Learning Pose Specific Representations by Predicting Different Views

Wenjie Niu

June 18,2018

Abstract

The labeled data required to learn pose estimation for articulated objects is difficult to provide in the desired quantity, realism, density, and accuracy. To address this issue, we develop a method to learn representations, which are very specific for articulated poses, without the need for labeled training data. We exploit the observation that the object pose of a known object is predictive for the appearance in any known view. That is, given only the pose and shape parameters of a hand, the hands appearance from any viewpoint can be approximated. To exploit this observation, we train a model that - given input from one view - estimates a latent representation, which is trained to be predictive for the appearance of the object when captured from another viewpoint. Thus, the only necessary supervision is the second view. The training process of this model reveals an implicit pose representation in the latent space. Importantly, at test time the pose representation can be inferred using only a single view. In qualitative and quantitative experiments we show that the learned representations capture detailed pose information. Moreover, when training the proposed method jointly with labeled and unlabeled data, it consistently surpasses the performance of its fully supervised counterpart, while reducing the amount of needed labeled samples by at least one order of magnitude.[4]

1. Introduction

In this work we aim to estimate the pose of the hand given a single depth image. For this task, the best performing methods have recently relied heavily on models learned from data [1],[3],[8],[5]. Even methods which employ a manually created hand model to search for a good fit with the observation, often employ such a data-driven part as initialization or for error correction [2],[6],[7],[9]. Unfortunately, data-driven models require a large amount of labeled data, covering a sufficient part of the pose space, to work well.

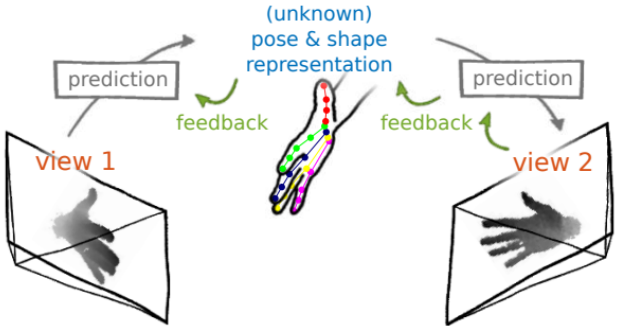

Figure 1. **Sketch for learning a pose specific representation from unlabeled data.** We learn to predict a low-dimensional latent representation and, subsequently, a different view of the input, solely from the latent representation. The error of the view prediction is used as feedback, enforcing the latent representation to capture pose specific information without requiring labeled data.

However, for the task of estimating the pose of articulated objects, like the human hand, it is especially expensive to provide accurate annotations for a sufficient amount of real world data. The articulated structure and specific natural movements of the hand frequently cause strong self-occlusions. Together with the many 3D points to be annotated, this makes the annotation procedure a huge effort for human annotators.

A largely unexplored direction to cope with this challenge is to exploit unlabeled data, which is easy to obtain in large quantities. We make a step towards closing this gap and propose a method that can exploit unlabeled data by making use of a specific property of the pose estimation task. We rely on the observation that pose parameters1 are predictive for the object appearance of a known object from any viewpoint. That is, given the pose parameters of a hand, the hands appearance from any viewpoint can be estimated. The observation might not seem helpful upfront, since it assumes the pose - which we want to estimate - to be given. However, the observation becomes helpful if we capture the scene simultaneously from different viewpoints.

By employing a different camera view, we can guide the training of the pose estimation model (see Fig. 1). The guid-

Table 1. **Pre-training from unlabeled data.** Mean joint error and standard deviation on the NYU-CS dataset for different pretraining methods and numbers of labeled samples, *n*.

n	Autoencoder	PreView(Ours)	
100	48.0 0.76	33.4 1.18	30.4%
1,000	47.2 0.29	29.6 0.32	37.3%
10,000	47.3 0.08	29.0 0.14	38.7%
43,640	47.1 0.08	29.0 0.09	38.4%

ance relies on the fact that from any set of pose parameters, which accurately specify the pose and rough shape of the hand, we necessarily need to be able to predict the hands appearance in any other view. Hence, by capturing another view, this additional view can be used as a target for training a model, which itself guides the training of the underlying pose representation.

More specifically, the idea is to train a model which - given the first camera view - estimates a small number of latent parameters, and subsequently predicts a different view solely from these few parameters. The intuition is that the small number of parameters resemble a parameterization of the pose. By learning to predict a different view from the latent parameters, the latent parameters are enforced to capture pose specific information. Framing the problem in this way, a pose representation can be learned just by capturing the hand simultaneously from different viewpoints and learning to predict one view given the other.

Given the learned low-dimensional pose representation, a rather simple mapping to a specific target (*e.g.*, joint positions) can be learned from a much smaller number of training samples than required to learn the full mapping from input to target. Moreover, when training jointly with labeled and unlabeled data, the whole process can be learned end-to-end in a semi-supervised fashion, achieving similar performance with one order of magnitude less labeled samples. Thereby, the joint training regularizes the model to ensure that the learned pose representation can be mapped to the target pose space using the specified mapping.

We show the specificity of the learned representation and its predictiveness for the pose in qualitative and quantitative experiments. Trained in a semi-supervised manner, the proposed method consistently outperforms its fully supervised counterpart, as well as the state-of-the-art in hand pose estimation - even if all available samples are labeled. For the more practical case, where the number of unlabeled samples is larger than the number of labeled samples, we find that the proposed method performs on par with the baseline, even with one order of magnitude less labeled samples.

2. Experiments

The results on the NYU-CS dataset are shown in Tab. 1. We compare our method to pre-training using an autoencoder because of its close relation. In particular, the autoencoders target is the input view, whereas our method aims to predict a different view. For a fair comparison, we use the same architecture, *i.e.*, the same number of parameters and training algorithm for the autoencoder and the proposed method for predicting different views (*PreView*).

References

[1] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 1933. 1

[2] P. Krejov, A. Gilbert, and R. Bowden. Guided optimisation through classification and regression for hand pose estimation. *Computer Vision and Image Understanding*, 2016. 1

[3] M. Oberweger, P. Wohlhart, and V. Lepetit. Training a feedback loop for hand pose estimation. In *IEEE International Conference on Computer Vision*, 2015. 1

[4] G. Poier, D. Schinagl, and H. Bischof. Learning pose specific representations by predicting different views. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[5] J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan. Depth-based hand pose estimation: Data, methods, and challenges. In *IEEE International Conference on Computer Vision*, 2015. 1

[6] D. Tang, J. Taylor, P. Kohli, C. Keskin, T. K. Kim, and J. Shotton. Opening the black box: Hierarchical sampling optimization for estimating human hand pose. In *IEEE International Conference on Computer Vision*, 2015. 1

[7] J. Taylor, L. Bordeaux, T. Cashman, B. Corish, C. Keskin, T. Sharp, E. Soto, D. Sweeney, A. Topalian, and A. Topalian. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *Acm Transactions on Graphics*, 2016. 1

[8] S. Xiao, W. Yichen, L. Shuang, T. Xiaoou, and S. Jian. Cascaded hand pose regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1

[9] Q. Ye, Y. Shanxin, and K. Tae-Kyun. Spatial attention deep net with partial PSO for hierarchical hybrid hand pose estimation. In *European Conference on Computer Vision*, 2016. 1