1 Introduction

A keypoint-based shape and pose representation is attractive because of its simplicity and ease of handling. In this paper, we consider that only one image per object is available in both during training and inference. We also assume that only minimalistic supervision in the form of 2D keypoints and objects’ categories are available. Proposed method is an end-to-end multi class deep non-rigid structure-from-motion (NRSfM) that only takes an image as input and outputs canonical 3D shape and camera pose.

2 Related Work

C3DPO[1] learns the factorization of the object deformation and viewpoint change. Transversal property through a separate canonicalization network. Procrustean regression is used to determine unique motions and shapes[2]. Also an end-to-end method using a CNN that can output 3D location of human keypoints for objects of multiple categories, in an end-to-end manner by only 2D keypoints annotations for supervision. Results produce the 2D locations and the semantics of the objects. Human pose estimation is also tackled in[3], where the authors propose a cyclic-loss and discriminator. Recently [4] extended Procrustean formulation with autoencoders and proposed a method that can infer 3D shapes. Most methods accept 2D keypoints as input rather than images and tackle the problem of obtaining 3D keypoint locations from a single image using a separate keypoint detector, such as a stacked hourglass network [5].

3 Method

- The method is in the framework of Non-Rigid Structure from Motion (NRSfM). Specifically, the 3D keypoint locations are represented as linear combination of a dictionary S. The matrix EEG is formed from the basis vectors where the coefficients for the basis vectors are encoded by α. The camera rotation matrix R is estimated together with S and α. τ represents the projection matrix.

\[ \min_{\alpha, R} \sum \mathcal{L}(Y, \tau R S) \]

- The proposed method takes an image as input and estimates the 2D keypoints as an intermediate step as well as producing a context vector to guide the 3D estimation.

- The method is end-to-end but operationally can be examined in 2 parts: 2D-3D (lifter) network and the 2D estimation network. For lifter, we propose cut-off formulation produces huge improvements and the overall end-to-end method surpasses the SOTA.

5 Conclusion

We study estimating 3D pose and shape from a single image for objects of multiple categories, in an end-to-end manner by only using 2D keypoint annotations for supervision. Results show that end-to-end training and the use of contextual information improve the performance substantially. Our method is the first of its kind, providing a framework that can be applied to any dataset. We outperform all the compared methods on three datasets.

References