GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB —Extended Abstract—*

Franziska Mueller[1,2] Florian Bernard[1,2] Oleksandr Sotnychenko[1,2] Dushyant Mehta[1,2]
Srinath Sridhar[3] Dan Casas[4] Christian Theobalt[1,2]

[1] MPI Informatics [2] Saarland Informatics Campus [3] Stanford University [4] Univ. Rey Juan Carlos

Abstract

We address the highly challenging problem of real-time 3D hand tracking based on a monocular RGB-only sequence. Our tracking method combines a convolutional neural network with a kinematic 3D hand model, such that it generalizes well to unseen data, is robust to occlusions and varying camera viewpoints, and leads to anatomically plausible as well as temporally smooth hand motions. For enhancing the training data for our CNN we propose a geometrically consistent image-to-image translation network. To be more specific, we use a neural network that translates synthetic images to "real" images, such that the so-generated images follow the same statistical distribution as real-world hand images while preserving geometric properties (such as hand pose). We demonstrate that our hand tracking system outperforms the current state-of-the-art on challenging RGB-only footage.

1. Introduction

Estimating the 3D pose of the hand is a long-standing goal in computer vision with many applications such as in virtual/augmented reality and human–computer interaction. While there is a large body of existing works that consider marker-free image-based hand tracking or pose estimation, many of them require depth cameras [7, 5] or multi-view setups [1] which are less ubiquitous, more expensive, and does not work in all scenes. In contrast, we address these issues and propose a new algorithm for *real-time skeletal 3D hand tracking* with a *single color camera* that is robust under object *occlusion and clutter*. Recent developments that consider monocular RGB markerless hand tracking [4, 10] come with clear limitations. For example, the approaches by Simon *et al*. [4] estimates only 2D joint locations. Recently, Zimmermann and Brox [10] presented a 3D hand pose estimation method from monocular RGB which, however, only obtains relative 3D positions and struggles with occlusions.

Inspired by recent work on hand and body tracking [3], we combine CNN-based 2D and 3D hand joint predictions with a kinematic fitting step to track hands in global 3D from monocular RGB. The major issue of such (supervised) learning-based approaches is the requirement of suitable *annotated* training data. While it is feasible to manually annotate 2D joint locations in single-view RGB images, it is impossible to accurately annotate in 3D due to the inherent depth ambiguities. One way to overcome this issue is to leverage existing multi-camera methods for tracking hand motion in 3D [1]. However, the resulting annotations would lack precision due to inevitable tracking errors. Some works render synthetic hands for which the perfect ground truth is known [3, 10]. However, CNNs trained on synthetic data may not always generalize well to real-world images. Hence, we propose a method to *generate suitable training data* by performing image-to-image translation between synthetic and real images. Using this enhanced data, a regression network for 2D and normalized 3D joint positions can be trained. Since the full global 3D pose can be derived from neither output in isolation, we fit a kinematic hand model in an energy minimization framework which combines all available information. In summary, our main contributions are:

- The first real-time system that tracks *global 3D hand pose* from unconstrained monocular RGB images.
- A novel geometrically consistent GAN that performs *pose-preserving* image-to-image translation.
- Based on this network, we are able to *enhance synthetic hand image datasets* such that the statistical distribution resembles real-world hand images.
- A *new RGB dataset* with annotated 3D hand joint positions. We overcome existing datasets in terms of size (>260k frames), image fidelity, and annotation precision.

*full paper accepted as Spotlight at CVPR 2018 [2]

2. Method

Our proposed approach can be divided into two parts as shown in Fig. 1: (1) enhancing our hand pose training data by image-to-image translation with a novel GAN, (2) real-time hand pose estimation using a combination of CNN-based joint position regression and kinematic model fitting.

GANerating Training Data: We impose two strong requirements on our data enhancement method. First, we want to be able to train on *unpaired images* so that we can easily collect a large-scale real hands dataset. Second, we need the algorithm to preserve the pose of the hand such that the annotations of the synthetic images are still valid for the translated images. To this end, we leverage the seminal work on CycleGANs [9], which successfully learns various image-to-image translation tasks with unpaired examples. We extend it with a *geometric consistency loss* which improves the results in scenarios where we only want to learn spatially localized (*e.g.* only the hand part) image-to-image conversions, producing pose-preserving results with less texture bleeding and sharper contours. Once this network is trained, we can use it to translate any synthetically generated image into a "real" image while preserving the perfect (and inexpensive) ground truth annotation. We denote images as "real" (in quotes), or *GANerated*, when we refer to synthetic images after they have been processed by our translation network such that they follow the same statistical distribution as real-world images.

Joint Regression and Model Fitting: Using annotated RGB images produced by our GAN, we train a CNN, which we call *RegNet*, that jointly regresses image-space 2D and root-relative 3D hand joint positions. For achieving a tighter coupling of 2D and 3D predictions, we propose to use a projection layer *ProjLayer*. While the skeletal hand model in combination with the 2D predictions are sufficient to estimate the global translation of the hand, the relative 3D positions resolve the inherent ambiguities in global rotation and articulation which occur in the 2D positions.

In the kinematic model fitting step, we are interested in estimating the 26 degrees of freedom Θ (3 for global translation, 3 for global rotation, 20 joint angles) which describe the full 3D hand pose. Thus, we minimize the energy

$$E(\Theta) = E_{\mathrm{2D}}(\Theta) + E_{\mathrm{3D}}(\Theta) + E_{\mathrm{limits}}(\Theta) + E_{\mathrm{temp}}(\Theta)\,, \quad (1)$$

where $E_{\mathrm{2D}}(\Theta)$ and $E_{\mathrm{3D}}(\Theta)$ are inverse kinematics constraints trying to pose the skeleton such that its joints match the 2D and 3D predictions, and $E_{\mathrm{limits}}(\Theta)$ and $E_{\mathrm{temp}}(\Theta)$ are regularizers for joint angle limits and temporal smoothness.

3. Experiments

We quantitatively and qualitatively evaluate our method and compare our results with other state-of-the-art methods on a variety of publicly available datasets. For that, we use the Percentage of Correct Keypoints (PCK) score, a popular criterion to evaluate pose estimation accuracy. PCK defines a candidate keypoint to be correct if it falls within a circle (2D) or sphere (3D) of given radius around the ground truth. In Fig. 2 we show qualitative results on community or vintage RGB video. In particular, we show 3D hand tracking in YouTube videos, which demonstrates the generalization of our method.

Ablative study: In Fig. 3 (left) we compare the accuracy when training our joint regression network *RegNet* with different types of training data: synthetic images only, synthetic images plus color augmentation, and synthetic images in combination with GANerated images, where for the latter we also considered additionally using the *ProjLayer* in *RegNet*. While we evaluated the *RegNet* on the entire Stereo dataset [8] comprising 12 sequences, we did *not train on any* frame of the dataset for this test. We show that training on purely synthetic data leads to poor accuracy. While color augmentation on synthetic images improves the results, our GANerated images significantly outperform standard augmentation techniques which validates the argument for using GANerated images.

Comparison to state-of-the-art: Fig. 3 (center) evaluates our detection accuracy on the Stereo dataset, and compares it to existing methods [8, 10]. We followed the same evaluation protocol used in [10], *i.e.* we train on 10 sequences and test it on the other 2. Our method outperforms all existing methods. Additionally, we test our approach *without* training on any sequence of the Stereo dataset, and demonstrate that we still outperform some of the existing works. This demonstrates the generalization of our approach. Fig. 3 (right) shows the 2D PCK, in pixels, on the Dexter+Object [6] and EgoDexter [3] datasets. We significantly outperform Zimmerman and Brox (Z&B) [10], which fails under difficult occlusions. Note that we cannot report 3D PCK since [10] only outputs root-relative 3D, and these datasets do not have root joint annotations.

4. Conclusion

Most existing works either consider 2D hand tracking from monocular RGB, or they use additional inputs, such as depth images or multi-view RGB, to track the hand motion in 3D. Our proposed approach goes one step ahead with regards to several dimensions: our method obtains the *absolute* 3D hand pose from monocular RGB by kinematic model fitting, is *more robust* to occlusions, and *generalizes better* due to enrichment of our synthetic data such that it resembles the distribution of real hand images. Our experimental evaluation demonstrates these benefits as our method significantly outperforms previous work [10].

Acknowledgements: This work was supported by the ERC Starting Grant CapReal (335545). Dan Casas was supported by a Marie Curie Individual Fellow, grant 707326.

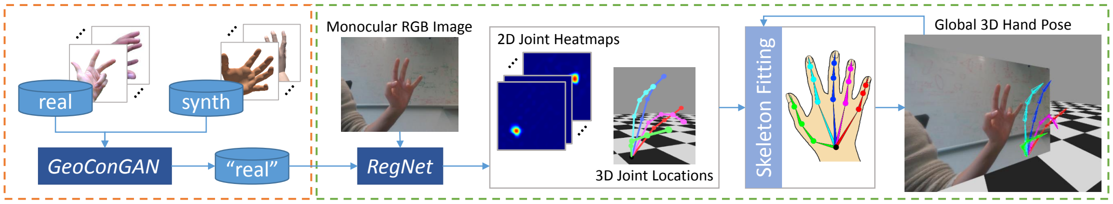

Figure 1: Pipeline of our real-time system for monocular RGB hand tracking in 3D.

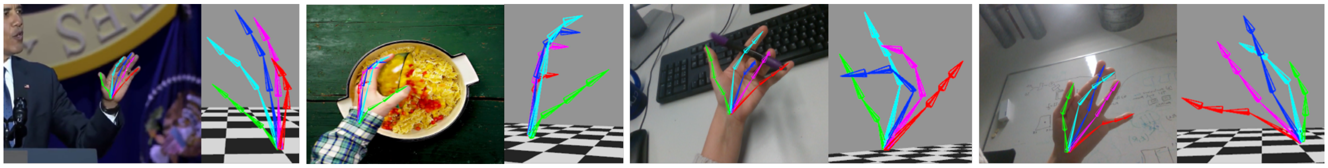

Figure 2: Qualitative results on YouTube videos (left) and a live webcam stream (right).

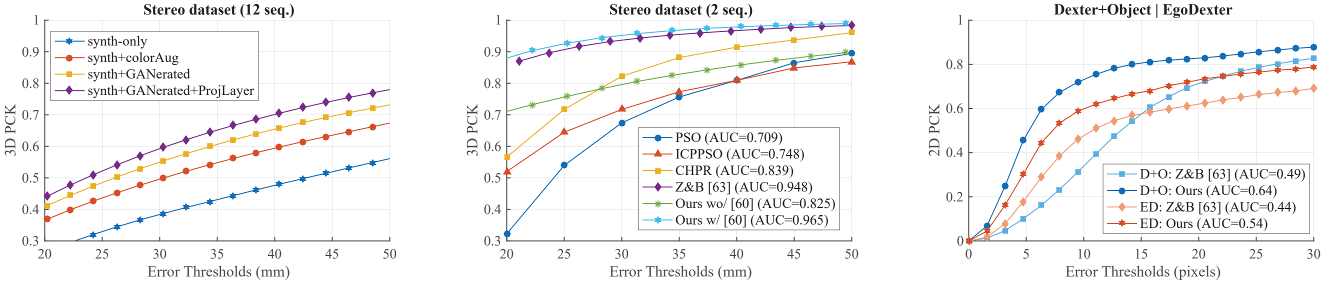

Figure 3: Quantitative evaluation. (left) Ablative study on different training options. (center, right) 3D and 2D PCK comparison with state-of-the-art methods on publicly available datasets.

References

[1] L. Ballan, A. Taneja, J. Gall, L. V. Gool, and M. Pollefeys. Motion Capture of Hands in Action using Discriminative Salient Points. In *European Conference on Computer Vision (ECCV)*, 2012.

[2] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[3] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *International Conference on Computer Vision (ICCV)*, 2017.

[4] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[5] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt. Fast and Robust Hand Tracking Using Detection-Guided Optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[6] S. Sridhar, F. Mueller, M. Zollhöefer, D. Casas, A. Oulasvirta, and C. Theobalt. Real-time Joint Tracking of a Hand Manipulating an Object from RGB-D Input. In *European Conference on Computer Vision (ECCV)*, 2016.

[7] J. Taylor, L. Bordeaux, T. Cashman, B. Corish, C. Keskin, T. Sharp, E. Soto, D. Sweeney, J. Valentin, B. Luff, et al. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics (TOG)*, 35(4):143, 2016.

[8] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*, 2016.

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision (ICCV)*, 2017.

[10] C. Zimmermann and T. Brox. Learning to Estimate 3D Hand Pose from Single RGB Images. In *International Conference on Computer Vision (ICCV)*, 2017.