Learning from Synthetic Humans

Cheng Guan

June 10, 2018

1 Introduction

Convolutional Neural Networks provide signficant gains to problems with large amounts of training data. In the field of human analysis, recent datasets [1, 5] now gather a sufficient number of annotated images to train networks for 2D human pose estimation [4]. Other tasks such as accurate estimation of human motion, depth and body-part segmentation are lagging behind as manual supervision for such problems at large scale is prohibitively expensive. Images of people have rich variation in poses, clothing, hair styles, body shapes, occlusions, viewpoints, motion blur and other factors. Many of these variations, however, can be synthesized using existing 3D motion capture (MoCap) data and modern tools for realistic rendering. Provided suf?cient realism, such an approach would be highly useful for many tasks as it can generate rich ground truth in terms of depth, motion, body-part segmentation and occlusions.

Although synthetic data has been used for many years, realism has been limited. In this work we present SURREAL: a new large-scale dataset with syntheticallygenerated but realistic images of people. Images are rendered from 3D sequences of MoCap data. To ensure realism, the synthetic bodies are created using the SMPL body model , whose parameters are fit by the MoSh method given raw 3D MoCap marker data. We randomly sample a large variety of viewpoints, clothing and lighting. SURREAL contains more than 6 million frames together with ground truth pose, depth maps, and segmentation masks. We show that CNNs trained on synthetic data allow for accurate human depth estimation and human part segmentation in real RGB images, see Figure 1. Here, we demonstrate that our dataset, while being synthetic, reaches the level of realism necessary to support training for multiple complex tasks. This opens up opportunities for training deep networks using graphics techniques available now. SURREAL dataset is publicly available together with the code to generate synthetic data and to train models for body part segmentation and depth estimation .

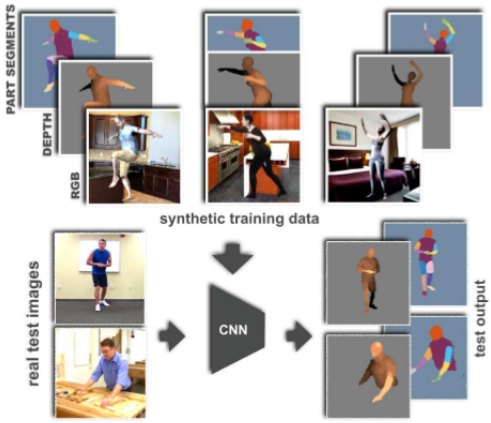

Figure 1: We generate photo-realistic synthetic images and their corresponding ground truth for learning pixel-wise classification problems: human part segmentation and depth estimation. The convolutional neural network trained only on synthetic data generalizes to real images sufficiently for both tasks. Real test images in this figure are taken from MPII Human Pose dataset

2 Related Work

Knowledge transfer from synthetic to real images has been recently studied with deep neural networks. Dosovitskiy *et al* [3]. learn a CNN for optical flow estimation using synthetically generated images of rendered 3D moving chairs. Peng *et al*. [6] study the effect of different visual cues such as object/background texture and color when rendering synthetic 3D objects for object detection task. Similarly, explores rendering 3D objects to perform viewpoint estimation. Fanello render synthetic infrared images of hands and faces to predict depth and parts. Recently, Gaidon have released the Virtual KITTI dataset with synthetically generated videos of cars to study multi-object tracking.

Several works focused on creating synthetic images of human bodies for learning 2D pose estimation , 3D pose estimation , pedestrian detection , and action recognition . Pishchulin generate synthetic images with a game engine. Some professors deform 2D images with a 3D model. More recently, Rogez and Schmid use an image-based synthesis engine to augment existing real images. Ghezelghieh render synthetic images with 10 simple body models with an emphasis on upright people; however, the main challenge using existing MoCap data for training is to generalize to poses that are not upright.

The closest work to this paper is [2], where the authors render large-scale synthetic images for predicting 3D pose with CNNs. Our dataset differs from [2] by having a richer, per-pixel ground truth, thus allowing to train for pixel-wise predictions and multi-task scenarios. In addition, we argue that the realism in our synthetic images is better , thus resulting in a smaller gap between features learned from

synthetic and real images. The method in [2] heavily relies on real images as input in their training with domain adaptation. This is not the case for our synthetic training. Moreover, we render video sequences which can be used for temporal modeling.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 1

[2] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3D pose estimation. In *3DV*, 2016. 1, 2

[3] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 1

[4] R. Okada and S. Soatto. Relevant feature selection for human pose estimation and localization in cluttered images. In *ECCV*, 2008. 1

[5] G. L. Oliveira, A. Valada, C. Bollen, W. Burgard, and T. Brox. Deep learning for human part discovery in images. In *ICRA*, 2016. 1

[6] X. Peng, B. Sun, K. Ali, and K. Saenko. Learning deep object detectors from 3D models. In *ICCV*, 2015. 1