FULLY CONVOLUTIONAL NERUAL NETWORK FOR BODY PART SEGEMENTATION

David Frank, Richard Kelley & David Feil-Seifer
Department of Computer Science
University of Nevada, Reno
Reno, NV, USA
{davidfrank,rkelley}@unr.edu, dave@cse.unr.edu

ABSTRACT

This paper presents the foundation of a new system for human body segmentation. It is based on a Fully Convolutional Neural Network that uses depth images as input and produces a per-pixel labeling of the image where each pixel has been labeled as a body segment of interest or as non-person. The training data are fully synthetic which allow for large amounts of data to be generated in a relatively short period of time. By using a GPU accelerated implementation of the convolutional neural network, the system is capable of segmenting an image in 8.5 milliseconds. This work will form the basis for more robust system in the future that will be suitable for finding pose skeletons in more cluttered environments.

1 INTRODUCTION

Real time human pose detection is a common issue within HRI (Micelli et al., 2011) and many techniques exist to solve this problem (Moeslund et al., 2006). This work aims to achieve real-time human pose detection by combining two approaches for identifying detail from image data. Shotten et al. (2011) showed the effectiveness of using depth data and the use of synthetic training data. Long et al. (2014) showed that convolutional neural networks can create dense, semantic labels for images. This paper demonstrates a convolutional neural network trained on synthetic data shows promising results for leveraging advances in deep learning for the task of human pose detection.

2 DATA GENERATION

Figure 1: Example labeled images from the data set

All of the data used for training the classifiers were synthetic similar to Shotton et al. (2013) and Keskin et al. (2011). The dataset had 100000 labeled images and was generated in approximately three days. The open source program Blender was used to render all images; the human model was generated by the program MakeHuman.

Each image featured a single human model that was automatically labeled with 6 different body segments with a distinction for left and right for each, giving 12 labels for body parts. The body segments were: Head, torso, upper arm, lower arm, upper leg, and lower leg. Examples are shown in 1.

For the automatically generated data and labels, the body type of the human did not vary; future automatically generated data sets should employ multiple body types. The model was given a ran-

dom pose by rotating each joint within a predefined range of possible values based on the range of motion for that joint. The pose was then checked to ensure that there were no self-collisions e.g., hands passing through the torso. The model stayed the same distance from the sensor and was facing the sensor at all times. This approximated the entertainment senario from Shotten et al. (2011). The depth data were extracted and used as the input for the FCN; the body segments were used as labels for training.

3 NETWORK

3.1 STRUCTURE

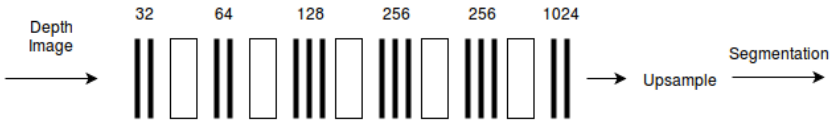


Figure 2: An outline of the fully convolutional network. Convolutional layers are shown as vertical lines with the number of feature planes they contain above them. Max pooling layers are shown as rectangles (all used kernels of 2x2)

The basic structure of the network was similar to Long et al. (2014); the network in that paper used only convolutional layers and downsampling layers to create dense, semantic label predictions. One of the main differences from the network in Long et al. (2014) is that the number of feature planes in most layers have been reduced so that the network would fit into consumer grade GPUs that have at least 3GB of memory, such as a Nvidia GTX 780 or certain models of GTX 960. The skip layers were also not inculded, but may be added in the future. The network structure is detailed in 2. The final layer of this network was shaped such that it had the same height and width as the input data and each pixel had a feature plane for each class. This network was implemented using the deep learning library Torch7 (Collobert et al.).

3.2 TRAINING

For optimization, Stochastic Gradient Descent was used with a learning rate of $.1$, learning rate decay of $.001$, weight decay of $.0001$, and momentum of $.5$. The high initial learning rate allowed the network to quickly learn a basic representation of the data since no prior weights were used to initialize the network. The learning rate decay brought the learning rate down to a level more suitable for fine tuning on later images and epochs. The criterion being optimized was Negative Log Likelihood.

Training the network for one epoch on 85000 images took approximately 10 hours on a GTX 780. A single forward propagation took 8.5 milliseconds on a GTX 960, suitable for real time use.

4 RESULTS

(a) True segmentation (b) Network segmentation

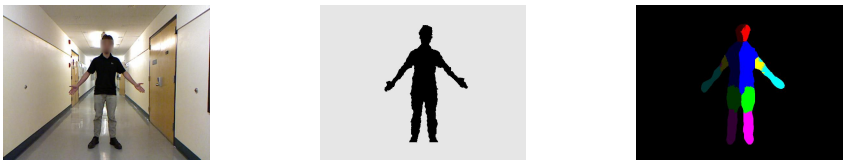
Figure 3: Network predictions for synthetic data

Table 1: Per class accuracy

Class	Non Person	Head Left	Torso Left	Upper Arm Left	Lower Arm Left	Upper Leg Left	Lower Leg Left
Accuracy %	99.15	70.55	91.11	72.89	73.37	80.52	66.31
Class		Head Right	Torso Right	Upper Arm Right	Lower Arm Right	Upper Leg Right	Lower Leg Right
Accuracy %		67.19	81.04	74.56	63.32	72.28	57.18

The results were scored using a pure accuracy measure. After three epochs, the network reached an overall accuracy of 97.25% on a test set of 15000 images, accuracy per class is shown in 1.Examples of the network segmentations are shown in 3. The example images show that confusion can occurs when separate body parts are in close proximity; these sorts of errors may be due to the large amount of down-sampling done in the network and could be mitigated by adding skip layers.

4.1 REAL DATA



(a) Camera view (b) Processed depth data (c) Network prediction

Figure 4: Network performance on real data

Several real depth images were taken with a Microsoft Xbox 360 Kinect. After the floor, walls, and ceiling were removed from the image by simple cropping and thresholding techniques, and then they were put through the trained network for segmentation. Ground truth labelings were not established for these images, but observations on network performance were made. The example 4 shows that the network was able to output sensible labelings for an image that contained noise and a human subject that did not exactly match the synthetic model used for training.

ACKNOWLEDGMENTS

This material is based in part upon work supported by: NASA Space Grant: NNX10AN23H, the Nevada Governor's Office of Economic Development (NV-GOED: OSP-1400872), and Flirtey Technology Pty Ltd., and by Cubix Corporation through use of their PCIe slot expansion hardware solutions and HostEngine. Software used in the implementation of this project include: Blender, MakeHuman, OpenEXR, and Torch7. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NASA, NV-GOED, Cubix Corporation, Blender Foundation, MakeHuman Team, Industrial Light & Magic, Deepmind Technologies, NYU, NEC Laboratories America, or IDIAP Research Institute.

REFERENCES

R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. *Neural Information Processing Systems*. URL http://ronan.collobert.com/pub/matos/2011_torch7_nipsw.pdf.

C. Keskin, F. Kirac, Y.E. Kara, and L. Akarun. Real time hand pose estimation using depth sensors. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 1228–1234, Nov 2011. doi: 10.1109/ICCVW.2011.6130391.

J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014. URL http://arxiv.org/abs/1411.4038.

V. Micelli, K. Strabala, and S. Srinivasa. Perception and control challenges for effective human-robot handoffs. *RSS 2011 RGB-D Workshop*, 2011.

T. Moeslund, A. Hilton, and V. Krger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 2006.

J. Shotten, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1297–1304, 2011.

J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12), 2013.