3D Pose Estimation from Color Images without Manual Annotations

Mahdi Rad¹, Markus Oberweger¹, and Vincent Lepetit²,¹

I. PROBLEM STATEMENT AND MOTIVATION

3D pose estimation is an important problem with many potential applications. However, 3D acquiring annotations for color images is a difficult task. To create training data, the annotating is usually done with the help of markers or a robotic system, which in both cases is very cumbersome, expensive, or sometimes even impossible, especially from color images. Another option is to use synthetic images for training. However, synthetic images do not resemble real images exactly. To bridge this domain gap, Generative Adversarial Networks or transfer learning techniques can be used but, they require some annotated real images to learn the domain transfer. To overcome these problems, we proposed a novel approach in [3]. Section II gives a short summary of our approach that uses synthetic data only, and Section III shows some results. For more details, we refer to [3].

II. OVERVIEW OF THE APPROACH

We propose a novel method that learns to predict a 3D pose from color images, without requiring labeled color images. Instead, it exploits labeled depth images. Depth maps are easier to generate than realistic color images, as illumination has no interference on depth, and as we will show, the gap between color images and depth maps, and the gap between real and synthetic depth maps can be dealt with easily. Fig. 1 shows an overview. 

Training: Our method is split into two steps, each one solving an easier problem than the original one. First, we use an RGB-D camera to capture pairs of color images and corresponding depth maps. Capturing such a set can be done by simply moving the camera around. We apply [4] to this set and learn to map the features from the color images to corresponding depth images. However, this mapping alone is not sufficient: A domain gap between the depth images captured by the RGB-D camera and the available labeled depth images remains, since the labeled depth images could be captured with another RGB-D camera or rendered synthetically. Fortunately, this remaining gap is easier to bridge than the domain gap between real and synthetic color images, since illumination and texture effects are not present in depth images. To handle it, we use Maximum Mean Discrepancy (MMD) [1] to measure and minimize the distance between the means of the features of the real and synthetic depth images mapped into a Reproducing Kernel Hilbert Space (RKHS).

III. DISCUSSION AND ILLUSTRATIVE RESULTS

Our approach is general, and can be applied to many applications, such as 3D hand pose estimation, human pose estimation, etc. Fig. 2 shows applications to 3D rigid object pose estimation and 3D hand pose estimation from color images, on the LINEMOD [2] and STB [5] datasets, respectively. Our method achieves performances comparable to state-of-the-art methods on popular benchmark datasets, without requiring any annotations for the color images.

ACKNOWLEDGMENT

This work was supported by the Christian Doppler Laboratory for Semantic 3D Computer Vision, funded in part by Qualcomm Inc.

REFERENCES