Is 2D Information Enough For Viewpoint Estimation?

Amir Ghodrati
amir.ghodrati@esat.kuleuven.be
Marco Pedersoli
marco.pedersoli@esat.kuleuven.be
Tinne Tuytelaars
tinne.tuytelaars@esat.kuleuven.be

KU Leuven, ESAT - PSI, iMinds
Leuven, Belgium

Context. Estimating the pose of objects is a classical problem in vision. It aims at predicting a discrete or continuous viewpoint. Recent top performing methods for viewpoint estimation use 3D information. These 3D annotations are expensive and not really available for many classes.

What does this paper demonstrate. We show that a very simple 2D architecture (in the sense that it does not make any assumption or reasoning about the 3D information of the object) generally used for object classification, if properly adapted to the specific task, can provide top performance also for pose estimation. More specifically, we demonstrate how a 1-vs-all classification framework based on a Fisher Vector (FV) [1] pyramid or convolutional neural network (CNN) based features [2] can be used for pose estimation. In addition, suppressing neighboring viewpoints during training seems key to get good results.

The pipeline. Our method takes as input a detection bounding box, extracts features and assigns to the bounding box a pose. The estimation of the pose is done with a one-vs-all classifier of a discrete set of viewpoints.

- Detection: we use the deformable part models (DPM). We train our viewpoint estimation on the detected objects.
- Feature Extraction: we extract dense SIFT descriptors from the output of the detector. They are enriched by augmenting the location of the patch centre with respect to the upper-left corner of the bounding box, normalized by its size.
- Pose representation: We compare two representations commonly used in visual classification: Fisher Vector [1] + spatial pyramid matching and convolutional neural network based features [2].
- Learning: we consider each viewpoint as a different class. In this scenario an important difference with a standard 1-vs-all multi-class problem is that nearby viewpoints are generally visually very correlated. In the experimental results we show that eliminating nearby poses from negative samples always improves the viewpoint estimation. We call this procedure neighboring viewpoint suppression or briefly `nv-suppression`.

Experimental Evaluation. We evaluated our method on four datasets: Annotated faces-in-the-wild (AFW), EPFL multi-view car dataset, PASCAL3D+ and 3DObject dataset.

In table 1, we evaluate the performance of different features and encodings. we clearly notice that Bag-of-Words (`BoW`) representation is the poorest method for pose representation. The best representation on both datasets is `fisher` with spatial pyramid `spm`. Also embedding spatial information in the low-level (`sift+loc`) is still advantageous. Finally, CNN-based features, `decaf`, performs quite good as well, especially considering their much lower dimensionality.

		EPFL (8 poses)	**AFW (13 poses)**
Feature Type	Encoding	MPPE	FVP±15
`sift`	`BoW`	54.8%	49.4%
`sift`	`fisher`	68.2%	54.3%
`sift`	`fisher+spm`	80.1%	69.7%
`sift+loc`	`fisher+spm`	**81.8%**	**70.3%**
`decaf`	-	72.0%	67.9%

Table 1: An evaluation with training and testing data from output of detector on the EPFL car dataset and AFW faces dataset. MPPE is computed as the average of the diagonal of the confusion matrix. FVP±15 is the fraction of faces that are within ±15 degrees error interval, counting missed detections as infinite error.

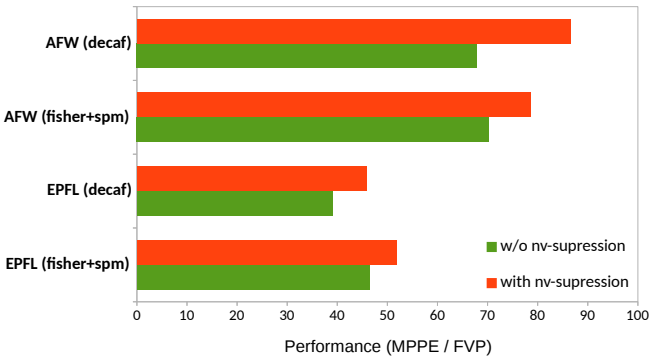

Figure 1: The effect of `nv-supression` using 36 poses for EPFL and 13 poses for AFW dataset.

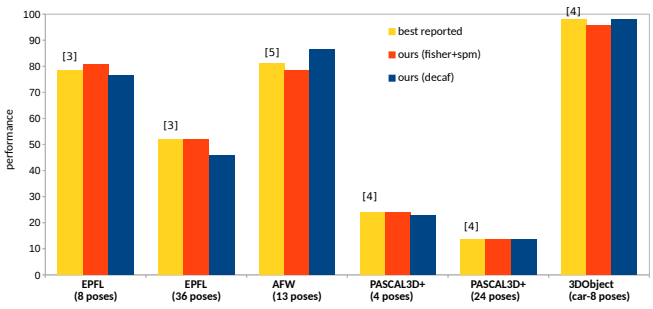

Figure 2: Viewpoint estimation in terms of MPPE, FVP(±15), mean AVP (Average Viewpoint Precision) and MPPE for EPFL, AFW, PASCAL3D+ and 3DObjects datasets respectively.

Figure 1 shows the effect of the neighboring viewpoints supression (`nv-suppression`). Its advantage is quite evident for the finer binning pose estimation for both types of features.

Comparison with state-of-the-art. Figure 2 shows the results of our methods and the current state-of-the-art on four datasets.

Conclusion. Through an extensive evaluation we can clearly see that for the fine-grained task of pose estimation, in contrast to common believe, the very simple framework based on the extraction of modern features (`decaf`) or in combination with modern encodings (`fisher+spm`) can in most of the cases get similar results as the 3D methods previously proposed and designed specifically for the problem of pose estimation.

References

[1] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *NIPS*, 1999.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[3] B. Pepik, P. Gehler, M. Stark, and B. Schiele. 3d2pm–3d deformable part models. In *ECCV*, 2012.

[4] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In *CVPR*, 2012.

[5] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.