GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB

Wenjie Niu

June 16,2018

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 training our CNN we propose a novel approach for the synthetic generation of training data that is based on 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. For training this translation network we combine an adversarial loss and a cycle-consistency loss with a geometric consistency loss in order to preserve geometric properties (such as hand pose) during translation. We demonstrate that our hand tracking system outperforms the current state-of-the-art on challenging RGB-only footage.[5]

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 (VR/AR) [3],[6] and human computer interaction [2],[4]. 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],[10],[8] or multi-view setups [9],[1]. However, in many applications these requirements are unfavorable since such hardware is 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 RGB-only markerless hand tracking come with clear limitations. For example, the approach by Simon et al. achieves the estimation of 3D joint locations within a multiview setup; however in the monocular setting only 2D joint locations are estimated.

In Fig. 1 we demonstrate that our method is also compatible with community or vintage RGB video. In particular, we show 3D hand tracking in YouTube videos, which demonstrates the generalization of our method.

References

[1] L. Ballan, A. Taneja, L. V. Gool, and M. Pollefeys. Motion capture of hands in action using discriminative salient points. In *European Conference on Computer Vision*, 2012. 1

[2] A. M. Feit, A. M. Feit, C. Theobalt, and A. Oulasvirta. Investigating the dexterity of multi-finger input for mid-air text entry. In *ACM Conference on Human Factors in Computing Systems*, 2015. 1

[3] T. Lee and T. Hllerer. Multithreaded hybrid feature tracking for markerless augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 15:355–368, 2008. 1

[4] A. Markussen, M. R. Jakobsen, Hornb, and K. k. Vulture: A mid-air word-gesture keyboard. In *ACM Conference on Human Factors in Computing Systems*, 2014. 1

[5] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[6] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn. User-defined gestures for augmented reality. In *Human-Computer Interaction*, 2013. 1

[7] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, and Y. Wei. Accurate, robust, and flexible real-time hand tracking. In *ACM Conference on Human Factors in Computing Systems*, 2015. 1

[8] 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*, 2016. 1

[9] S. Sridhar, H. Rhodin, H. P. Seidel, A. Oulasvirta, and C. Theobalt. Real-time hand tracking using a sum of anisotropic gaussians model. In *International Conference on 3D Vision*, 2014. 1

[10] J. Taylor, L. Bordeaux, T. Cashman, B. Corish, C. Keskin, T. Sharp, E. Soto, D. Sweeney, A. Topalian, and A. Topalian.

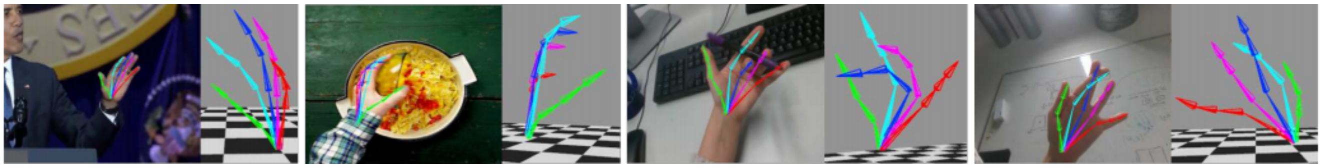

Figure 1. We present an approach for real-time 3D hand tracking from monocular RGB-only input. Our method is compatible with unconstrained video input such as community videos from YouTube (left), and robust to occlusions (center-left). We show real-time 3D hand tracking results using an off-the-shelf RGB webcam in unconstrained setups (center-right, right).

Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics*, 35:143, 2016. 1