# Reasoning about Object Relations for Object Pose Classification

José Oramas M.
KU Leuven, ESAT-PSI, iMinds

Luc De Raedt
KU Leuven, CS-DTAI

Tinne Tuytelaars
KU Leuven, ESAT-PSI, iMinds

## 1. Introduction

Object pose or viewpoint classification is an important problem for a wide range of applications. This problem has been traditionally approached from a local perspective which exploits intrinsic features of the objects such as shape or color [8, 9, 12, 15]. In this paper we follow the line of [2, 7, 14] which exploits relations between objects. However, different from these works, in addition to predicting the occurrence of an object instance, we also predict its pose. Moreover, we reason in a 3D representation of the scene assuming we know the ground plane, not in the 2D image space. In addition, instead of using symbolic spatial relations (e.g. *in-front-of, close, near, far*) we use continuous measures to define relations between objects as in [1, 13]. Finally, different from existing work, we explore the use of relations defined in an *object-centered* Frame of Reference. We formulate pose classification as a Within-Network classification problem which consists on making a prediction about an object based on the neighboring objects.

## 2. Proposed Method

In order to measure the level to which an object fits in a group of objects, first, we need to define *relations* between objects. Here, we limit ourselves to purely pairwise relations. We define these relations in an *object-centered* perspective by changing the location and orientation of the frame of reference (FoR). First an object $o_i$ is selected and the frame of reference is centered on it with the Z-axis facing in the frontal direction of the object. Then, we measure the relative location and pose of each of the other objects $o_j$, one at a time, producing a relational descriptor $r_{ij} = (rx_{ij}, ry_{ij}, rz_{ij}, r\theta_{ij})$. In practice we ignore $ry_{ij}$ since all the objects we consider are found on the ground plane so $ry_{ij} = 0$ in all cases.

**Allocentric Pose Classification:** with *allocentric pose classification*, we refer to classifying the pose $\theta_i$ of an object $o_i$ purely based on the objects in its neighborhood $N_i$. In our experiments, $N_i$ is the set containing all the other objects $o_j$ in the scene. This pose is estimated as $\theta_i^* = \arg\max_{\theta_i}(pRN(o_i|N_i))$, where $\theta_i$ belongs to the discrete set of possible poses and $pRN(o_i|N_i)$ is a probabilistic Relational Neighbor classifier (pRN) as introduced in [10]. pRN is a simple method that can take advantage of the underlying structure between elements in a network. It has been successfully used, on text datasets, for social network analysis, etc. This classifier operates in a node-centric fashion meaning that it processes one object $o_i$ at a time based on a set of $m$ objects $o_j$ in its neighborhood $N_i$. It is defined as $pRN(o_i|N_i) = \frac{1}{Z}\sum_{j\in N_i} p(o_i|o_j)p(\hat{o_j})$.

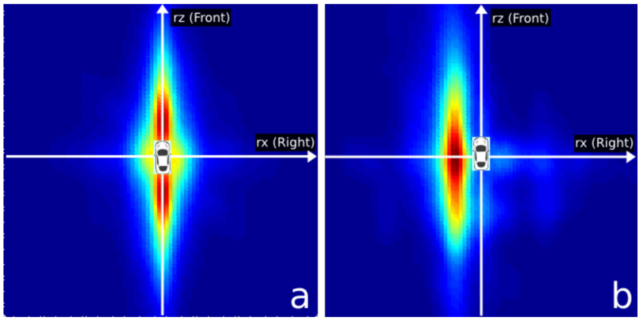

Figure 1. Distribution of object-centered relations for cars with the same pose (a) and opposite pose (b) respectively .

This classifier is composed by three terms: $p(o_i|o_j)$, which expresses the influence of the neighboring object $o_j$ on the object $o_i$; the term $p(\hat{o_j})$ which measures the confidence on the neighbor $o_j$; and the normalization term $Z$. In our setting, we define the influence term $p(o_i|o_j)$ as $p(o_i|r_{ij})$. Using Bayes' rule we estimate $p(o_i|r_{ij})$ as the posterior:

$$p(o_i|r_{ij}) = \frac{p(r_{ij}|o_i)p(o_i)}{p(r_{ij}|o_i)p(o_i) + p(r_{ij}|\neg o_i)p(\neg o_i)} \quad (1)$$

To obtain the components of Eq.1, first, we run the local detector on a validation set producing a set of hypotheses per image. Then we label the hypotheses as true positives (TP) or false positives (FP) based on the Pascal VOC matching criterion [3]. We define pairwise relations $r_{ij}$ between the hypotheses reported for each image. Relations are divided in two groups. One group contains relations in which both participants are TP hypotheses and the second group contains relations in which at least one participant is a FP hypothesis. Finally, the relations on these groups are used via Kernel Density Estimation (KDE) to estimate $p(r_{ij}|o_i)$ and $p(r_{ij}|\neg o_i)$ respectively. This method captures the statistics of typical configurations. For instance, when

| **8 Poses** | | Real | | |
|---|---|---|---|---|
| ideal (RC) | chance (RC) | LC [9] | RC | LC+RC |
| 0.47 | 0.13 | 0.27 | 0.20 | **0.30** |
| **16 Poses** | | Real | | |
| ideal (RC) | chance (RC) | LC [5] | RC | LC+RC |
| 0.37 | 0.06 | 0.55 | 0.27 | **0.57** |

Table 1. Mean Pose Classification Performance for the Ideal and Real Scenarios (MPPE values per method). LC (Local Classifier), for their respective baselines. RC (Relational Classifier).

applied on top of *OC* relations, it effectively encodes that cars with the same pose tend to be one behind the other - as when driving in the same lane, while cars with opposite poses are more likely to be driving on the left - as in opposite lanes (see figure 1). The priors $p(o_i)$ and $p(\neg o_i)$ of the object occurring or not at the given location, are estimated as the percentage of TP hypotheses and FP hypotheses in the validation set, respectively. We combine the response of the Local and Relational classifiers following the method proposed in [13]. For more details please refer to [11].

## 3. Evaluation

We run experiments on the KITTI dataset [4] focusing on the car class. To evaluate pose classification we show the Mean Average Precision in Pose Estimation (MPPE) as presented in [6, 8, 9, 12, 15]. MPPE is computed as the average of the diagonal of the class-normalized confusion matrix of the pose classifier. We evaluate the performance on the classification of 8 and 16 poses.

Our experiments aim to answer the question: “What is the effect of considering object relations for the task of object pose classification?”. The first experiment considers the ideal scenario where the local object detector and pose estimator are $100\%$ accurate for the objects in the neighborhood. The pose of each object is then predicted based on the ground truth locations and poses from objects in its neighborhood. This will show the upper limit of the performance that the Relational Classifier (RC) used for allocentric pose classification can achieve. The second experiment uses, as local classifiers (LC), the pose-aware object detectors from [9] and [5] to obtain the initial object hypotheses. These local classifiers predict 8 and 16 poses respectively. The objective of this experiment is to evaluate the performance of both, local and relational, classifiers alone and the combination of the two for the task of object pose classification in realistic settings.

**Discussion:** Table 1 shows that, in an ideal scenario, the allocentric pose classifier takes advantage of finer discretization of object poses. While the absolute number is lower for the 16 poses classifier, with twice as many output labels this is a significantly harder problem. This experiment shows the upper limits in performance that can be expected from allocentric pose classification using local detectors [5, 9]. Based only on context information, it is not possible to accurately classify the object's pose. At the same time, this upper bound is similar or even higher than what current state-of-the-art local detectors can obtain, and therefore using context information to improve pose classification seems promising. Our experiments also shows that it is possible, in a real scenario to classify, at least to some extent, the pose of objects by looking at the poses and locations of other objects – even if these poses and locations are noisy themselves. While the performance of the relational classifier alone is lower than the one obtained by the local classifier, it is significantly above the chance levels. Moreover, the combination of both local and relational classifier brings a mean improvement, over the local classifier, of $2.5\%$ and $1.7\%$, on [9] and [5] respectively.

**Acknoledgments:** This work is supported by the ERC grant 240530 COGNIMUND, the VASI Project and the BOF Research Fund KUL 3E100864.

## References

[1] R. G. Cinbis and S. Sclaroff. Contextual object detection using set-based classification. In *ECCV*, 2012. 1

[2] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 2011. 1

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 Results. 1

[4] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2

[5] A. Geiger, C. Wojek, and R. Urtasun. Joint 3d estimation of objects and scene layout. In *NIPS*, 2011. 2

[6] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and pose estimation. In *CVPR*, 2011. 2

[7] C. Li, D. Parikh, and T. Chen. Automatic discovery of groups of objects for scene understanding. In *CVPR*, 2012. 1

[8] J. Liebelt and C. Schmid. Multi-view object class detection with a 3d geometric model. In *CVPR*, 2010. 1, 2

[9] R. J. Lopez-Sastre, T. Tuytelaars, and S. Savarese. Deformable part models revisited: A performance evaluation for object category pose estimation. In *ICCV WS*, 2011. 1, 2

[10] S. A. Macskassy and F. J. Provost. Classification in networked data: A toolkit and a univariate case study. *JMLR*, 2007. 1

[11] J. Oramas M, L. De Raedt, and T. Tuytelaars. Allocentric pose estimation. In *ICCV*, 2013. 2

[12] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In *CVPR*, 2012. 1, 2

[13] R. Perko and A. Leonardis. A framework for visual-context-aware object detection in still images. *CVIU*, 2010. 1, 2

[14] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 1

[15] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *ICCV*, 2007. 1, 2