Occlusion Patterns for Object Class Detection

Bojan Pepik[1] Michael Stark[1,2] Peter Gehler[3] Bernt Schiele[1]

[1]Max Planck Institute for Informatics, [2]Stanford University, [3]Max Planck Institute for Intelligent Systems

Abstract

Despite the success of recent object class recognition systems, the long-standing problem of partial occlusion remains a major challenge, and a principled solution is yet to be found. In this paper we leave the beaten path of methods that treat occlusion as just another source of noise – instead, we include the occluder itself into the modelling, by mining distinctive, reoccurring occlusion patterns *from annotated training data. These patterns are then used as training data for dedicated detectors of varying sophistication. In particular, we evaluate and compare models that range from standard object class detectors to hierarchical, part-based representations of occluder/occludee pairs. In an extensive evaluation we derive insights that can aid further developments in tackling the occlusion challenge.*

Figure 1. Detections on the KITTI dataset [9]. (Left) True positive detections by our occluded objects detector. Even hard occlusion cases are detected. (Right) True positives by the DPM [7].

1. Introduction

Object class recognition has made remarkable progress in recent years [6], both on the level of individual classes [4, 7] and on the level of entire visual scenes [22, 2]. Reminiscent of the early days of computer vision, 2D bounding box-based localization has been generalized to more fine-grained object class representations capable of predicting poses [1, 23], viewpoints [17], 3D parts [15], and fine-grained categories [18].

Despite these achievements towards more accurate object hypotheses, *partial occlusion* still poses a major challenge to state-of-the-art detectors [4, 7], as becomes apparent when analyzing the results of current benchmark datasets [6]. While there have been attempts to tackle the occlusion problem by integrating detection with segmentation [8] and latent variables for predicting truncation [20, 21] resulting in improved recognition performance, all these attempts have been tailored to specific kinds of detection models, and not been widely adopted by the community.

Curiously, what is also common to these approaches is that they focus entirely on the occluded object – *the occludee* – without any explicit notion of the cause of occlusion. While this approach is more general than assuming any specific type of occlusion, it also complicates the distinction between weak, but visible evidence for an object and an occluder. In this paper we therefore follow a different route, by treating the *occluder* as a first class citizen in the occlusion problem. In particular, we start from the observation that certain types of occlusions are more likely than others: consider a street scene with cars parked on either side of the road (as in Fig. 1). Clearly, the visible and occluded portions of cars tend to form patterns that repeat numerous times, providing valuable visual cues about both the presence of individual objects and the layout of the scene as a whole.

Based on this observation, we chose to explicitly model these *occlusion patterns* by leveraging fine-grained, 3D annotations of a recent data set of urban street scenes [9]. In particular, we mine reoccurring spatial arrangements of objects observed from a specific viewpoint, and model their distinctive appearance by an array of specialized detectors. To that end, we evaluate and compare two different models: i) a single-object class detector specifically trained to detect occluded objects from multiple viewpoints, occluded by various occluders, ii) a hierarchical double-object detector explicitly trained for accurate occluder/occludee bounding box localization. As baselines we include a standard, state-

of-the-art object class detector [7] as well as a recently proposed double-person detector [19] in the evaluation, with sometimes surprising results (Sect. 5).

Our paper makes the following contributions. First, we approach the challenging problem of partial occlusions in object class recognition from a different angle than most recent attempts by treating causes of occlusions as first class citizens in the model. Second, we propose three different implementations of this notion of varying complexity, ranging from easily implementable out-of-the-box solutions to powerful, hierarchical models of occluder/occludee pairs. And third, in an extensive experimental study we evaluate and compare these different techniques, providing insights that we believe to be helpful in tackling the partial occlusion challenge in a principled manner.

2. Related work

Sensitivity to partial occlusion has so far mostly been considered a lack in robustness, essentially treating occlusion as "noise rather than signal"[1]. As a result, modelling has typically focused on different ways of preventing noisy (i.e. presumably occluded) image evidence from impacting detection confidence estimates in a negative way. Among the most successful implementations are integrated models of detection and segmentation using structured prediction and branch-and-bound [8], latent occlusion variables in a max-margin framework [20], and boosting [21].

The notion that multiple visual entities that occlude each other can possibly be beneficial for recognition has mostly arisen from the perspective of context-modelling. Small objects have been demonstrated to be easier to detect in the presence of larger ones that are more reliably found [12], detected musical instruments and sports tools have been shown to enable easier human pose estimation and vice versa [24], groups of people hint on the presence of individuals [5, 23], and frequent arrangements of objects have been shown to support identification of individual objects [13].

Only recently, [19] leveraged the joint appearance of multiple people for robust people detection and tracking by training a double-person detector [7] on pairs of people rather than single humans. While our evaluation includes their model as a baseline, we systematically evaluate and contrast different ways of modelling occluders as first class citizens, and propose a more expressive, hierarchical model of occluder/occludee pairs that outperforms their model in certain configurations.

In the realm of deformable part models [10] has considered part-level occlusion in the form of dedicated "occlusion" candidate parts that represent generic occlusion features (such as a visible occlusion edge). In contrast, our models capture the specific, distinctive appearance of various occluders separately, and also leverage their distinctive spatial layout w.r.t. the occludee.

On the scene-level occlusion has been tackled with quite some success in terms of recognition performance by drawing evidence from partial object detections in probabilistic scene models [22, 14]. While these models can reason about occluder/occludee in principle, their level of detail is limited by the chosen object class representation – in both cases standard 2D bounding box-based detectors are used [7] which clearly fail to capture interactions between objects that are not box-shaped.

An entirely different avenue has been taken in the context of robotics applications, where prior distributions over expected occlusions can be analytically derived for heavily constrained, indoor scenes [11].

3. Occlusion patterns

Our approach to modelling partial occlusions is based on the notion of *occlusion patterns*, i.e., re-occurring arrangements of objects that occlude each other in specific ways and that are observed from a specific viewpoint. Note that a similar approach has been taken in the poselet framework [3], but in the context of human body pose estimation and the resulting problem of dealing with self-occlusion.

Specifically, we limit ourselves to pairs of objects, giving rise to occlusion patterns on the level of single objects (*occludees*) and double objects (*occluder-occludee* pairs).

3.1. Mining occlusion patterns

We mine occlusion patterns from training data by leveraging fine-grained annotations in the form of 3D object bounding boxes and camera projection matrices that are readily available as part of the KITTI dataset [9]. We use these annotations to define a joint feature space that represents both the relative layout of two objects taking part in an occlusion and the viewpoint from which this arrangement is observed by the camera. We then perform clustering on this joint feature space, resulting in an assignment of object pairs to clusters that we use as training data for the components of mixture models, as detailed in Sec. 4.

Feature representation. We use the following properties of occlusion patterns as features in our clustering: i) occluder left/right of occludee in image space, ii) occluder and occludee orientation in 3D object coordinates, iii) occluder is/is not itself occluded, iv) degree of occlusion of occludee.

Rule-based clustering. We found that a simple, greedy clustering scheme based on repeatedly splitting the training data according to fixed rules (e.g. based on assigning the viewing angle of the occluder to one of a fixed number

[1] J. Malik, invited talk, CVPR'12

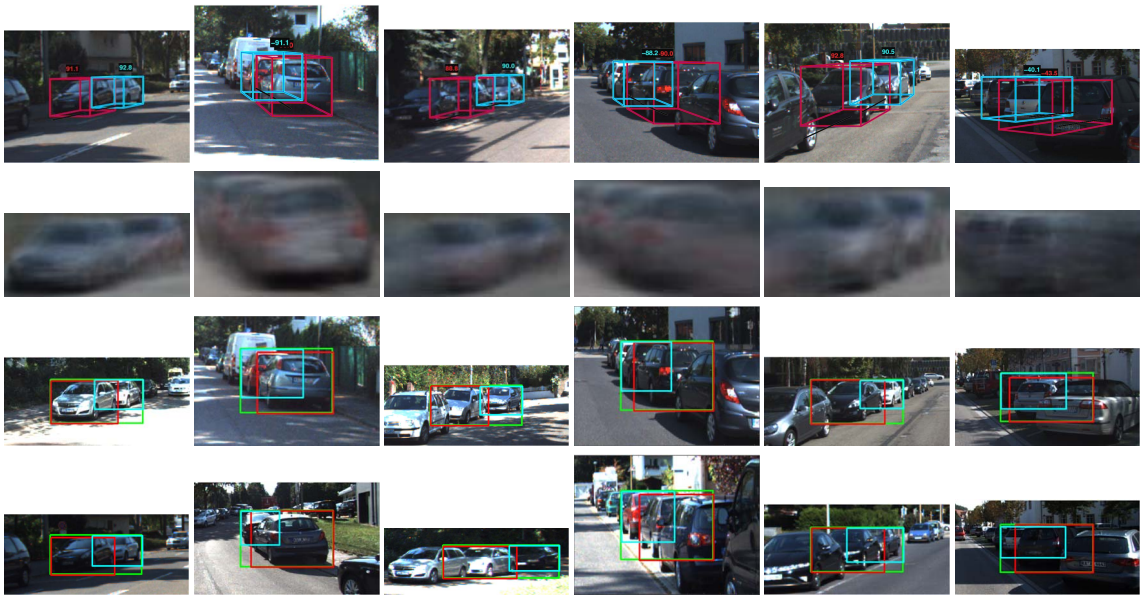


Figure 2. Visualization of mined *occlusion patterns* (occluder-occludee pairs). Top to bottom: 3D bounding box annotations provided by KITTI [9] for the cluster centroid along with the objects azimuth (row (1)), the corresponding average image over all cluster members (row (2)), two cluster members with corresponding 2D bounding boxes of occluder, occludee, and their union (rows (3) - (4)). Occlusion patterns span a wide range of occluder-occludee arrangements: resulting appearance can be well aligned (leftmost columns), or diverging (rightmost columns) – note that occluders are sometimes themselves occluded.

of predetermined bins) resulted in sufficiently clean clusters. Figure 2 visualizes a selection of occlusion patterns mined from the KITTI dataset [9]. As shown by the average images over cluster members (row (2)), some occlusion patterns are quite well aligned, which is a prerequisite for learning reliable detectors from them (Sec. 5.2).

4. Occlusion pattern detectors

In the following, we introduce three different models for the detection of occlusion patterns, each based on the well known and tested deformable part model (DPM [7]) framework. We propose two qualitatively different types of models. The first type (Section 4.2) focuses on *individual* occluded objects, by dedicating distinct mixture components to different single-object occlusion patterns. The second type (Section 4.3) models *pairs* of objects in occlusion interaction, i.e. modelling both occluder and occludee. For the second model we propose two different variants (a symmetric and an a-symmetric one).

4.1. Preliminaries

We briefly recap the basics of the DPM model as implemented in [7]. The DPM is a mixture of C star shaped log-linear conditional random fields (CRF), all of which have a root p_0 and a number of latent parts $p_i, i = 1, \ldots, M$. All parts are parameterized through their *l*eft, *r*ight, *t*op and *b*ottom extent (l, r, t, b). This defines both position and aspect ratio of the bounding box. Root and latent parts are singly connected through pairwise factors. The energy of a part configuration $p = (p_0, \ldots, p_M)$ given image evidence I for mixture component c is then

$$E_c(p; I) = \sum_{i=0}^{M} \langle v_i^c, \phi(p_i; I) \rangle + \sum_{i=1}^{M} \langle w_i^c, \phi(p_0, p_i) \rangle. \quad (1)$$

Each component has its own set of parameters (v^c, w^c) for unary and pairwise factors. The collection of those $c = 1, \ldots, C$ define the set of parameters that are learned during training. Training data is given as a set of N tuples $(I_n, y_n), n = 1, \ldots, N$ of pairs of images I and object annotations y, consisting of bounding boxes (l_n, r_n, t_n, b_n) and coarse viewpoint estimates.

4.2. Single-object occlusion patterns – OC-DPM

We experiment with the following extension of the DPM [7]. In addition to the original components $c = 1, \ldots, C_{visible}$ that represent the appearances of instances of an object class of interest, we introduce additional mixture components dedicated to representing the distinctive appearance of *occluded* instances of that class. In particular, we reserve a distinct mixture components, for each of the *occludee* members of clusters resulting from our occlusion pattern mining step (Sec. 3).

4.3. Double-object occlusion patterns

While the single-object occlusion model of Sec. 4.2 has the potential to represent distinctive occlusion patterns in

the data, modelling occluder and corresponding occludee *jointly* suggests a potential improvement: intuitively, the strong evidence of the occluder should provide strong cues as to where to look for the occludee. In the following we capture this intuition by designing two variants of a hierarchical occlusion model based on the DPM [7] framework. In these models occluder and occludee are allowed to move w.r.t. a spatial models much like parts in the DPM [7]. The two models vary in their choice of topology of the associated spatial deformations. We note that a similar route has been explored by [19], but in the context of people tracking.

4.3.1 Double-objects with joint root – Sym-DPM

The first double-object occlusion pattern detector is graphically depicted in Fig. 3 (b,e). The idea is to join two star shaped CRFs, one for the *occluding* object $\overline{p}_0$, and one for the *occluded* object $\underline{p}_0$ by an extra common root part $p_0 = (l, r, t, b)$. As training annotation for the root part we use the tightest rectangle around the union of the two objects, see the green bounding boxes in Fig. 2. The inclusion of this common root part introduces three new terms to the energy, an appearance term for the common root $\langle v^c_{joint}, \phi(p_0; I)\rangle$ and two pairwise deformation terms

$$\langle \underline{w}, \phi(\underline{p}_0, p_{joint})\rangle + \langle \overline{w}, \phi(\overline{p}_0, p_{joint})\rangle \tag{2}$$

with new parameters $\underline{w}, \overline{w}$. For these pairwise terms we use the same feature function ϕ as for all other root-latent part relations in the DPM, basically a Gaussian factor on the displacement around an anchor point.

This model retains the properties of being singly connected and thus warrants tractable exact inference. Because of the form of the pairwise term one can still use the distance transform for efficient inference. We will refer to this model as Sym-DPM. During training we have annotations for three parts $\underline{p}_0, \overline{p}_0, p_0$, while all others remain latent.

4.3.2 Double-objects without joint root – Asym-DPM

The second double-object model is a variation of Sym-DPM, where the common root part is omitted (Fig. 3 (c,f)). Instead, we directly link *occluder* and *occludee*. This relationship is asymmetric – which is why we refer to this model as Asym-DPM – and follows the intuition that the occluder can typically be trusted more (because it provides unhampered image evidence).

4.4. Training

All models that we introduced are trained using the structured SVM formulation as done for the DPM in [16]. To avoid cluttering the notation we write the problem in the following general form

$$\begin{aligned} \min_{\beta,\xi\geq 0} \quad & \frac{1}{2}\|\beta\|^2 + \frac{C}{N}\sum_{n=1}^{N}\xi_n \\ \text{sb.t.} \quad & \max_h \langle \beta, \phi(I_n, y_n, h_n)\rangle - \max_{h'} \langle \beta, \phi(I_n, y', h')\rangle \\ & \ldots \geq \Delta(y_n, y') - \xi_n, \forall y' \in \mathcal{Y}. \end{aligned} \tag{3}$$

For the models considered here, β refers to all their parameters $(v, w, \underline{w}, \overline{w})$ for all components c, y to the bounding box annotations per example (can be 1 or 2), and h to the latent part placements. For simplicity we will use y as the bounding box annotation that could comprise one or two annotations/detections. This problem is a quadratic problem and can be solved using the CCCP algorithm, alternating between fixing the latent part assignments h', and updating the parameters. The latter step involves detecting high scoring bounding boxes and latent part assignments (y', h') using loss-augmented inference $(y', h') = \operatorname{argmax}_{y,h}\langle \beta, \phi(I_n, y', h')\rangle + \Delta(y_n, y')$.

The most important change w.r.t. learning a DPM through SSVM compared to [16] is that the loss now has to take into account the possibility of multiple annotations and predictions. We use the standard intersection over union loss Δ_{VOC} for a pair of bounding boxes y, y'

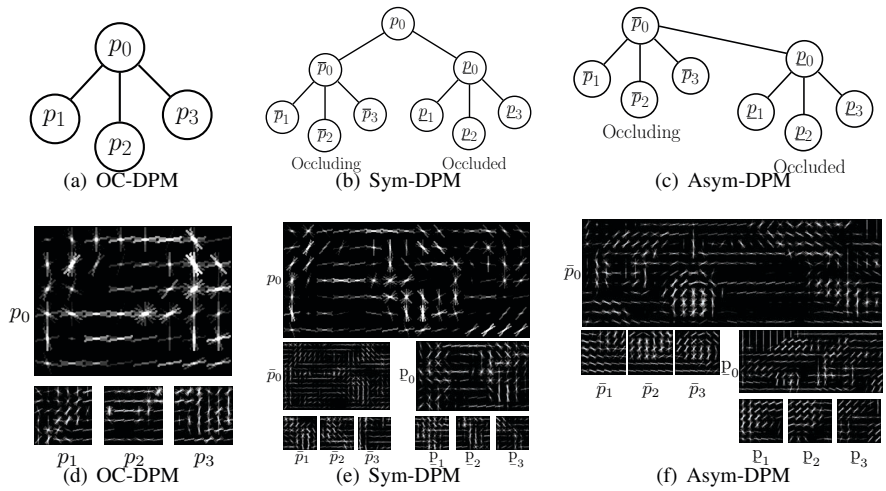
$$\Delta_{VOC}(y, y') = (1 - \frac{y \cap y'}{y \cup y'}). \tag{4}$$

and modify it in the following way. There are four different cases that have to be distinguished, 1 or 2 objects in the annotation and 1 or 2 objects that are being predicted.

In case the model predicts a single bounding box y only (decided through the choice of the component) the loss is the intersection over union loss between $\Delta(y_n, y)$ in case there is one annotation and $\Delta(\overline{y}_n, y)$ in case of an occlusion annotation. This of course is not ideal, since in case there is a second occluded object that is not being predicted, this will result in a false negative detection.

When two bounding boxes are predicted $\overline{y}, \underline{y}$ the loss is computed as either $\Delta(y_n, \overline{y})$ in case there is a single annotation or as the average $0.5\Delta(\overline{y}_n, \overline{y}) + 0.5\Delta(\underline{y}_n, \underline{y})$ between occluding and occcluded object. Again this is a proxy only, since the case of two detections but only one present in the annotation would result in a false positive.

As explained, our loss is not a perfect match since it does not penalize all false positives/negatives. We still believe it is a better proxy than the Hinge loss and found while experimenting with different implementations of Δ that the choice of Δ has only a small influence on the test time performance. This is consistent with the findings of [16] who report that the "correct" structured loss function for the DPM that takes into account the bounding box prediction rather than using the Hinge loss for classification [7]



(a) OC-DPM (b) Sym-DPM (c) Asym-DPM

(d) OC-DPM (e) Sym-DPM (f) Asym-DPM

Figure 3. Visualization of a single component of the three different occlusion models. (a) OC-DPM, (b) the Sym-DPM with root component for each object $\overline{p}_0, \underline{p}_0$ and a joint root variable p_0, (c) Asym-DPM as Sym-DPM but without a joint root variable. All models are shown with only three latent parts to avoid overloading the figure. The bottom row (d),(e),(f) show the learnt filters for the respective models. Note that for the Sym-DPM we place the joint root p_0 at half the resolution in the pyramid.

gives a consistent but rather small improvement. Our implementation of the loss function is capturing both single and double object detections simultaneously.

Detection and non-maximum suppression Test time inference in all mentioned models is tractable and efficient because they still are singly connected and allow the use of the distance transform. As usual we compute the max-marginal scores for the root components, p_0, and $\underline{p}_0, \overline{p}_0$ resp. Non-maximum suppression is done in the standard way.

5. Experimental evaluation

In the following, we give a detailed analysis of the various methods based on the notion of occlusion patterns that we introduced in Sect. 3. In a series of experiments we consider both results according to classical 2D bounding box-based localization measures, as well as a closer look at specific occlusion cases. We commence by confirming the ability of our models to detect occlusion patterns in isolation 5.2, and then move on the task of object class detection in an unconstrained setting, comprising both un-occluded and occluded objects of varying difficulty 5.3.

5.1. Data set

We chose the recently proposed KITTI data set [9] as the testbed for our evaluation, since it provides a large variety of challenging, real-world imagery of occlusion cases of different complexity, and comes with fine-grained annotations (manual 3D BBs of Lidar scans)that support a detailed analysis. It contains 7481 images of street scenes with accompanying Lidar scans, acquired from a moving vehicle.

	#objects	#occluded objects	%
Car	28521	15231	53.4
Pedest.	4445	1805	40.6
Cycles	1612	772	44.5

Table 1. KITTI dataset statistics on objects and occlusions

It is divided into 151 distinct sequences with varying duration. The sequences mostly depict inner-city scenes, but also contain rural and highway passages. In all experiments we limit ourselves to a thorough evaluation of the *Car* object class (since it occurs most often), but give additional results on the *Pedestrian* class, highlighting that our approach generalizes to non-rigid objects.

Protocol. In all experiments we perform k-fold cross-validation on the publicly available data set portion in all experiments ($k = 3$). We successively train models on two folds, evaluate them on the other fold, and afterwards aggregate the per-fold results on the level of detections.

Occlusion statistics. The KITTI dataset [9] is a rich source of challenging occlusion cases, as shown in Tab. 1. It contains thousands of objects of which almost half are occluded, e.g. 53.4% of 28521 Car objects). From Fig. 4 (a), we see that many of these are occluded to a substantial degree (the mode is around 60% occlusion). Further, Fig. 4 (b) confirms our intuition that occlusions tend to form patterns: the distribution over relative orientations of occluder-occludee pairs of cars is highly peaked around two modes.

In all our experiments on *Car* (*Pedestrian*) we train our occlusion models with 6 (6) components for visible objects and 16 (15)[2] components for occlusion patterns. We obtain these numbers after keeping the occlusion pattern clusters which have at least 30 positive training examples.

[2]The numbers vary for different folds

5.2. Detecting occlusion patterns

We commence by evaluating the ability of our models to reliably detect occlusion patterns in isolation, since this constitutes the basis for handling occlusion cases in a realistic detection setting (Sect. 5.3). In particular, we contrast the performance of our models (OC-DPM, Sym-DPM, and Asym-DPM) with two baselines, the standard deformable part model [7], unaware of occlusions, and our implementation of the recently proposed double-person detector [19], which we adapt to the *Car* setting.

Double-object occlusion patterns. We first consider the joint detection of occlusion patterns in the form of object pairs (occluder and occludee). For that purpose, we limit our evaluation to a corresponding subset of the test data, i.e. images that contain occlusion pairs, which we determine from the available fine-grained annotations (we run the occlusion pattern mining of Sect. 3 with parameters that yield a single cluster). This targeted evaluation is essential in order to separate concerns, and to draw meaningful conclusions about the role of different variants of occlusion modelling from the results. Fig. 5 (left) gives the corresponding results, comparing the performance of two variants of our Sym-DPM model (normal, in black, and a variant with object-level templates at doubled HOG resolution, red) to the double-person detector of [19] (magenta). We make the following observations: first, we observe that all detectors achieve a relatively high recall of over 90% – note that this can not be trivially achieved by lower detection thresholds, since different occlusion patterns result in largely different aspect ratios, which our models counter by dedicating a discrete set of distinct components to them. Second, we observe that our Sym-DPM performs on a comparable level to the baseline [19] (55.9% vs. 58.5% AP), and dominates in its double-resolution variant (60.6% AP).

Single-object occlusion patterns. Based on the setup of the previous experiment we turn to evaluating our occlusion pattern detectors on the level of individual objects (this comprises both occluders and occludees from the double-object occlusion patterns). To that end, we add our single-object detectors to the comparison, namely, our Asym-DPM (orange), our OC-DPM (cyan), and the deformable part model [7] baseline (green). Fig. 5 (right) gives the corresponding results. Clearly, all explicit means of modelling occlusion improve over the DPM [7] baseline (53.7% AP) by up to a striking 20.3% AP (OC-DPM, cyan, 74% AP). Equally, the recall improves drastically from approx. 70% to over 80%. As concerns the relative performance of the different occlusion models, we observe a different ordering compared to the double-object occlusion pattern case: the double-object baseline [19] (blue, 61% AP) performs slightly better than our double-resolution Sym-DPM (red, 57.9% AP), followed by our Asym-DPM (orange, 56.4% AP), and our normal Sym-DPM (black, 54.0 AP). Curiously, the arguably simplest model, our OC-DPM, outperforms all other models by at least to 13% AP.

Summary. To summarize, we conclude that detecting occlusion patterns in images is in fact feasible, achieving both sufficiently high recall (over 90% for both single- and double-object occlusion patterns) and reasonable AP (up to 74% for single-object occlusion patterns). We consider this result viable evidence that occlusion pattern detectors have the potential to aid recognition in the case of occlusion (which we examine and verify in Sect. 5.3). Furthermore, careful and explicit modelling of occluder and occludee characteristics helps for the joint detection of double-object patterns (our hierarchical Sym-DPM model outperforms the flat baseline [19]). For the single-object case, however, the simplest model OC-DPM outperforms all others by a significant margin.

5.3. Occlusion patterns for object class detection

In this section we apply our findings from the isolated evaluation of occlusion pattern detectors to the more realistic setting of unconstrained object class detection, again considering the KITTI dataset [9] as a testbed. Since the focus is again on occlusion, we consider a series of increasingly difficult scenarios for comparing performance, corresponding to increasing levels of occlusion (which we measure based on 3D annotations and the given camera parameters). Specifically, we consider the following six scenarios: the full, unconstrained data set (Fig. 8 (a)), the data set restricted to at most 20% occluded objects (Fig. 8 (b)), restricted to objects occluded between 20 and 40% (Fig. 8 (c)), between 40 and 60% (Fig. 8 (d)), between 60 and 80% (Fig. 8 (e)), and beyond 80% (Fig. 8 (f)).

Modeling unoccluded objects. In order to enable detection of occluded as well as unoccluded object instances, we augment our various occlusion pattern detectors by additional mixture components for unoccluded objects.

Results - full dataset. On the full data set (Fig. 8 (a)) we observe that the trends from the isolated evaluation of occlusion patterns (Sect. 5.2) transfer to the more realistic object class detection setting: while the double-object occlusion pattern detectors are comparable in terms of AP (Asym-DPM, orange, 52.3%; Sym-DPM, blue, 53.7%), our OC-DPM achieves the best performance (64.4%), improving over the next best double-object occlusion pattern detector Sym-DPM by a significant margin of 10.7%.

Surprisingly, the DPM [7] baseline (green, 62.8% AP) beats all double-object occlusion pattern detectors, but is in turn outperformed by our OC-DPM (cyan, 64.4%). While the latter improvement seems modest at first glance, we

point out that this corresponds to obtaining 1000 more true positive detections, which is approximately the number of cars (1250) in the entire Pascal VOC 2007 trainval set.

In comparison to [19] (53.9%), Sym-DPM and Asym-DPM provide similar performance. All double-object detectors have proven to be very sensitive to the non-maxima supression scheme used and suffer from score incomparability among the double and single object components. We intend to address this issue in future work.

On the *Pedestrian* class (Fig. 8 (g)) OC-DPM (37.2%) outperform the DPM (36.2%), confirming the benefit of our occlusion modelling, while Sym-DPM (31.4%) outperforms the Asym-DPM (29.4%).

Results - occlusion. We proceed by examining the results for increasing levels of occlusion (Fig. 8 (b-f)), making the following observations. First, we observe that the relative ordering among double-object and single-object occlusion pattern detectors is stable across occlusion levels: our OC-DPM (cyan) outperforms all double-object occlusion pattern detectors, namely, Sym-DPM (blue) and Asym-DPM (orange).Second, the DPM [7] baseline (green) excels at low levels of occlusion (77.2% AP for up to 20% occlusion, 37% AP for 20 to 40% occlusion), performing better than the double-object occlusion pattern detectors for all occlusion levels. But third, the DPM [7] is outperformed by our OC-DPM for all occlusion levels above 40% by significant margins (12.9%, 21.5%, and 4.4% AP, respectively).

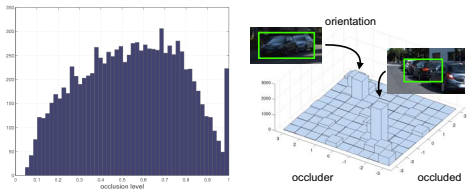
The same trend can be observed for the *Pedestrian* class: for occlusions between 60 and 80% OC-DPM (5.7%) outperforms DPM (5.0%) (Fig. 8 (h)). Asym-DPM (3.0%) outperforms the Sym-DPM (2.7%).

Summary. We conclude that occlusion pattern detectors can in fact aid detection in presence of occlusion, and the benefit increases with increasing occlusion level. While, to our surprise, we found that double-object occlusion pattern detectors were not competitive with [7], our simpler, single-object occlusion pattern detector (OC-DPM) improved performance for occlusion by a significant margin.

5.4. Discussion

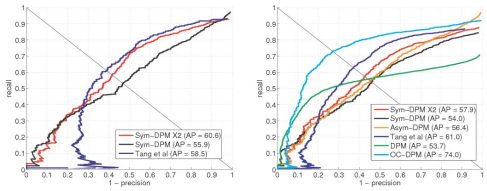
In the course of our evaluation, we have gained a number of insights which we discuss in the following.

Biased occlusion statistics. From our experience, the poor performance of double-object occlusion detectors on the KITTI dataset [9] (Sect. 5.3), which is in contrast to [19]'s findings for people detection, can be explained by the distribution over occlusion patterns: it seems biased towards extremely challenging "occluded occluder" cases. We found a large fraction of examples in which double-objects appear in arrangements of a larger number of objects (e.g. row of cars parked on the side of the road), where the occluder is itself occluded – these cases are not correctly represented by occluder-occludee models. In these



(a) Occlusion histogram (b) Orientation histogram

Figure 4. Occlusion (a), orientation (b) histogram



(a) Double object detection (b) Single object detection

Figure 5. (a) Joint , (b) single *Car* detection results

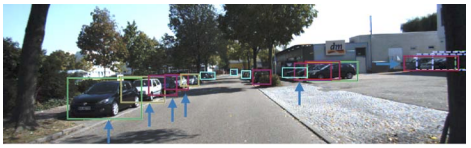


Figure 6. Examples of non tight BB annotations

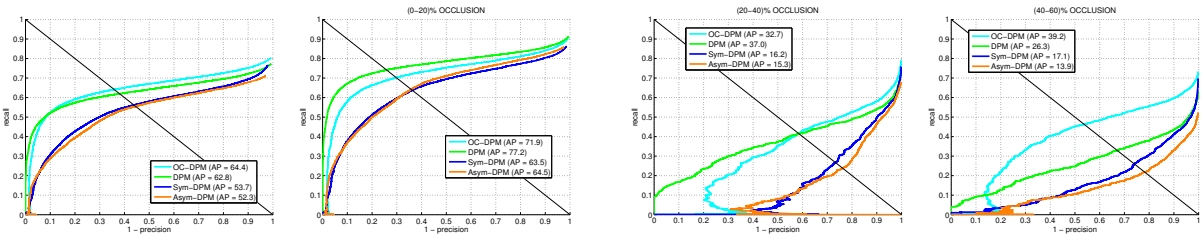
cases it proves less robust to combine possibly conflicting pairwise detections (Asym-DPM, Sym-DPM) into a consistent interpretation than aggregating single-object occlusion patterns (OC-DPM). As a result, single-object models ([7], OC-DPM) tend to be more robust against this bias, resulting in improved performance.

Annotation noise. We also found that the KITTI dataset [9] contains a significant number of occluded objects that are not annotated, supposingly due to being in the Lidar shadow, and hence missing 3D ground truth evidence for annotation. While there is a reserved "don't care" region label for these cases, this seldomly overlaps sufficiently with the object bounding box in question. This is particularly true for our best performing OC-DPM model, for which the first ≈ 70 false positive detections are of that nature, resulting in a severe under-estimation of its performance in Sect. 5.3 (Fig.7 shows examples).

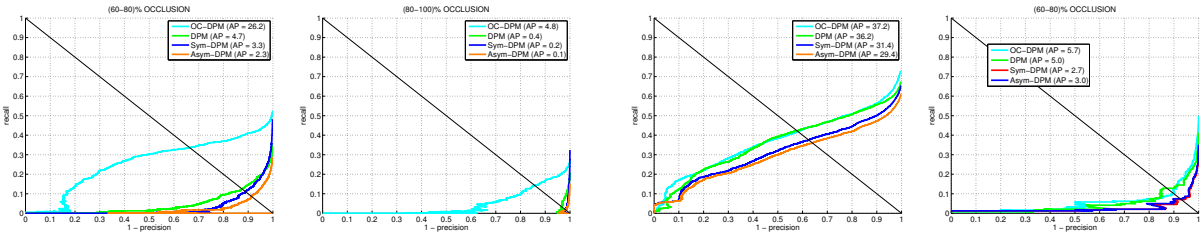
Overlap criterion. In line with the previous argument we believe the overlap threshold of 70% intersection-over-union [6] proposed by the KITTI dataset [9] is hardly compatible with the accuracy of the annotations in many cases (Fig. 6 gives examples), which is why we report results for the less challenging but more robust overlap of 50%.

6. Conclusions

We have considered the long-standing problem of partial occlusion by making *occluders* first class citizens in modelling. In particular, we have proposed two different mod-



(a) *Car*: Full dataset (b) *Car*: Occlusion level 1 (c) *Car*: Occlusion level 2 (d) *Car*: Occlusion level 3



(e) *Car*: Occlusion level 4 (f) *Car*: Occlusion level 5 (g) *Pedestrian*: Full dataset (h) *Pedestrian*: Occlusion level 4

Figure 8. Detection performance for class *Car* on (a) the full dataset, (b)-(f) increasing occlusion levels from $[0-20]\%$ to $[80-100]\%$. Detection performance on class *Pedestrian*, (g) full set, (h) $[60-80]\%$ occlusion.

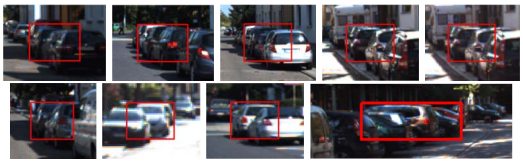


Figure 7. Valid detections on unannotated objects

els for detecting distinctive, reoccurring occlusion patterns, mined from annotated training data. Using these detectors we could improve over the performance of a current, state-of-the-art object class detector over an entire dataset of challenging urban street scenes, but even more so for increasingly difficult cases in terms of occlusion. Our most important findings are: i) reoccurring occlusion patterns can be automatically mined and reliably detected, ii) they can aid object detection, and iii) occlusion is still challenging also in terms of dataset annotation.

Acknowledgements. This work has been supported by the Max Planck Center for Visual Computing and Communication.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

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

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *ECCV*, 2010.

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[7] P. F. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010.

[8] T. Gao, B. Packer, and D. Koller. A segmentation-aware object detection model with occlusion handling. In *CVPR*, 2011.

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

[10] R. Girshick, P. Felzenszwalb, and D. McAllester. Object detection with grammar models. In *NIPS*, 2011.

[11] E. Hsiao and M. Hebert. Occlusion reasoning for object detection under arbitrary viewpoint. In *CVPR*, 2012.

[12] L. Karlinsky, M. Dinerstein, D. Harari, and S. Ullman. The chains model for detecting parts by their context. In *CVPR*, 2010.

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

[14] D. Meger, C. Wojek, B. Schiele, and J. J. Little. Explicit occlusion reasoning for 3D object detection. In *BMVC*, 2011.

[15] B. Pepik, P. Gehler, M. Stark, and B. Schiele. 3D2PM - 3D deformable part models. In *ECCV*, 2012.

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

[17] S. Savarese and L. Fei-Fei. 3D generic object categorization, localization and pose estimation. In *ICCV*, 2007.

[18] M. Stark, J. Krause, B. Pepik, D. M. andJ. J. Little, B. Schiele, and D. Koller. Fine-grained categorization for 3D scene understanding. In *BMVC*, 2012.

[19] S. Tang, M. Andriluka, and B. Schiele. Detection and tracking of occluded people. In *BMVC*, 2012.

[20] A. Vedaldi and A. Zisserman. Structured output regression for detection with partial occulsion, 2009.

[21] X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *ICCV*, 2009.

[22] C. Wojek, S. Walk, S. Roth, and B. Schiele. Monocular 3D scene understanding with explicit occlusion reasoning. In *CVPR*, 2011.

[23] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *CVPR*, 2012.

[24] B. Yao and L. Fei-Fei. Grouplet: a structured image representation for recognizing human and object interactions. In *CVPR*, 2010.