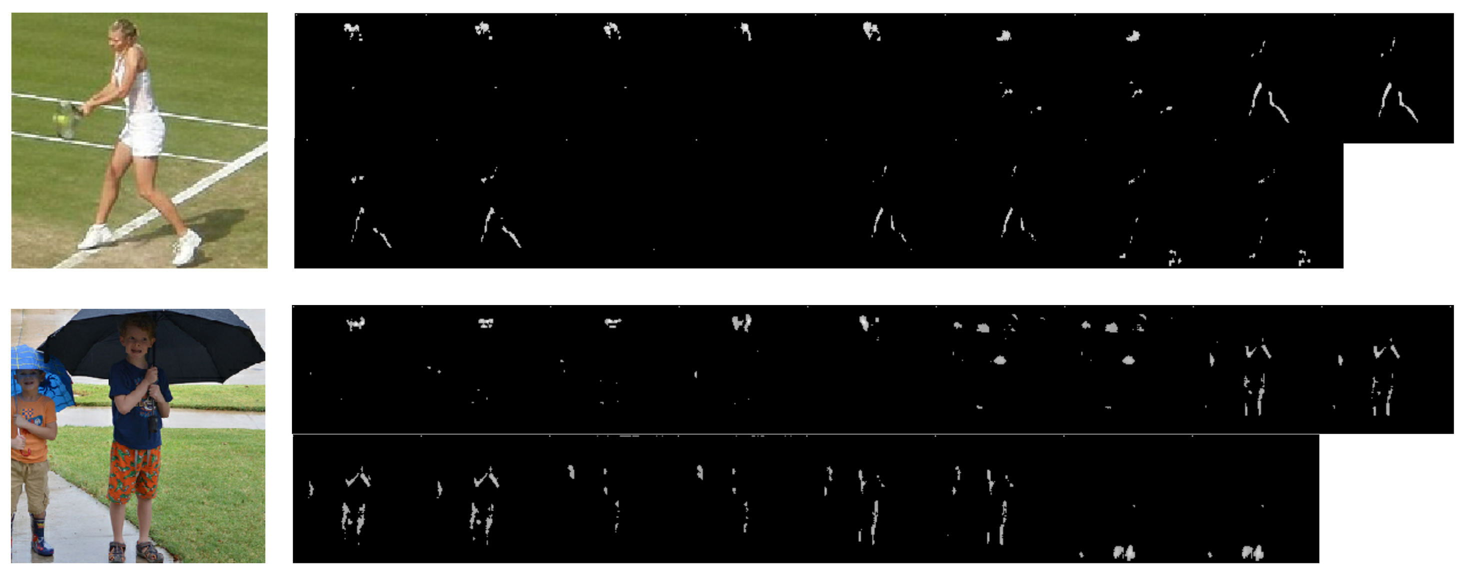
Training Methods

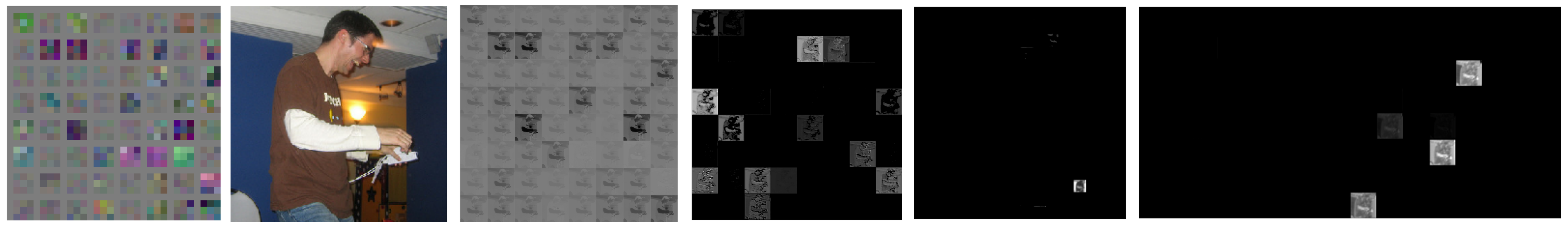
- Softmax Cross-Entropy over 1-hot binary key point masks
- RMSProp Optimizer
- Learning rate decay – exponential staircase decay schedule

Preliminary Results – still workin' on it!

Accuracy - measured by thresholded distance to ground truth key points:
Training Accuracy – 20.1 %
Validation Accuracy – NA
Test Accuracy – NA

It's learning…kinda





1st layer weights

Activations layer by layer

Some good features…but mostly just a bunch of dead neurons

References:
[1] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation."
[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks."
[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," Mar. 2017.
[4] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh, "Pose Machines: Articulated Pose Estimation via Inference Machines."
[5] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," Mar. 2016.
[6] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional Pose Machines," Jan. 2016.
[7] L. Pishchulin *et al.*, "DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation."
[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields *."
[9] J. Tompson, A. Jain, Y. Lecun, and C. Bregler, "Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation."
[10] T. Pfister, J. Charles, and A. Zisserman, "Flowing ConvNets for Human Pose Estimation in Videos."