Cross-modal Deep Variational Hand Pose Estimation

Wenjie Niu

June 20,2018

Abstract

The human hand moves in complex and highdimensional ways, making estimation of 3D hand pose configurations from images alone a challenging task. In this work we propose a method to learn a statistical hand model represented by a cross-modal trained latent space via a generative deep neural network. We derive an objective function from the variational lower bound of the VAE framework and jointly optimize the resulting cross-modal KLdivergence and the posterior reconstruction objective, naturally admitting a training regime that leads to a coherent latent space across multiple modalities such as RGB images, 2D keypoint detections or 3D hand configurations. Additionally, it grants a straightforward way of using semisupervision. This latent space can be directly used to estimate 3D hand poses from RGB images, outperforming the state-of-the art in different settings. Furthermore, we show that our proposed method can be used without changes on depth images and performs comparably to specialized methods. Finally, the model is fully generative and can synthesize consistent pairs of hand configurations across modalities. We evaluate our method on both RGB and depth datasets and analyze the latent space qualitatively.[3]

1. Introduction

Hands are of central importance to humans in manipulating the physical world and in communicating with each other. Recovering the spatial configuration of hands from natural images therefore has many important applications in AR/VR, robotics, rehabilitation and HCI. Much work exists that tracks articulated hands in streams of depth images, or that estimates hand pose [1],[2],[4],[6] from individual depth frames. However, estimating the full 3D hand pose from monocular RGB images only is a more challenging task due to the manual dexterity, symmetries and selfsimilarities of human hands as well as difficulties stemming from occlusions, varying lighting conditions and lack of accurate scale estimates. Compared to depth images the RGB case is less well studied.

Recent work relying solely on RGB images [5] proposes a deep learning architecture that decomposes the task into several substeps, demonstrating initial feasibility and providing a public dataset for comparison. The proposed architecture is specifically designed for the monocular case and splits the task into hand and 2D keypoint detection followed by a 2D-3D lifting step but incorporates no explicit hand model. Our work is also concerned with the estimation of 3D joint-angle configurations of human hands from RGB images but learns a cross-modal, statistical hand model. This is attained via learning of a latent representation that embeds sample points from multiple data sources such as 2D keypoints, images and 3D hand poses. Samples from this latent space can then be reconstructed by independent decoders to produce consistent and physically plausible 2D or 3D joint predictions and even RGB images.

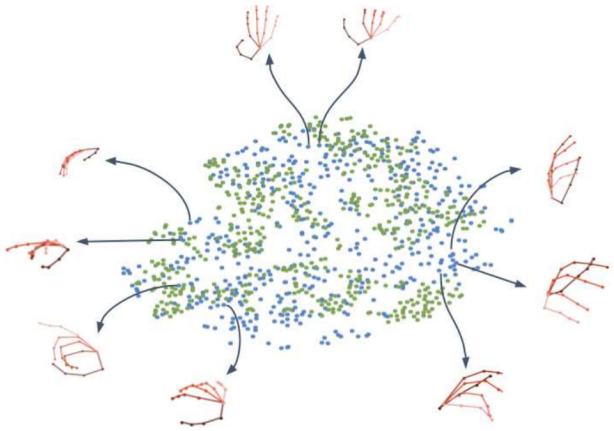

Figure 1. **Cross-modal latent space.** t-SNE visualization of 500 input samples of different modalities in the latent space. Embeddings of RGB images are shown in blue, embeddings of 3D joint configurations in green. Hand poses are decoded samples drawn from the latent space. Embedding does not cluster by modality, showing that there is a unified latent space. The posterior across different modalities can be estimated by sampling from this manifold.

In this work we propose to learn a single, unified latent space via an extension of the VAE framework. We provide a

derivation of the variational lower bound that permits training of a single latent space using multiple modalities, where similar input poses are embedded close to each other independent of the input modality. Fig. 1 visualizes this learned unified latent space for two modalities (RGB & 3D). We focus on RGB images and hence test the architecture on different combinations of modalities where the goal is to produce 3D hand poses as output. At the same time, the VAE framework naturally allows to generate samples consistently in any modality.

We deploy the VAE framework that admits cross-modal training of such a hand pose latent space by using various sources of data representation, even if stemming from different data sets both in terms of input and output. Our cross-modal training scheme, illustrated in Fig. 2, learns to embed hand pose data from different modalities and to reconstruct them either in the same or in a different modality.

Fig. 2, illustrates our proposed architecture for the case of RGB based handpose estimation. In this setting we use two encoders for RGB images and 3D keypoints respectively. Furthermore, the architecture contains two decoders for RGB images and 3D joint configurations.

References

[1] M. Oberweger and V. Lepetit. Deepprior++: Improving fast and accurate 3D hand pose estimation. In *IEEE International Conference on Computer Vision Workshop*, 2017. 1

[2] M. Oberweger, P. Wohlhart, and V. Lepetit. Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*, 2015. 1

[3] A. Spurr, J. Song, S. Park, and O. Hilliges. Cross-modal deep variational hand pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[4] D. Tang, H. J. Chang, A. Tejani, and T. Kim. Latent regression forest: Structured estimation of 3D articulated hand posture. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1

[5] D. Tang, H. J. Chang, A. Tejani, and T. Kim. Learning to estimate 3D hand pose from single RGB images. In *International Conference on Computer Vision*, 2017. 1

[6] C. Wan, A. Yao, and L. V. Gool. Hand pose estimation from local surface normals. In *European Conference on Computer Vision*, 2016. 1

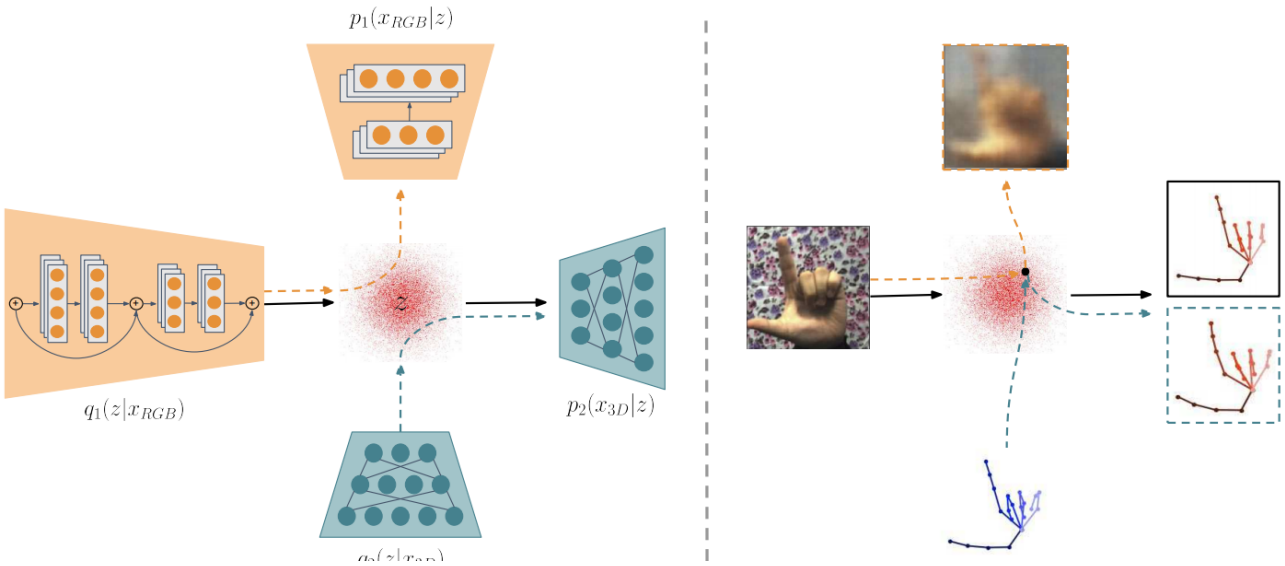

Figure 2. **Schematic overview of our architecture.**Left: a cross-modal latent space z is learned by training pairs of encoder and decoder q, p networks across multiple modalities (e.g., RGB images to 3D hand poses). Auxilliary encoder-decoder pairs help in regularizing the latent space. Right: The approach allows to embed input samples of one set of modalities (here: RGB, 3D) and to produce consistent and plausible posterior estimates in several different modalities (RGB, 2D and 3D).