A Coarse-to-Fine Model for 3D Pose Estimation and Sub-category Recognition

Roozbeh Mottaghi[1], Yu Xiang[2,3], Silvio Savarese[3]
[1]Allen Institute for AI. [2]University of Michigan-Ann Arbor. [3]Stanford University.

Despite the fact that object detection, 3D pose estimation, and sub-category recognition are highly correlated tasks, they are usually addressed independently from each other because of the huge space of parameters. To jointly model all of these tasks, we propose a coarse-to-fine hierarchical representation, where each level of the hierarchy represents objects at a different level of granularity. The hierarchical representation prevents performance loss, which is often caused by the increase in the number of parameters (as we consider more tasks to model), and the joint modeling enables resolving ambiguities that exist in independent modeling of these tasks.

Our coarse to fine hierarchical model is shown in Figure 1. The coarsest level of the hierarchy reasons about the basic-level categories (e.g., *cars* vs. other categories) and provides a rough discrete estimate for the viewpoint. As we go down the hierarchy, the level of granularity changes, and more details are added to the model. For instance, for *car* recognition, at one level we reason about sub-categories such as *SUV*, *sedan*, *truck*, etc., while at a finer level we distinguish different types of *SUVs* from each other. Also, we have a more detailed viewpoint representation (continuous viewpoint) in the layers below.

There are advantages of this coarse-to-fine hierarchical representation. First, tasks at different levels of granularity can benefit from each other. For instance, if there is ambiguity about the viewpoint of the object, knowing the sub-category might help resolving the ambiguity or reduce the uncertainty in viewpoint estimation. Second, different types of features are required for different tasks that we consider. For instance, a feature that is most discriminative for distinguishing *cars* from other categories is not necessarily useful for distinguishing different types of *SUVs*. The hierarchical representation provides a principled framework to learn feature weights for different tasks jointly. Finally, we can better leverage the structure of the parameters so the performance does not drop as we increase the complexity of the model (or equivalently, the layers of the hierarchy).

Our hierarchical model is a hybrid random field as it contains discrete (e.g., sub-category) and continuous (e.g., continuous viewpoint) random variables. We employ a particle-based method to handle the mixture of continuous and discrete variables in the model. During learning, the parameters of the model in all layers of the hierarchy are estimated jointly. Inference is also a joint estimation of the object location, and its continuous viewpoint, sub-category and finer-sub-category.

Our results show that our hierarchical model is effective in joint estimation of object location, 3D pose and (finer-)sub-category information. Also, the performance typically does not drop significantly or even improves as we increase the complexity of the model. Moreover, the hierarchical model provides significant improvement over a flat model that uses the same set of features.

Dataset. For our experiments, we use PASCAL3D+ [1] dataset, which provides continuous viewpoint annotations for 12 rigid categories in PASCAL VOC 2012. We augment three categories (*aeroplane*, *boat*, *car*) of PASCAL3D+ with sub-category and finer-sub-category annotations. We consider 12, 12, and 60 finer-sub-categories for *aeroplane*, *boat*, and *car* categories, respectively. We group finer-sub-categories into 4, 4, and 8 sub-categories, respectively. For instance, the sub-categories we consider for *cars* are *sedan, SUV, truck, race*, etc., and the finer-sub-categories represent different types of *sedans* or *SUVs*. For each finer-sub-category, we have a corresponding 3D CAD model, and for annotation we assign the instance in the image to the most similar CAD model.

Results. In the paper, we provide results and comparisons for the following cases:

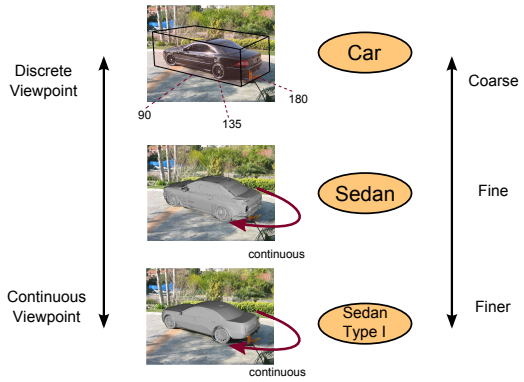

Figure 1: A coarse-to-fine hierarchical representation of an object. The top-layer captures high-level information such as a discrete viewpoint and a rough object location, while the layers below represent the object more accurately using continuous viewpoint, sub-category, and finer-sub-category information.

- We provide results for estimation of viewpoint, sub-categories and finer-sub-categories.
- We compare our 3-layer hierarchy with a flat model that uses the same set of features and we show that the hierarchy provides a significant improvement.
- We compare the results of our continuous viewpoint estimation with the results of the discrete version of our model and we show more accurate segmentation is obtained by the continuous version.
- In addition to the azimuth estimation results, we evaluate how well the method performs on elevation and distance estimation.

Some qualitative results are shown in Figure 2.

Figure 2: The result of object detection, 3D pose estimation, and (finer-)sub-category recognition. We show the projection of the 3D CAD model corresponding to the estimated finer-sub-categories according to the estimated continuous viewpoint. The magenta text is the estimated sub-category. Note that the 3D CAD model might not be the exact model for objects in PASCAL images.

[1] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.