Allocentric Pose Estimation

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

Luc De Raedt
KU Leuven, CS-DTAI

Tinne Tuytelaars
KU Leuven, ESAT-PSI, iMinds

Abstract

The task of object pose estimation has been a challenge since the early days of computer vision. To estimate the pose (or viewpoint) of an object, people have mostly looked at object intrinsic features, such as shape or appearance. Surprisingly, informative features provided by other, external elements in the scene, have so far mostly been ignored. At the same time, contextual cues have been shown to be of great benefit for related tasks such as object detection or action recognition. In this paper, we explore how information from other objects in the scene can be exploited for pose estimation. In particular, we look at object configurations. We show that, starting from noisy object detections and pose estimates, exploiting the estimated pose and location of other objects in the scene can help to estimate the objects' poses more accurately. We explore both a camera-centered as well as an object-centered representation for relations. Experiments on the challenging KITTI dataset show that object configurations can indeed be used as a complementary cue to appearance-based pose estimation. In addition, object-centered relational representations can also assist object detection.

1. Introduction

Object pose or viewpoint estimation is an important problem for a wide range of applications, including robotics and road safety systems. Various methods for tackling this problem have been proposed [16, 20, 21, 27, 31], yet it is still far from being solved. Especially in 'real-world' scenarios, like the one depicted in the KITTI dataset [10], with lots of clutter, occlusions, etc. results are still relatively poor. Context information has been used successfully for object detection [5, 14, 25] in various forms (stuff, things and scene related cues). This has been effective in clarifying ambiguous scenarios. Yet, to the best of our knowledge, context information has not yet been exploited for pose estimation.

Imagine you are given the task of predicting the pose of the objects below the yellow circles in Fig.1. Even when there is no access to intrinsic features of the objects, the overall configuration of surrounding objects provides a strong cue to predict their pose. This can be considered a Collective Classification problem [32] in which the class (pose) of one object influences that of another. We face two challenges towards solving this problem. First, we need a method to define informative relations between objects. These relations should be robust to viewpoint changes and general enough to be applicable to different classes of objects (i.e. not using class-specific features). Second, a method to discover and reason about configurations of objects should be adopted. In this paper, we explore how information from other objects in the scene can be exploited for the task of pose estimation. In particular, we look at configurations of "Things". We show that, even when starting from a noisy pose estimator, results can be improved by looking at configurations. Considering the first challenge, robust and informative relations, we explore both a camera-centered and an object-centered representation for relations. Related to the second challenge, we use a simple, yet powerful, method to reason about the configuration of objects. We capture statistics of typical objects configurations using kernel density estimation, and combine this information using collective classification, more specifically a Relational Neighbor classifier [23].

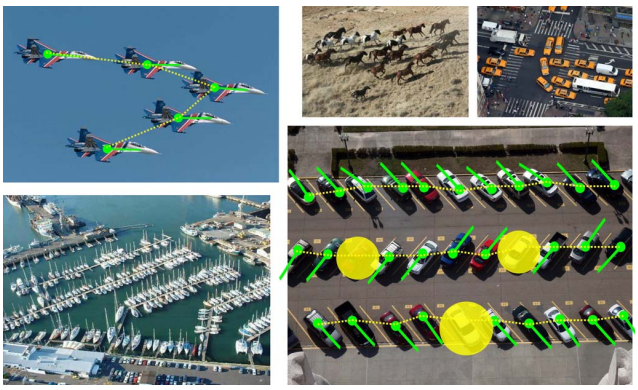

Figure 1. The natural or "desired" configurations in which objects occur in the world often provide strong cues of their pose. For instance, it is not difficult to guess the pose of the cars below the yellow circles by only looking at the rest.

The main contributions of our work are: First, we show that considering configurations between objects can be ben-

eficial for pose estimation: the proposed collective classification method complements state-of-the-art local pose estimation methods. Second, we show the influence of the Frame of Reference (FoR) – i.e. object-centered or camera-centered, used to define relations between objects for object pose estimation and detection. To our knowledge this is the first attempt to exploit relations defined between object entities via collective classification for the task of pose estimation. Additionally, we show our scheme can also improve object detection results. The paper is organized as follows: section 2 presents related work. The following three sections show how we define and learn relations between objects in the scene, and how we combine them with the evidence from local detectors. In section 6 we provide implementation details, while section 7 describes the experimental results. Finally, we draw conclusions in section 8.

2. Related Work

Some object classes can be easily recognized based on their material, color and texture, while others are characterized predominantly by their shape or appearance. Based on this particularity, Forsyth et al. [8] introduced the division of object classes into "Things" and "Stuff". In this paper we focus on object classes with defined shape and appearance, the "Things", and methods exploiting relations and configurations between them to predict their pose.

Several pose estimation methods have been proposed in the literature. All of these rely on intrinsic characteristics of the object class. In the traditional processing pipeline for pose estimation, first, candidate regions to host object instances are proposed. Secondly, an appearance descriptor is computed in the area of each candidate region. Finally, based on a pre-trained model, each descriptor is classified as one of the possible poses the object may take. Following this pipeline, methods have evolved from modeling 2D views of the classes of interest (e.g. [21]) to reasoning about object parts in the 3D space [16, 20, 27, 31].

Recently, methods related to structure from motion such as [1, 2, 12, 37] aim at understanding the full scene layout. They assume that correspondences between scene elements such as points, regions and objects across image views or sequences introduce constraints in the scene behind the images. These correspondences are exploited and among the different tasks these methods target, they also perform 3D pose estimation. These methods have shown impressive qualitative results. Yet they rely on the availability of image sequences or stereo pairs. Similar to these works we define relations between scene elements. However, instead of defining relations between different scene element types such as points, regions or objects, we focus on relations between object instances. Additionally, we drop the requirement of multiple images for the extraction of evidence - we only assume the ground plane to be known.

In recent years, learning relations between "Things" has gained popularity in the computer vision community, particularly to assist the task of object detection. Early work [5, 6, 28, 33, 36] represented objects as regions in the image. Then, by learning qualitative 2D spatial relations (e.g. top-left, far-left) between them, hypotheses in unlikely areas were filtered out. Extending this idea, [9, 18, 24] went beyond object categories and also take the appearance of the objects into account. More recently, [19, 30] use discriminant relations between objects to learn the collective appearance of related objects in order to guide the detection of the individual objects. Similar to these works, we learn relations between object instances. 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 asuming we know the ground plane, not in the 2D image space. Additionally, instead of using symbolic spatial relations (e.g. *in-front-of, close, near, far*) we use continuous measures to define relations between entities as in [4, 28, 29]. Finally, different from existing work, we explore the use of relations defined in an *object-centered* Frame of Reference. For simplicity, we focus in this paper on the relational aspect of Collective Classification, i.e. the Within-Network classification problem, leaving for future work the analysis of different methods that can be applied for collective inference. Within-Network classification consists of making a prediction about an object based on the neighboring objects.

3. Relations between Objects

We believe that the pose of an element is not only affected by its individual behavior but also by its behavior towards other elements in the scene. This idea is inspired by "Psychological Allocentrism" which states that elements tend to be interdependent, defining themselves in terms of the group they are part of, and behaving according to the norms of the group [17, 34]. Allocentric elements appear to see themselves as an extension of their in-group. Based on this description, our method takes into account the group consistency of each element relative to the group defined by the other elements in the scene.

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 two different ways, by changing the location and orientation of the frame of reference (FoR). This results in *camera-centered (CC)* and *object-centered (OC)* relations. We define *object-centered* relations between objects as follows. 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 (see Figure 2b). Then, we measure the relative location and pose of each of the other objects o_j, one at a time, producing a rela-

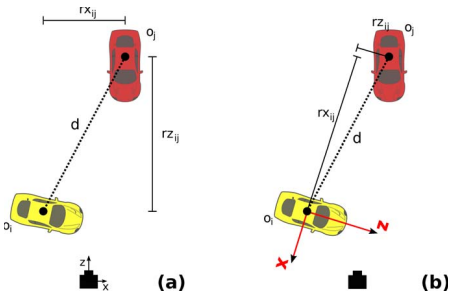

Figure 2. Spatial relations between objects. a) Camera-centered relations, b) object-centered relations. Note the difference between relative X and Z values.

tional descriptor $r_{ij} = (rx_{ij}, ry_{ij}, rz_{ij}, r\theta_{ij})$. For an image with m objects a total of $(m(m-1))$ pairwise relations are extracted. 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. As a baseline, we also perform experiments with *camera-centered* relations, as used traditionally. For these, we use the same relational descriptor as above, yet with everything measured relative to a frame of reference attached to the camera (see Figure 2a). Note that $rx_{ij}^{CC} \neq rx_{ij}^{OC}$ and $rz_{ij}^{CC} \neq rz_{ij}^{OC}$, but $r\theta_{ij}^{CC} = r\theta_{ij}^{OC}$.

4. Learning

4.1. Allocentric Pose Estimation

With *allocentric pose estimation*, we refer to the task of estimating the pose θ_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 follows:

$$\theta_i^* = \arg\max_{\theta_i}(pRN(o_i|N_i)) \tag{1}$$

where θ_i belongs to the discrete set of possible poses and $pRN(\theta_i|N_i)$ is a probabilistic Relational Neighbor classifier (pRN) as introduced in [23]. pRN is a simple method, yet with strong representative power, 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, author collaboration detection and suspicion scoring. 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 follows:

$$pRN(o_i|N_i) = \frac{1}{Z}\sum_{j\in N_i} p(o_i|o_j)p(\hat{o_j}) \tag{2}$$

This classifier is composed by three terms: $p(o_i|o_j)$, which expresses the influence of the neighboring object o_j on the unknown object o_i; the term $p(\hat{o_j})$ which measures the confidence on the neighbor o_j; and the normalization term $Z = \sum_{j\in N_i} p(\hat{o_j})$.

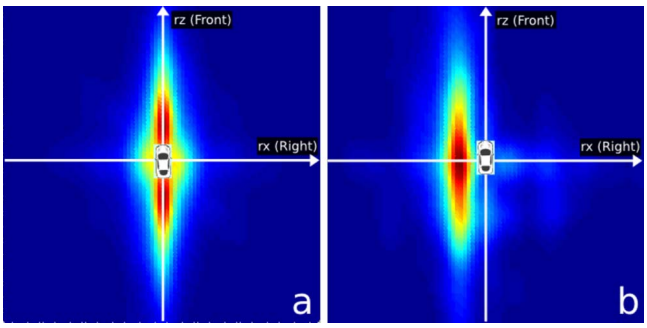

Figure 3. Distribution of object-centered relations for cars with the same pose (a) and opposite pose (b) respectively .

In a perfect scenario, where all the objects are accurately detected, the term $p(\hat{o_j}) = 1$, since we are certain of their occurrence, and the normalizer Z corresponds the number of neighboring objects. 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:

$$p(o_i|r_{ij}) = \frac{p(r_{ij}|o_i)p(o_i)}{p(r_{ij}|o_i)p(o_i) + p(r_{ij}|\neg o_i)p(\neg o_i)} \tag{3}$$

To obtain the components of Eq.3, first, we run the local detector on a validation set with annotated objects 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 [7]. We follow the procedure of Sec.3 to 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 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 3). 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.

4.2. Working with noisy detections

In practice, state-of-the-art object detectors are not perfect and produce many false hypotheses. Moreover, the location of true predictions are also noisy, while the pose is often simply wrong. For these reasons the confidence on the hypotheses predicted by the local detector should be considered during the voting procedure of Eq.2, via $p(\hat{o_j})$. We define the term $p(\hat{o_j})$ in two different ways depending on

the objective of the classification. In this paper we focus mainly on the task of object pose estimation and as a side experiment in re-ranking object detections (see Sec.7.2).

Pose Estimation: For the task of object pose estimation, we estimate $p(\hat{o_j}) \sim p(\hat{\theta_j})$ aiming to compensate for the noise in the poses used to compute r_{ij}. Since the scores given as output by state-of-the-art pose aware detectors are indications of the localization of the object rather than of its pose, we exploit the information from the confusion matrix of the pose estimator. Given a 3D object o_j with estimated continuous pose $\hat{\theta_j}$ (see Sec. 6), we estimate $p(\hat{o_j})$ by performing a linear interpolation to its nearby discrete poses θ_{low} and θ_{top} using their corresponding responses $p(\theta_{low})$ and $p(\theta_{top})$ from the diagonal of the confusion table.

$$p(\hat{\theta_j}) = p(\theta_{low}) + (p(\theta_{top}) - p(\theta_{low}))\frac{(\hat{\theta_j} - \theta_{low})}{(\theta_{top} - \theta_{low})} \quad (4)$$

Object detection: For this task, we need to put more emphasis on the occurrence of the object rather than its pose. For this reason, we estimate $p(\hat{o_j})$ through a *probabilistic local classifier* that takes into account the detection score s_j of the predicted hypothesis $\hat{o_j}$. We consider the posterior of the object occurrence given its detection score as the output of the local classifier, $p(\hat{o_j}) = p(o_j|s_j)$. We compute this posterior following the procedure of [28]:

$$p(o_j|s_j) = \frac{p(s_j|o_j)p(o_j)}{p(s_j|o_j)p(o_j) + p(s_j|\neg o_j)p(\neg o_j)} \quad (5)$$

To obtain the components of this equation we perform a procedure similar to the one done for Eq.3 up to the point were hypotheses are labeled as TPs or FPs. Then, considering the TP and FP hypotheses we compute the conditionals $p(s|o)$ and $p(s|\neg o)$ respectively based on KDE. Finally, the priors $p(o)$ and $p(\neg o)$ are estimated in the same way as in Eq.3. As a result, $p(o_j|s_j)$ will express the probability of a hypothesis being correct given its detection score. This procedure allows us to plug-in any standard object detector in our method.

5. Modeling consistency between local appearance and allocentric behavior

At this point, we have two methods to estimate the probability of a certain pose for an object hypothesis o_i: based on its intrinsic features, as evaluated by a traditional pose estimator, and based on its neighborhood N_i, respectively. The reader should note the "competitive" behavior of these two methods. While the local classifier (*lc*) pulls the decision towards individual features, the relational classifier (*rc*) (Eq.2) pulls it towards the collective feature of group fitting. Given the "competitive" nature of these classifiers, local and relational, we need to find a method to reconcile them.

To achieve this we follow a method similar to [28]. First, we collect the responses of the local (Eq.5) and relational (Eq.2) classifiers on a validation set, giving us score pairs $S = (s_{lc}, s_{rc})$ for each object hypothesis o. Then we group these score pairs for TP and FP hypotheses. At test time, we estimate $p(S|o)$ and $p(S|\neg o)$ via KDE. These terms are used in the equation $p(o|S) = p(S|o)p(o)/(\sum_{(o,\neg o)} p(S|o)p(o))$ to estimate the desired posterior.

6. Implementation Details

The focus of this paper is on the study of how relations between objects can assist the task of object pose estimation. For this reason rather than proposing our own object detector and pose estimator we use state-of-the-art detectors to acquire evidence of objects in the scene. To show the generality of our method, we build on two different detectors / pose estimators, namely those proposed in [21] and [12]. Both methods are based on the popular deformable parts model of [26], and both of them jointly tackle the problems of object detection and pose estimation. We use them as off-the-shelve detectors with default parameters. These detectors, separately, feed our framework with confidence scores, locations (2D bounding box) and poses of object hypotheses discretized into 8 and 16 partitions respectively. Then, using a stereo pair and the algorithm for efficient large-scale stereo matching proposed in [11] we obtain a 3D point cloud of the scene. To obtain the 3D location of the object, we project the point cloud into the image plane and take as location the 3D point at the bottom center of the bounding box predicted by the detector. For the 3D size of the object (used purely for visualization purposes), we use the mean width, length and height of 3D annotations in the training data. Though this is not very accurate, it is an approximation that worked well in practice. Reasoning about the relative location of objects permits the usage of alternative methods (e.g. [3], [15] and [22]) that focus on building 2.5D-3D scene representations from still images in cases where stereo pairs are not available. It should be noted that the stereo pairs are used solely to estimate the 3D location of the objects and not to derive information (e.g. 3D shape) that can be used to estimate the pose of the object. Additionally, this dependence on relative location rather than shape/volume, regions or scene class, sets our work in the middle between works based on 2.5D and works from Holistic Scene Understanding.

For the pose, the detectors provide a discrete angle α of the object as seen by the camera. From this angle we obtain a continuous azimuth angle θ in the world coordinate frame by back projecting the object o on the ground plane. To measure the certainty of this estimation during testing, we perform a linear interpolation of the estimated azimuth angle using the closest discrete pose angles and the confusion

table of the local pose estimator as discussed in Sec. 4.2. Since one of our objectives is to evaluate the influence of the frame of reference for defining informative relations, we define relations using both *CC* and *OC* FoRs. The procedure is directly applied for the case of *camera-centered relations*. For *object-centered relations* an additional step is required where the FoR should be centered in the trajector object before any relation attribute can be measured (see Section 3).

When performing Kernel Density Estimation, $f(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h})$, K is a gaussian kernel, x_i represents each of the n observations (detection score or pairwise relations) gathered from the annotated images, and h is the bandwidth value. This h value is obtained in a data-driven fashion using Silverman's Rule of Thumb [35]. For the case of Multivariate KDE, we employ kernel products.

7. Evaluation

7.1. Dataset

Most pose estimation datasets do not include groups of objects in images. Usually, there is just a single object in the main focus of the picture. This reduces the datasets in which the proposed method could be evaluated to one, the KITTI benchmark [10]. We run experiments on the object detection set of KITTI with *car* as the class of interest. We evaluate the influence of the FoR when defining relations between objects in both ideal (annotated) and real (estimated) world settings. For the ideal setting, the dataset provides 3D location and pose vectors for the objects. For the real setting, it provides stereo pairs for each scene and object annotations that allow us to build methods to learn and evaluate the configurations between object instances. Additionally, the multiple cars occurring in each image provide a challenging realistic scenario with occlusions and clutter that will be useful to evaluate our proposed allocentric pose estimator. We evaluate against all the object annotations despite their occlusion level and considered images with more than two objects. We split the training set of the KITTI dataset [10] into four subsets. The first quarter of the set is used for training the relational classifier and estimating the pose estimator confusion matrix. The second is used for validation and learning the combination of the local and the relational classifier. The third and fourth quarters are used for testing. We run experiments in 5 different splits of data.

7.2. Experiments

7.2.1 Pose Estimation

To evaluate pose estimation we show the Mean Average Precision in Pose Estimation (MPPE) as presented in [13, 20, 21, 27, 31]. MPPE is computed as the average of the diagonal of the class-normalized confusion matrix of the pose classifier. In our work MPPE is computed from hypotheses that are assumed correct based on the Pascal VOC intersection/union criterion [7], as in prior work. Both the baseline and our method start from the same initial set of hypotheses. We report results in four sets of experiments.

Method	testSet
Ideal Local Classifier (8 poses)	0.47
Ideal Local Classifier (16 poses)	0.37

Table 1. Pose Estimation Performance in the Ideal Setting (MPPE values per method).

Ideal Scenario Experiment: The first experiment aims at answering the question: "How much information about the object's pose can be obtained based on the locations and poses of objects in its neighborhood?". To this end, we consider the ideal scenario, where the local object detector and pose estimator are 100% accurate for the objects in the neighborhood. In this scenario all the objects of interest in the scene have been detected and their pose has been accurately predicted. For this experiment we use ground-truth annotations from the dataset. The pose of each object is then predicted based on the ground truth locations and poses in its neighborhood. The objective of this experiment is to present the upper limit of the performance that the Relational Classifier (RC) used for allocentric pose estimation can achieve in an ideal setting on the current dataset. We compare 2 ideal allocentric pose estimators that are able to predict 8 and 16 poses respectively.

Discussion: Table 1 shows that, in an ideal scenario, the allocentric pose estimator 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 estimation using local detectors [12, 21]. Based only on context information, it is not possible to accurately estimate 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 (see below), and therefore using context information to improve pose estimation results seems promising.

Real Scenario Experiment: This experiment starts from the local detectors [12, 21] introduced in Section 6. We define object-centered relations between the 3D hypotheses in the scene (i.e. the 2D object detection back projected onto the ground plane) and perform pose estimation based on the method proposed in Section 3. The objective of this experiment is to evaluate: a) the performance of the local pose estimators, b) the performance of pose estimation based on object relations alone, and c) the changes in performance brought by the method proposed in Sec.5 for modeling the consistency of local and relational classifiers. We report results on two sets. The first set runs on the raw output of the baseline detectors, while the second set adds a 3D Non-Maximum Suppression (3DNMS) pre-processing

testSet			testSet (3DNMS)		
LC [21]	RC	LC+RC	LC [21]	RC	LC+RC
0.27	0.20	**0.30**	0.29	0.20	**0.31**
testSet			testSet (3DNMS)		
LC [12]	RC	LC+RC	LC [12]	RC	LC+RC
0.55	0.27	**0.57**	0.57	0.24	**0.58**

Table 2. Mean Pose Estimation Performance in the Real Scenario (MPPE values per method). LC (Local Classifier), for their respective baselines. RC (Relational Classifier).

step to remove overlapping hypotheses. Given a set of 3D hypotheses o_i we suppress all the hypotheses that are closer than a threshold value t. This value is heuristically estimated from the training set, by estimating the mean width of the objects of interest. Any object closer than a factor of 0.8 is assumed to overlap and is suppressed.

Discussion: The results of this experiment (see table 2) show it is possible, also in a real scenario to estimate, 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 level (12.5% for the 8-poses [21] setting and 6.25% for the 16-poses [12] setting). Moreover, the combination of both local and relational classifier brings a mean improvement, over the local classifier, of 2.5% and 1.7% with standard deviation 0.7% and 0.6%, on [21] and [12] respectively. This indicates that the obtained improvement is indeed significant. Additionally, this shows that information encoded by our allocentric pose estimator is complementary to the local detectors and can help in scenarios where evidence of multiple object instances can be obtained. As depicted in Fig.4 the inclusion of object configurations helps to fix some of the, initially, wrongly estimated poses. This example also shows, that even for the case of FP hypotheses, our method predicts poses for objects that could have occurred in such locations.

We additionally tried a variation of this setting where pose information is ignored when defining relations between object instances. As a result, reasoning will be performed based purely on relative locations between objects. As expected, allocentric pose estimation in this setting has lower performance. In fact, its performance is close to chance level and is 15% lower than the setting where relations include pose information. Given these observations, we conclude that object pose information plays an important role when modeling configurations between object instances and that it is an intrinsic feature that must be considered in future algorithms that take into account contextual features for reasoning. This is also a strong evidence that we are dealing with a true collective classification problem as the pose of one object depends on the pose of the other

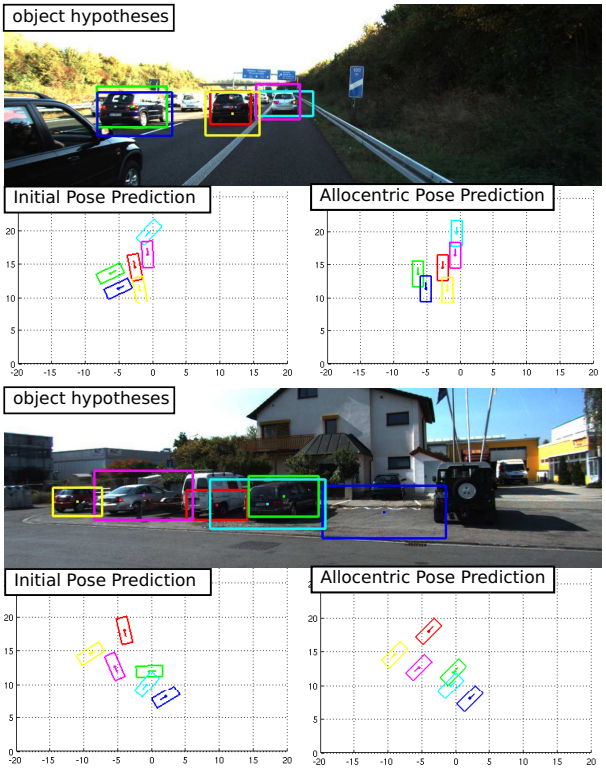

Figure 4. Effect of considering object configurations for pose estimation. Per set: Top image, hypotheses reported by the detector; bottom left, in bird's eye view, initial pose prediction given by the standard pose estimator; bottom right, in bird's eye view, pose prediction when considering object configurations.

ones. This motivates the use of pRN.

7.2.2 Object Verification

While in this paper we focus on the task of pose estimation, the configuration of objects and their poses in a neighborhood around a given object can also be exploited for object verification, i.e. to correct errors of the object detector. This is tested in the next experiment. We define *Object Verification* as the task of re-ranking the set of hypotheses given by a detector in such a way that the most likely hypotheses get a higher score. For this task we need a relational classifier that predicts the occurrence of an object o_i given the objects in its neighborhood N_i. We define this classifier similar to the pRN classifier (Eq.2) presented in Sec.4.2 where the weighting factor is assumed to be equal to $p(o_j|s_j)$ and is computed using Eq.5. The conditional $p(o_i|o_j)$ is estimated using Bayes' rule as in Eq.3 . The task of object verification is evaluated based on the criterion used in Pascal VOC [7]. We report results using Average Precision (AP) as performance metric on the testing set described before. Additionally, we report the performance of using traditional camera-centered relations and our proposed object-centered Relations. Again we show results for the two selected object de-

LC \ RC	none	CCRel.	OCRel
None	-	0.342	0.347
[21]	0.600	0.622	0.629
None	-	0.300	0.314
[12]	0.637	0.666	0.671

Table 3. Object Verification Performance (AP) related to the baseline [21] and [12]. LC (Local Classifier), RC (Relational Classifier), CCRel (Camera-Centered Relations), OCRel (Object-Centered Relations).

LC \ RC	none	CCRel.	OCRel
None	-	0.396	0.399
[21]	0.676	0.685	0.682
None	-	0.353	0.364
[12]	0.717	0.724	0.725

Table 4. Object Verification Performance (AP) related to the baseline [21] and [12] using 3DNMS. LC (Local Classifier), RC (Relational Classifier), CCRel (Camera-Centered Relations), OCRel (Object-Centered Relations).

tectors, relational classifiers based on them, and the combination of the two (Table 3). Considering the fact that we are reasoning in 3D Space, we repeat the previous object verification experiment adding a pre-processing 3DNMS step applied on the 3D hypotheses (Table 4).

Discussion: The change in performance brought by the combination of local and relational classifiers, over the local classifier alone, confirms that indeed the proposed relations assist the task of object verification. In our experiments we obtained mean improvements of 3% and 4% for [21] and [12] baselines respectively. Furthermore, it is remarkable how the relational classifiers (RC) are clearly above their respective chance levels, 24% and 14%, by 10% and 16% respectively. These chance levels correspond to the true positive - false positive ratio of the baselines [21] and [12], respectively Table 4 shows how using the "heuristic" 3DNMS step improves all the baselines by 7%. However, the improvement brought by contextual information in that case is reduced to 1% for both detectors. This can be explained by the fact that the increase in performance given by the 3DNMS makes the local classifier better, leaving less room for improvement. One might argue that our distance based 3DNMS is sub-optimal when compared with methods used in Holistic Scene Understanding for NMS based on volumetric overlap. However, our experiments presenting 3DNMS results should be considered as just a hint of additional advantages that can be obtained from reasoning in a 3D rather than a 2D space. Future work will address reasoning about the volumetric properties of objects and the effect of the re-estimated poses on the aspect ratios of the hypotheses initially predicted by the detector.

7.2.3 Object-centered or Camera-centered

To analyze the effect of the FoR when defining relations between objects, we evaluated the performance of the relational classifier with camera-centered relations and object-centered relations respectively (Sec. 3). As in the previous experiments, we present results in an ideal and realistic scenario. Furthermore, we add an experiment on the realistic scenario where we apply 3DNMS as a preprocessing step. This complements the experiments in Sec 7.2.2 involving these types of relations and will provides us an overview of their effect in such tasks.

Relations	Ideal	Real	Real (3DNMS)
[21] CCRel.	0.44	0.20	0.19
[21] OCRel.	0.47	0.20	0.20
[12] CCRel.	0.32	0.24	0.22
[12] OCRel.	0.37	0.27	0.24

Table 5. Effect of the Frame of Reference when defining relations for pose estimation (MPPE values per method). CCRel (Camera-Centered Relations), OCRel (Object-Centered Relations).

Discussion: On the KITTI dataset, the difference between the object-centered and camera-centered settings seems to be minimal for object detection (Table 3). While the object-centered setting does not depend on the camera viewpoint and therefore can be expected to generalize better to different camera setups (e.g. surveillance cameras as opposed to cameras mounted on a vehicle), the camera viewpoints in the KITTI dataset are consistent, and therefore the camera-centered setting works equally well than the object-centered one. On the pose estimation problem, previous experiments proved that pose information plays an important role when defining relations. Here object-centered relations bring an improvement of $\sim$2% over their camera-centered counterparts (Table 5).

8. Conclusions

In this paper we presented an early attempt to reason about object configurations to estimate and refine object poses. Even when, in isolation, allocentric pose estimation does not solve the object pose estimation problem, experimental results suggest that the proposed method complements local pose estimators. Furthermore, its performance, above chance levels, makes it a good alternative for cases where local information about the unknown object is unavailable (i.e. when augmenting a scene with virtual objects). Experiments also prove the relevance of pose information when describing relations between object instances; a feature that has been largely ignored in existing work that exploits contextual information, even in the context of object verification. This stresses the use of relative pose information as a feature to describe object relations. Though there is room for improvement, our results support

our hypothesis that there is something to gain from object configurations when predicting object poses. Complementing this, experiments show how defining relations from an object-centered perspective can increase performance in object pose estimation and detection. Future work will focus on two directions: first, the combination of allocentric pose estimation with more advanced local pose estimators that can reason about the 3D geometry of the object; and second, the use of more advanced relational classifiers and collective classification methods to reason about object configurations.

Acknoledgments: This work is supported by the FP7 ERC grant 240530 COGNIMUND, the VASI Project, and the DBOF Special Research Fund KUL_3E100864.

References

[1] S. Y. Bao, M. Bagra, Y.-W. Chao, and S. Savarese. Semantic structure from motion with points, regions, and objects. In *CVPR*, 2012. 2

[2] S. Y. Bao and S. Savarese. Semantic structure from motion. In *CVPR*, 2011. 2

[3] S. Y.-Z. Bao, M. Sun, and S. Savarese. Toward coherent object detection and scene layout understanding. In *CVPR*, 2010. 4

[4] R. G. Cinbis and S. Sclaroff. Contextual object detection using set-based classification. In *ECCV*, 2012. 2

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

[6] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, June 2009. 2

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 Results. 3, 5, 6

[8] D. A. Forsyth, J. Malik, M. M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson, and C. Bregler. Finding pictures of objects in large collections of images. In *ECCV*, 1996. 2

[9] C. Galleguillos, B. McFee, S. Belongie, and G. R. G. Lanckriet. Multi-class object localization by combining local contextual interactions. In *CVPR*, 2010. 2

[10] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1, 5

[11] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *ACCV*, 2010. 4

[12] A. Geiger, C. Wojek, and R. Urtasun. Joint 3d estimation of objects and scene layout. In *NIPS*, 2011. 2, 4, 5, 6, 7

[13] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and pose estimation. In *CVPR*, 2011. 5

[14] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008. 1

[15] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006. 4

[16] D. Hoiem, C. Rother, and J. M. Winn. 3d layoutcrf for multi-view object class recognition and segmentation. In *CVPR*, 2007. 1, 2

[17] L. G. Hulbert, M. L. Corrła da Silva, and G. Adegboyega. Cooperation in social dilemmas and allocentrism: a social values approach. *European Journal of Social Psychology*, 2001. 2

[18] A. Jain, A. Gupta, and L. S. Davis. Learning what and how of contextual models for scene labeling. In *ECCV*, 2010. 2

[19] C. Li, D. Parikh, and T. Chen. Automatic discovery of groups of objects for scene understanding. In *CVPR*, 2012. 2

[20] J. Liebelt and C. Schmid. Multi-view object class detection with a 3d geometric model. In *CVPR*, 2010. 1, 2, 5

[21] 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, 4, 5, 6, 7

[22] F. Lv, T. Zhao, and R. Nevatia. Camera calibration from video of a walking human. *TPAMI*, 2006. 4

[23] S. A. Macskassy and F. J. Provost. Classification in networked data: A toolkit and a univariate case study. *JMLR*, 2007. 1, 3

[24] T. Malisiewicz and A. A. Efros. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS*, 2009. 2

[25] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001. 1

[26] D. M. P. Felzenszwalb, R. Girshick. Cascade object detection with deformable part models. In *CVPR*, 2010. 4

[27] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In *CVPR*, 2012. 1, 2, 5

[28] R. Perko and A. Leonardis. A framework for visual-context-aware object detection in still images. *CVIU*, 2010. 2, 4

[29] M. Ristin, J. Gall, and L. van Gool. Local context priors for object proposal generation. In *(ACCV)*, 2012. 2

[30] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 2

[31] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *ICCV*, 2007. 1, 2, 5

[32] P. Sen, G. Namata, M. Bilgic, and L. Getoor. Collective classification. In *Encyclopedia of Machine Learning*. 2010. 1

[33] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *NIPS*, 2004. 2

[34] H. C. Triandis and E. M. Suh. Cultural influences on personality. *Annual Review of Psychology*, 2002. 2

[35] M. Wand and M. Jones. Kernel smoothing, 1995. Chapman & Hall CRC. 5

[36] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *NIPS*, 2007. 2

[37] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele. Monocular Visual Scene Understanding: Understanding Multi-Object Traffic Scenes. *TPAMI*, 2013. 2