Viewpoints and Keypoints

Shubham Tulsiani and Jitendra Malik
University of California, Berkeley - Berkeley, CA 94720
{shubhtuls,malik}@eecs.berkeley.edu

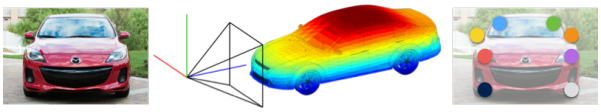

Figure 1: Alternate characterizations of pose in terms of viewpoint and keypoint locations

There are two ways in which one can describe the pose of the car in Figure 1 - either via its viewpoint or via specifying the locations of a fixed set of keypoints. The former characterization provides a global perspective about the object whereas the latter provides a more local one. In this work, we aim to reliably predict both these characterizations of pose for objects.

Our overall approach is motivated by the theory of global precedence - that humans perceive the global structure before the fine level local details [3]. It was also noted by Koenderink and van Doorn [2] that viewpoint determines appearance and several works have shown that larger wholes improve the discrimination performance of parts [4]. Focusing on the rigid object categories, we operationalize these observation by proposing an algorithm which first estimates viewpoint for the target object and leverages the predicted viewpoint to improve the local appearance based keypoint predictions.

Our proposed algorithm, as illustrated in Figure 2 has the following components -

Viewpoint Prediction : We formulate the problem of viewpoint prediction as predicting three euler angles (azimuth, elevation and cyclorotation) corresponding to the instance. We train a CNN based architecture which can implicitly capture and aggregate local evidences for predicting the euler angles to obtain a viewpoint estimate.

Local Appearance based Keypoint Activation : We propose a fully convolutional CNN based architecture to model local part appearance. We capture the appearance at multiple scales and combine the CNN responses across scales to obtain a resulting heatmap which corresponds to a spatial log-likelihood distribution for for each keypoint.

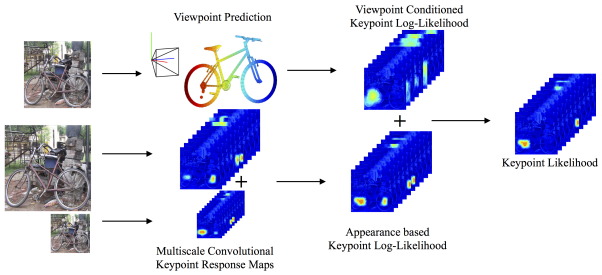

Figure 2: Overview of our approach. To recover an estimate of the global pose, we use a CNN based architecture to predict viewpoint. For each keypoint, a spatial likelihood map is obtained via combining multiscale convolutional response maps and combined with a likelihood conditioned on predicted viewpoint to obtain our final predictions.

Viewpoint Conditioned Keypoint Likelihood : We propose a viewpoint conditioned keypoint likelihood, implemented as a non-parametric mixture of gaussians, to model the probability distribution of keypoints given the viewpoint prediction. We combine it with the appearance based likelihood computed above to obtain our keypoint predictions.

Furthermore, inspired by the analysis of the detection methods presented by Hoeim *et al*. [1], we present an analysis of our algorithm's failure modes as well as the impact of object characteristics on the algorithm's performance.

References

[1] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *Computer Vision–ECCV 2012*, pages 340–353. Springer Berlin Heidelberg, 2012. 1

[2] J. J. Koenderink and A. J. van Doorn. The internal representation of solid shape with respect to vision. *Biological cybernetics*, 32(4):211–216, 1979. 1

[3] D. Navon. Forest before trees: The precedence of global features in visual perception. 1977. 1

[4] S. E. Palmer and N. M. Bucher. Configural effects in perceived pointing of ambiguous triangles. *Journal of Experimental Psychology: Human Perception and Performance*, 7(1):88, 1981. 1