Indian Institute of Technology, Kanpur
Department of Computer Science

Viewpoint Invariant Human Pose Estimation

Kamlesh Kumar Meena(14299) Gourav Katha(14254)
Supervisor: Prof. Vinay P Namboodiri

Abstract

We propose to work on a model to improve the accuracy to achieve viewpoint invariant model for human pose estimation from a single depth image.This approach tries to embed local patches into a learned viewpoint invariant feature space by computing 12 transformation parameters using spatial transformation network.The model learns to predict the poses in the presence of noise and occlusion by selectively extracting glimpses from each body-joints and predicting the presence of the same.This method uses convolutional(VGG-16) and recurrent(LSTM) network to predict joint locations. To achieve more accuracy we have used iterative error feedback technique which iteratively tries to move joint locations by learning from previous iteration. We finally train our model on Invariant-Top View Dataset (ITOP) which consists of 100K real-world depth images taken from various camera viewpoint.

Acknowledgements

I would like to express my regards to:

- Prof. Vinay P. Namboodiri
- Srinivas Rao
- Utsav Singh
- Prabuddha Chakraborty

Contents

Chapter 1

Introduction

1.1 Motivation

Due to wide popularity of surveillance cameras there has been increasing interest in estimating the pose of human beings. Since these surveillances uses depth sensors ,various efforts are made in using depth maps to efficiently determine the pose of human beings.In real world we face occlusion and other noise.Though human beings can easily determine human pose , we still are not able to determine the pose with high accuracy as achieved in other areas like face detection or recognition.

There has been large amount of work based on both generative and discriminative methods.In generative model we construct a skeleton using templates in top-down manner ,whereas in decriminative model we work in bottom-down manner by locating each body part and then constructing the skeleton.

1.2 Objectives

We tried to tackle the problems faced in the human pose estimation by working on a model which handles the occlusion present in the real life example .To make the model work as

viewpoint invariant we have embeded the local depth image feature into viewpoint invariant feature space.To increase its accuracy we applied iterative feedback model [**?**] to map high order temporal dependencies.Occlusion is handled by learning whether we have certain joint is present in the image and then learning the loss function accordingly.

We have used datasets with images taken from various direction like top,front or side.This dataset enable us to nicely learn and develop the viepoint transfer techniques.

Chapter 2

Related Work

Various methods have been proposed in human pose estimation like deformable part models [FGMR10] , pictorial structures[FH05] to model each body joint independently.Convolutional network which have provided state-of-art result in various vision problem failed to work with human pose estimation as pose represent lower dimensional manifold in the high dimensional input space.So [FZLW15, LLC14] first detect the body parts and then localizes the parts.Another work is based on enforcing global pose consistency on Markov random fields which represents human anatomical constraints[TJLB14].Graphical models are also popular in human pose estimation which encodes dependencies between outputs[CY14].

Since availability of depth maps for input image use of iterative closest point algorithms[GPKT12, GWK05] and database lookups [YWY$^+$11] is widely motivated. [HWLX15] imposes kinematic constraints for improving human pose estimation.With this kernel methods with kinematic chain structures and template fitting models have been proposed.Various discriminative approaches have shown good results.Body segmentation from a single depth image using random forest classifier is used to predict body part location[SSK$^+$13].Other works extending this work includes hough forests[GSK$^+$11], random ferns[HSBA15] , and random tree walks[YJLSHDY15].

Chapter 3

Model

3.1 Model Architecture

3.1.1 Input

The input to our model is depth image of dimension 240 x 320.We extract a set of 15 patches from the image with each centered around the predicted location of the joints.We denote the pose by locating 15 joints.Each of the patches is foveated from the center to form a glimpse.Each glimpse is generated by the predicted location from previous iteration y_{t-1}.We initialize the location to the average pose y_0.

3.1.2 Viewpoint Invariant feature space

Each glimpse with dimension (160,160) is converted to voxel($x^{'} \in R^{H \times W \times D}$) with height H,width W and depth D using the depth maps. A voxel is simply a 3D representation of depth map.Now we use spatial transformation network(STN) to generate a set of 12 3D-transformation parameter for each voxel.These transformation parameters are then used to generate a sampling grid $G \in R^{H \times W \times D \times 3}$.This sampling grid transforms the voxel to viewpoint invariant feature space.

Let each coordinate of G is represented as $G_{ijk} = (x_{ijk}, y_{ijk}, z_{ijk})$.Let V represent the map of the voxel in viepoint invariant space.We define V as below:

$$V_{ijk} = \sum_{a=1}^{H} \sum_{b=1}^{W} \sum_{c=1}^{D} x'_{abc} ker\left(\frac{a - x_{ijk}}{H}\right) ker\left(\frac{b - y_{ijk}}{W}\right) ker\left(\frac{c - z_{ijk}}{D}\right)$$

$$ker(\cdot) = max(0, 1 - |\cdot|)$$

Finally we convert V,our 3D viewpoint invariant feature to 2D viewpoint invariant feature denoted as $U \in R^{H \times W}$ as follows:

$$U_{ij} = \sum_{c=1}^{D} V_{ijc}$$

3.1.3 Neural Networks

We stack the viewpoint invariant feature for each glimpse to feed input tensor of form $H \times W \times J$.We have VGG-16 architecture to process the viewpoint invariant feature obtained.As mentioned before, directly regressing of activations of dense layers is difficult we use a iterative refinement technique.Each iteration is influenced by its previous iteration.We use recurrent network with LSTM [HS97] module to provide recurrent connections between iterations.This helps us in modelling higher-order temporal dependencies.

3.2 Loss Calculation

We have argued to handle occlusion in our model.For this we formulate our optimization problem as multi-task learning problem.We determine whether a body part visible or occluded and predict location to the correct location of the body parts.

3.2.1 Joints Detection

We define a loss function for predicted visibility of joints as follows:

$$Ł_{\alpha} = \sum_{j=1}^{J} \alpha_j log(p_j) + (1 - \alpha_j) log(1 - p_j)$$

$$\alpha_i = \begin{cases} 1 & \text{if } i^{th} \text{ joint is visible} \\ 0 & \text{otherwise a} \end{cases}$$

Here p represent the unnormalized log probabilities generated by the LSTM.

3.2.2 Joint Location

As our ultimate goal is to correctly predict the location of each joint handling occlusion and viewpoint invariances.The loss function we have used to regress the offsets for joint locations is follow:

$$Ł_{\beta} = \sum_{j=1}^{J} l\{\alpha_j = 1\} \left\| \hat{\beta}_j - \beta_j \right\|_2^2$$

Her $l\{\alpha_j = 1\}$ indicates that a particular joint is visible.Hence we only backpropagate pose error if that joint is predicted as visible.

3.2.3 Total Loss

Combining the above mentioned loss function, the global loss function is given by:

$$Ł = Ł_{\alpha} + Ł_{\beta}$$

Chapter 4

Experiments

4.1 Training parameters

Our model has following parameters for training:

Voxel height=40

Voxel width=40

Voxel depth=40

Batch size=10

No of Joints=15

Number of Iterations=4

Learning Prameter= 0.01

Intrinsic camera calibration parameter=0.0035

VGG-16 is used for extracting features from 3D images.In VGG-16,last fully connected layer and softmax is removed and no of nodes reduces to 2048 and output is passed to LSTM layer containing 2048 LSTM units. Weights of network is intialized from a Gaussian with mean=0, standard deviation=0.001 Optimization is done with Adam optimiszer[KB14] [Note: Intrinsic Camera parametrs is used to calulate real world coordinates from single depth images]

4.2 Dataset

We use HDF5 data-set which is available on available on "https://www.albert.cm".This data set contains single depth images and corresponding point cloud of top and side views. Dimensions, attributes, and data types are listed below. The key refers to the (HDF5) dataset name. Let n denote the number of images.

Table 4.1: Depth Maps

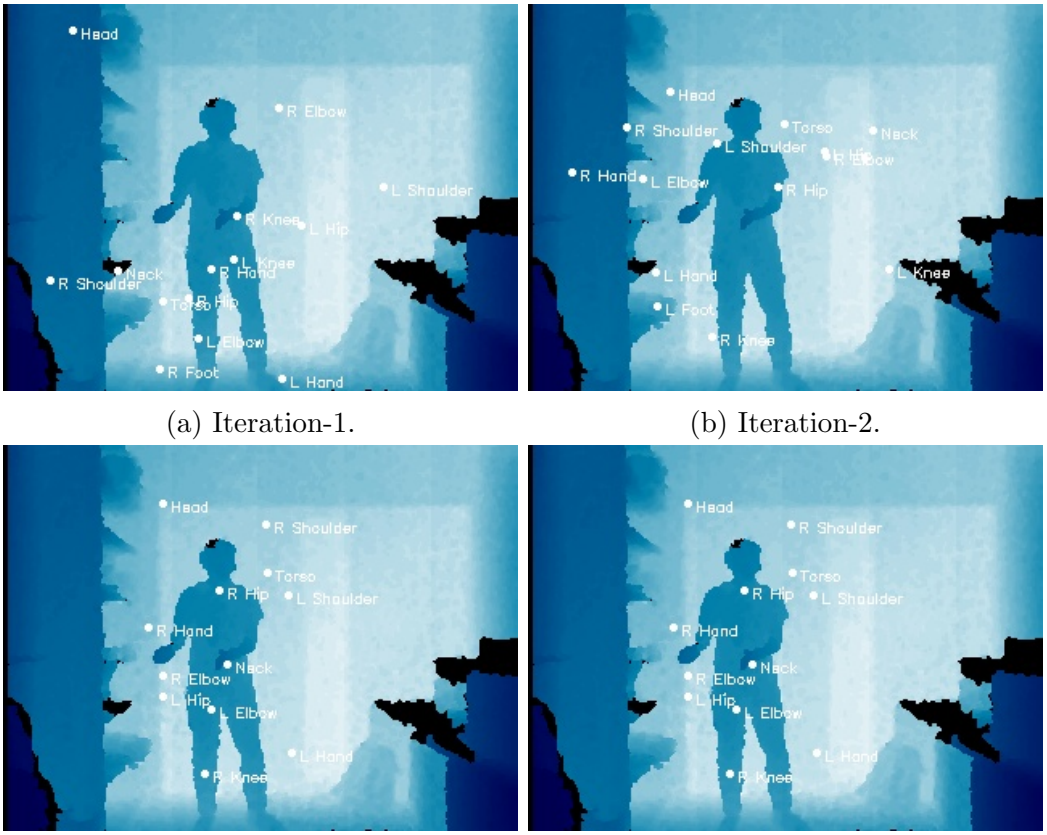
Case	Key	Dimensions	Data Type	Description
1	id	(n,)	8-bit unsigned integer (uint8)	This will show identity of images.
2	data	(n,240,320)	Half precision floating point (float16)	Point cloud containing 76,800 points (240x320) which is represented by a 3D tuple measured in real world meters (m)

Table 4.2: Point-Cloud

Case	Key	Dimensions	Data Type	Description
1	id	(n,)	8-bit unsigned integer (uint8)	This will show identity of images.
2	data	(n,76800,3)	Half precision floating point (float16)	Point cloud containing 76,800 points (240x320). Each point is represented by a 3D tuple measured in real world meters (m).

4.3 Results

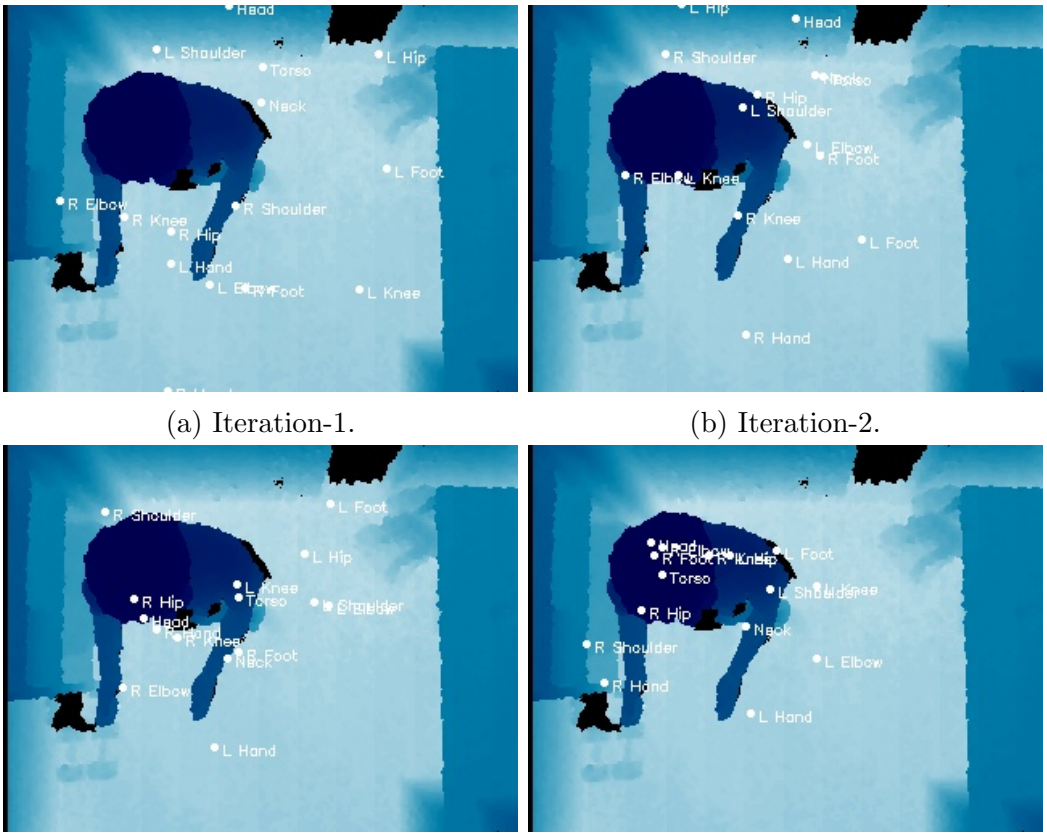
We have trained our model on single depth images HDF5 data-sets for top view and side view images through iterative feed-back learning.We used five iterations to get better results.In each iteration,Our model learns by using result of previous iteration. A summary of our results is given by figures for Side and top views:



(a) Iteration-1. (b) Iteration-2.

(c) Iteration-3. (d) Iteration-4.

Figure 4.1: Testing Result for Side-View-Image.



(a) Iteration-1. (b) Iteration-2.

(c) Iteration-3. (d) Iteration-4.

Figure 4.2: Testing Result for Top-View-Image.

Chapter 5

Conclusion

5.1 Summary of Project Achievements

Through this project we are able to understand a model which learns to work nicely with viewpoint invariant dataset.This model can be used in 3D pose estimation in camera survelliances in banks,hospital,offices etc. A modification of this model can be used to retreive images with similar poses in a dataset.

5.2 Future Work

Further work can be done to improve these result by removing approximation that we have used due to limited time and GPU resources.Since these method uses deep learning techniques the computational time is high compared to non deep learning techniques.Also matrix computations while performing mapping to viewpoint invariant feature space is expensive both in time and space.Work can be done to optimize these computational steps while maintaining the accuracy.

Bibliography

[CY14] Xianjie Chen and Alan L Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems*, pages 1736–1744, 2014.

[FGMR10] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[FH05] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79, 2005.

[FZLW15] Xiaochuan Fan, Kang Zheng, Yuewei Lin, and Song Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1347–1355, 2015.

[GPKT12] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun. Real-time human pose tracking from range data. In *European conference on computer vision*, pages 738–751. Springer, 2012.

[GSK+11] Ross Girshick, Jamie Shotton, Pushmeet Kohli, Antonio Criminisi, and Andrew Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 415–422. IEEE, 2011.

[GWK05] Daniel Grest, Jan Woetzel, and Reinhard Koch. Nonlinear body pose estimation from depth images. In *DAGM-Symposium*, volume 5, pages 285–292. Springer, 2005.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[HSBA15] Nikolas Hesse, Gregor Stachowiak, Timo Breuer, and Michael Arens. Estimating body pose of infants in depth images using random ferns. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 35–43, 2015.

[HWLX15] Li He, Guijin Wang, Qingmin Liao, and Jing-Hao Xue. Depth-images-based pose estimation using regression forests and graphical models. *Neurocomputing*, 164:210–219, 2015.

[KB14] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[LLC14] Sijin Li, Zhi-Qiang Liu, and Antoni B Chan. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 482–489, 2014.

[SSK+13] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[TJLB14] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, pages 1799–1807, 2014.

[YJLSHDY15] Ho Yub Jung, Soochahn Lee, Yong Seok Heo, and Il Dong Yun. Random tree walk toward instantaneous 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2467–2474, 2015.

[YWY+11] Mao Ye, Xianwang Wang, Ruigang Yang, Liu Ren, and Marc Pollefeys. Accurate 3d pose estimation from a single depth image. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 731–738. IEEE, 2011.