Proceedings of the OAGM Workshop 2018 DOI: 10.3217/978-3-85125-603-1-27

Being lazy at labelling for pose estimation

Georg Poier David Schinagl Horst Bischof
Institute for Computer Graphics and Vision
Graz University of Technology
Austria

Arguably, the best performing methods for real-time pose estimation heavily rely on models learned from data [1], [3]. Such data-driven models need to be trained on a large corpus of labeled data to deliver the expected results. Labeled data, however, is difficult to provide in the desired quantity, realism and accuracy. This is the case, in particular, for the task of 3D pose estimation of articulated objects. In this work we show that using a simple observation, which we exploit in a self-supervised training procedure, we are able to substantially reduce the amount of required labels. Using the proposed training procedure we are able to reach the same performance, even with one to two orders of magnitude less labeled samples.

The featured presentation covers the work published in [2]. Implementation source code and additional material can be found at `https://poier.github.io/PreView`.

I. Introduction to the task

In this work we focus on the exemplary task of 3D hand pose estimation from depth data. We want to learn a model, which – given a single depth image capturing a human hand – estimates the hand's pose. For this task the articulated structure and specific natural movements of the hand frequently cause strong self-occlusions. This not only makes the task more difficult, it also makes the currently necessary annotation procedure a huge effort for human annotators.

II. A simple observation

A largely unexplored direction to reduce the annotation effort is to exploit unlabeled data. This direction bears the advantage that unlabeled data for this task is easy to obtain in large quantities. Hence, we present a method that exploits unlabeled data by making use of a specific property of the pose estimation task. The method is based on the observation that pose parameters are predictive for the object appearance of a known object from any viewpoint. That is, given the pose parameters of a hand, the hand's appearance from any viewpoint can be estimated. This observation might not seem helpful upfront, since it assumes the pose – which we want to estimate – to be known. However, the observation becomes helpful if we capture the scene simultaneously from different viewpoints.

III. Exploitation of the observation

With a different camera view, we can guide the training of the pose estimation model. More specifically, 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. That is, by training a model which estimates a small number of latent parameters from the first camera view, and subsequently predicts a different view solely from these few parameters, these parameters become very predictive for the object pose. Setting the task up this way, a pose representation can be learned by simply capturing the hand simultaneously from different viewpoints and learning to predict one view given the other.

Using the low-dimensional pose representation learned from unlabeled data, 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, the model can easily be trained end-to-end – jointly with labeled and unlabeled data – in a semi-supervised fashion.

IV. Experimental results

Through an experimental evaluation we show that using the semi-supervised training procedure 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.

In a more practical experiment we investigate the case where the number of unlabeled samples is larger than the number of labeled samples and find that the proposed method performs on par with the baseline, even with one order of magnitude less labeled samples. This indicates that 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.

In additional qualitative and quantitative experiments, we investigate the representations learned without any labeled data. In this way, we find that the proposed training procedure vastly improves the specificity of the learned representation and its predictiveness for the pose compared to related approaches towards learning without labels.

References

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. CVPR*, 2017.
[2] G. Poier, D. Schinagl, and H. Bischof, "Learning pose specific representations by predicting different views," in *Proc. CVPR*, 2018.
[3] J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan, "Depth-based hand pose estimation: Data, methods, and challenges," in *Proc. ICCV*, 2015.