Deformable Part Models Revisited: A Performance Evaluation for Object Category Pose Estimation

Roberto J. López-Sastre[1] Tinne Tuytelaars[2] Silvio Savarese[3]
[1]GRAM, Dept. Signal Theory and Communications, University of Alcalá, robertoj.lopez@uah.es
[2]ESAT - PSI, K.U. Leuven, tinne.tuytelaars@esat.kuleuven.be
[3]Dept. of Electrical and Computer Engineering, University of Michigan, silvio@eecs.umich.edu

Abstract

Deformable Part Models (DPMs) as introduced by Felzenszwalb et al. *have shown remarkably good results for category-level object detection. In this paper, we explore whether they are also well suited for the related problem of category-level object pose estimation. To this end, we extend the original DPM so as to improve its accuracy in object category pose estimation and design novel and more effective learning strategies. We benchmark the methods using various publicly available data sets. Provided that the training data is sufficiently balanced and clean, our method outperforms the state-of-the-art.*

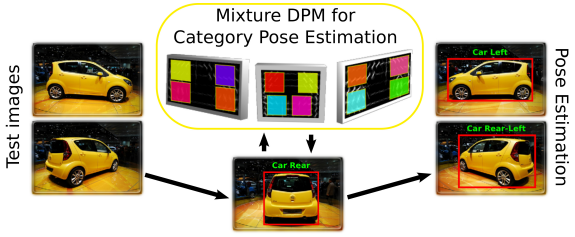


Figure 1. Our approach to the problem of object class pose estimation is based on the DPM of [5]. In our models each component represents a discrete viewpoint for an object class. Following this approach, we simultaneously perform object detection and pose estimation.

1. Introduction

Detecting object classes and estimating their poses are critical perceptual capabilities for an autonomous robot to interpret and interact with the world. For instance, for a robotic arm to grasp an object in a scene, it is crucial to accurately estimate the object pose from the camera view point, so the robot can successfully interact with it. These abilities are also crucial in autonomous navigation systems, scene layout understanding or activity recognition.

In spite of the three-dimensional nature of the world, the default approach for category-level object detection is still mostly based on independent 2D models, *i.e.* the combination of different detectors for different viewpoints (*e.g.* one detector for side views of cars, one for frontal cars, etc.). In [22] it was shown that this results in a large number of false positives, and better results can be obtained if the classifiers for different viewpoints are integrated. This leads to various multi-view object detectors, including [1, 7, 9, 12, 15, 17, 21]. Also the Deformable Part Model, or DPM for short, introduced recently by Felzenswalb *et al*. [5], combines different viewpoints, albeit in an implicit way. Instead of explicitly dividing the training data in different viewpoints at training time, this is done in a latent manner. The object category model consists of a mixture of components (as also used in [19, 23]), and training images are assigned to components without supervision. Combined with a flexible model for adding object parts, this system has demonstrated state-of-the-art results in object detection, and was the winning scheme for most recent PASCAL VOC challenges [3].

Related to the problem of category-level object detection is the problem of object category pose estimation, *i.e.* to determine the viewpoint for a given object instance. These two problems are tightly coupled and are often solved simultaneously.

In this paper, we evaluate the use of DPMs for object category pose estimation, motivated by the recent success in object detection by both DPMs [5] and discriminative template approaches [26]. To the best of our knowledge there have been no systematic explorations of the performance of DPMs in this context. Gu and Ren [7] have proposed a mixture of HOG [2] templates for the problem of viewpoint classification, but their model is not part-based, while for detection the parts have been shown to be a crucial component.

Therefore, this paper revisits the DPMs of [5], evaluating and improving their accuracy in object category pose estimation. This is achieved by: designing more effective train-

ing strategies, and introducing a *semi*-latent SVM learning methodology where the component label is not treated as latent, *i.e*. only the instantiations of the model in the feature pyramid are considered latent.

Figure 1 illustrates our approach. Our model is obtained as a mixture of components wherein each component corresponds to a different viewpoint. For a new test image, the selected component then automatically identifies the viewpoint. We have performed a thorough analysis, focusing on the following research questions: 1) How does the DPM of [5] perform in object pose estimation? 2) How can the DPM be adapted to improve the results? 3) How can differences in performance between different data sets be explained? 4) What is the ideal training strategy for a DPM in order to obtain an accurate pose estimator? 5) How does the number of components of a DPM affect the result? 6) How does the intra-view variability in training data modify the performance? and 7) Does improved pose estimation also result in better detection? We answer these questions based on a number of experiments aimed at evaluating pose estimation performance using various publicly available data sets, namely PASCAL VOC 2006, 2007 and 2010 [3], the 3D object categories of [17], the EPFL multi-view car data set [16] and the ICARO data set [13].

The rest of the paper is structured as follows. In Section 2 we describe related work on object category pose estimation. The DPM of [5] is reviewed in Section 3. In this section we also propose extensions to the DPM which allow object category pose estimation. The experimental setup and results are presented in Section 4 and 5 respectively. We conclude in Section 6.

2. Related Work

While the problem of recognizing single instances of 3D objects has been investigated for a long time in computer vision (see for example [14, 6]), only recently, researchers have started to address the problem of category-level 3D (or multi-view) object detection. Thomas *et al*. [22] incorporate shape and appearance information into a 3D object model for the class. An Implicit Shape Model (ISM) [11] is built for each viewpoint. Then, the method of [6] is used to connect regions in different views so as to transfer ISM votes from one view to its neighboring viewpoints. A similar approach using a 3D Implicit Shape Model has been explored by [1]. Kushal *et al*. [10] use a single appearance model for object parts across different viewpoints. Object classes are represented by Partial Surface Models (PSM), which are formed of dense, locally rigid assemblies of image features. Pairs of PSM which regularly co-occur near each other at consistent relative positions are linked. These local connections are then used to build a probabilistic graphical model for the geometry and appearance of the PSM. Yan *et al*. [24] propose a method that establishes spatial connections between views by mapping them directly to the surface of a 3D model, instead of using a mechanism for relating multiple 2D training views. Hoiem *et al*. [8] use a coarse 3D model of the object class to roughly match physical parts across instances at different viewpoints and to include a description of the color of the object. Liebelt and Schmid [12] present a method for including external 3D geometry from synthetic CAD models into a 2D part-based appearance detection method. Savarese and Fei-Fei [17, 18] represent an object category as a collection of view-invariant regions linked by transformations that capture the relative change of pose among parts. This model is further extended into a generative Bayesian model in [20, 21] so as to incorporate the ability to link parts across views and generate unseen views within a principled probabilistic framework. A variety of other approaches for category pose estimation have been explored as well (*e.g*. using view-specific Bayes classifiers [16], or mixed (2D and 3D) template matching [9]).

Discriminative approaches have also been proposed, mostly focusing on the problem of object localization (*e.g*. [2, 5]). The work of Gu and Ren [7] introduces a discriminative approach to the problem of object class pose estimation. Their approach is based on a mixture of HOG templates for joint category and pose estimation. However, they do not use parts in their formulation.

Part-based models have received large attention lately. The DPM model of [5], which stands out for being the winner on recent PASCAL VOC challenges, has demonstrated state-of-the-art results in object detection. Zhu *et al*. [25] incorporate into a DPM the notion of hierarchy of parts, and reformulate the training as a structural SVM learning problem. However, neither of these works directly considers the problem of object class pose estimation. In our work, we propose an extension and new learning strategy for the DPM that can explicitly cope with the problem of category pose estimation.

3. DPM for Object Category Pose Estimation

In this section, we first briefly review the DPM of Felzenszwalb *et al*. [5]. Then we extend the standard training pipeline and introduce a modified learning strategy more adequate for object class pose estimation.

In the DPM introduced in [5], an object class is modeled as a mixture over m components $(M_1, M_2, \ldots, M_m)$. The model for the c-th component M_c consists of a $2+n$-tuple $M_c = (R, P_1, P_2, \ldots, P_n, b)$, where: R represents the root filter, n is the number of parts, P_i is the model of part i, and b is a bias term. Each P_i is represented by a 3-tuple $P_i = (F_i, v_i, d_i)$, where F_i is the part filter, v_i defines the anchor position of the part i (relative to the root filter), and d_i encodes the deformation cost for the part. Root and part filters are computed on a pyramid of HOG features. An object hypothesis z specifies the mixture component as well

as the locations of both the root and part filters in the feature pyramid, *i.e.* $z = (c, p_0, \ldots, p_n)$, where c identifies the component. So, each p_i encodes the 2D position and the level in the pyramid for the filter i (with p_0 giving the position/pyramid level of the root filter). The score of a hypothesis, *i.e.* $s(z)$, is given by the scores of the filters at their locations minus a deformation cost that depends on the relative position of each part with respect to the root filter, plus the bias (we refer to [5] for further details). This score can be expressed in terms of a dot product, $s(z) = \beta_c \cdot \Phi(H, z)$, where $\beta_c = (R, F_1, \ldots, F_n, d_1, \ldots, d_n)$ is the model parameters vector for component c, and $\Phi(H, z)$ is the feature vector for a given hypothesis z in the feature pyramid H.

In our approach, for category-level pose estimation, we use a mixture model with m components. Each component c corresponds to a particular viewpoint of the class. An object hypothesis z specifies the component (*i.e.* pose) as well as the locations of the filters for the selected component. So, object detection and pose estimation are solved simultaneously.

3.1. Training DPMs for pose estimation

Felzenszwalb *et al.* [5] introduce a coordinate-descent algorithm for learning the model parameters: the Latent-SVM. The method simultaneously learns the object detector and parts detectors without part-level training annotation. The problem of part discovery is casted as a multiple instance SVM learning problem. It is possible to reformulate the DPM learning as a structural SVM learning problem [25], but here we follow the original learning paradigm proposed in [5].

The DPM in [5] has been carefully designed and optimized for object detection. Each component learned by the original implementation [4] is an asymmetric model. This bilateral asymmetry allows each component to specialize at the task of detecting left or right object orientation. This property improves the performance in object detection, but is not suitable to discriminate between *e.g.* frontal and rear poses of cars, which cannot be distinguished by their orientation. Furthermore, during training, the model ignores the pose labels given by some of the PASCAL VOC annotations. For initialization, images are assigned to components based on the aspect ratio of the bounding boxes.

In this work we propose to initialize the components by assigning the images based on the pose labels in the dataset. Moreover, we do not incorporate the bilateral asymmetry into the model. As in [5], we start training the root filters of each component independently. These root filters are combined into a mixture model with no parts, and we retrain the combined model parameters using latent detections and hard negatives, as it is described in the Latent-SVM learning. Parts are initialized using the heuristic detailed in [5], and learned following again the Latent-SVM approach. In the experiments (Section 5) the models trained following this pipeline are referred to as DPM.

3.2. Semi-latent SVM learning strategy

Toward the goal of improving pose estimation accuracy, we also experiment with an alternative, modified learning strategy. In the Latent-SVM training pipeline, a coordinate-descent approach is followed similar to expectation-maximization. It alternates between two steps: 1) relabel the positive examples, and 2) optimize the objective function over model parameters β. That is, the process starts optimizing the objective function over $Z(x_i)$, which specifies a latent value for each positive example x_i in the training set, *i.e.* the instantiations of the model in the feature pyramid. The objective of this step is to find the highest scoring object hypothesis with a root filter that significantly overlaps the positive example. In the second step, the goal is to optimize the objective function over β, *i.e.* update the model parameters. This two steps procedure needs a careful initialization.

When a mixture of models is trained during the step 1, both the component label and the instantiations of the filters are treated as latent variables. However, for pose estimation, we propose a *semi*-latent learning strategy. During the optimization, in step 1, only the instantiations of the filters are treated as latent, but not the component label, which is fixed to the pose label given in the training data. That is, for each positive example $z_i = (c, z'_i)$, the label component c is given, and $z'_i = (p_0, \ldots, p_n)$ encodes the positions of the filters in the feature pyramid. In step 1, we select the highest scoring latent value for each positive example as $z'_i = \arg\max_{z' \in Z(x_i)} \beta \cdot \Phi(x_i, z')$.

We combine this modification in the optimization process, with a new design of the training pipeline for category-level pose estimation, where the different components are merged gradually. As in the previous DPM version, we start training the root filters for each component independently. However, instead of taking negative examples from images without the objects of interest, we take as negatives the images of the opposite viewpoint to further help discriminate between poses, *e.g.* when training the frontal view, negatives are taken from the rear view. In a second step, opposite viewpoints are combined into a mixture model, and trained against negatives extracted from images that do not contain the object class of interest. For this training step and the rest, the modified optimization approach is used to retrain the model parameters. In a third step, all the components are combined into a global mixture model, and we retrain the model parameters again. Finally, parts are initialized and learned. The models trained with this modified training pipeline are referred to as mDPM in Section 5, which stands for *modified* DPM.

4. Experimental Setup

Object category pose estimation involves predicting both the bounding boxes of objects as well as their poses. To evaluate this process, we build on the PASCAL VOC Challenge guidelines [3] for evaluation of object detection using the interpolated average precision (AP). We first evaluate the bounding box with the standard 50% overlap criterion. Next we evaluate the estimated pose, which is considered correct if it coincides with the ground truth pose label. We then compute the Pose Estimation Average Precision (PEAP) similar to AP but using as positives only the correct detections *with* correct pose estimation. Additionally, we present results using confusion matrices. For these, we compute the mean of the elements on the main diagonal, a measure we refer to as Mean Precision in Pose Estimation (MPPE). The extensions to the DPM of [5] proposed in this paper are all built on top of the original software. [4].

4.1. Datasets

Traditionally, object category pose estimation systems are benchmarked using the annotated images in the PASCAL VOC 2006 [3] (only for the classes car and bicycle), and the more complete 3D object categories dataset introduced in [17]. Here, we also present results using the PASCAL VOC 2007 and 2010 datasets, the ICARO [13] set of images and the EPFL Multi-view car dataset [16].

For the experiments with PASCAL VOC 2006 and 2007, we select the `trainval` and `test` sets for training and testing respectively. However, for PASCAL VOC 2010, we use the `train` and `val` sets for training and testing, respectively. We only run experiments for the classes car and bicycle. Note that the PASCAL VOC datasets do not offer a viewpoint annotation for every object in the dataset. In a first set of experiments we only use those training images that come with pose information. Additionally, we have also manually annotated the discrete pose for all other images that come without this information (for the car and bicycle classes). We refer to these as PASCAL VOC Augmented sets[1].

We also run experiments using the 3D object categories dataset [17]. For each class, the dataset offers images of 10 different object instances with 8 different poses. In the experiments we follow the evaluation protocol described in [17]: 7 randomly selected objects are used for training, and 3 object instances for testing. 2D training bounding boxes are obtained from the provided segmentation masks. The negative examples are taken from the PASCAL VOC 2007 set. We use all the classes in the dataset, except the class monitor, because it contains too few image sequences.

[1]These augmented sets for the PASCAL VOC 2006, 2007 and 2010 can be downloaded from http://agamenon.tsc.uah.es/Personales/rlopez/data/pose-estimation/

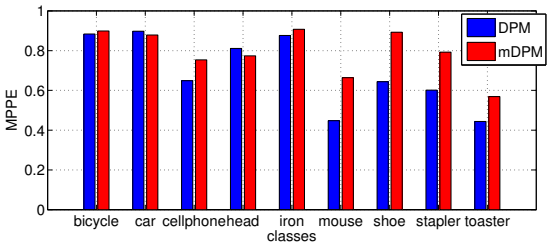


Figure 2. Mean Precision in Pose Estimation (MPPE) in the 3D object categories dataset for DPM and mDPM.

	MPPE				
Class	DPM	mDPM	[7]	[21]	[12]
bicycle	.88	.90	n/a	n/a	.75
car	.90	.89	.74	.67	.70

Table 1. A comparison between methods for classes car and bicycle in the 3D Object categories dataset.

ICARO [13] is a publicly available dataset especially designed for object category pose estimation. The dataset consists of 6419 images of 26 classes, with a total of 16221 annotated objects spread over 11 discrete viewpoints. We follow the evaluation protocol detailed in [13]. We use the `trainval` set of ICARO for training and the `test` set for testing.

Finally, we use the EPFL Multi-view car dataset, which was introduced in [16]. It contains around 2000 images, belonging to 20 different car models. All the cars appear centred in the images, without occlusions, and under 16 poses (covering the whole 360 degree range). We follow the experimental setup proposed in [16] using only the 10 first sequences for training, and the rest for testing.

5. Experiments

5.1. 3D Object categories

This dataset is widely acknowledged as a challenging benchmark for category-level pose estimation. We train and test models of 8 components, and compare the performance with state-of-the-art results reported by Gu and Ren [7], Sun *et al*. [21] and Liebelt and Schmid [12].

Figure 2 shows a comparison between DPM and mDPM. Except for classes car and head, the mDPM obtains significantly better results than the DPM in pose estimation[2]. The average MPPE for mDPM is 79.2%, while for DPM it is 69.5%, resulting in an average increase in performance of mDPM over DPM of 9.7%. Furthermore, Table 1 shows that the DPM-based methods proposed in this paper obtain significantly better results than the state-of-the-art.

[2]Note that we compare pose estimation results that have been obtained on top of different detections, *i.e*. each method estimates the pose based on its own object detections. This holds for all experiments reported. This is the way the comparisons have been done in the literature so far.

Method	AP								
	bike	car	cell	head	iron	mouse	shoe	stapler	toaster
DPM	.91	1	.62	.87	.78	.71	.88	.73	.9
mDPM	.91	.96	.43	.76	.53	.41	.78	.32	.54

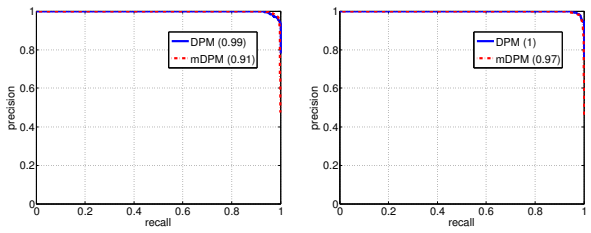
Table 2. AP in 2D object detection for DPM and mDPM in the 3D object categories dataset.

For object detection, Table 2 shows the AP obtained by DPM and mDPM. For most of the classes, the DPM version performs better than the mDPM. The modification introduced in mDPM, makes this approach more suitable for pose estimation, but less effective in object detection for this dataset. We believe this difference in AP performances stems from the different training pipelines. Recall that DPM follows the latent-SVM formulation described in [5], and this work has shown state-of-the-art results in 2D object detection.

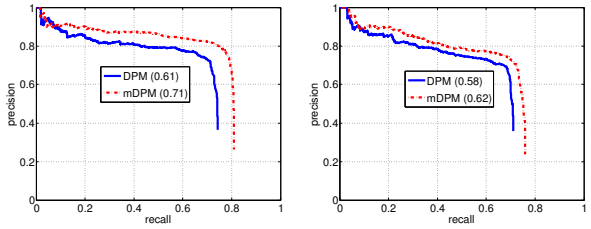
5.2. ICARO

The objective of this experiment is threefold: 1) explore how DPM and mDPM object class pose estimators perform with images where the objects appear in more realistic scenes (occluded, not centred, with clutter, etc.), 2) analyze how the number of training images affects the performance, and 3) explore how the intra-view variability in the training and testing sets influences the results. To do so, we use the ICARO dataset [13] for learning a 8 viewpoint estimator for the classes car, laptop and motorbike. In the experiments, we increase the number of training images from 50 to 250 (distributed over all viewpoints). Figure 3 shows the results of this experiment. On average, the MPPE is higher for the mDPM (0.46) than for the DPM (0.42). For quite low number of training images, reasonable results are obtained, but, as expected, they improve when more training data is available.

If we compare the results obtained for the class car in 3D object categories and in ICARO, we notice that the MPPE drastically decreases from 89% to 45%. We relate this drop of performance to several reasons. The objects in ICARO appear not centred, with occlusions and with background clutter. Moreover, the intra-view variability of the training set is higher than in the 3D object categories dataset, which makes both the detection and pose estimation problems harder. Figure 4 shows the confusion matrices for class car and method mDPM. Note that there is a relatively high confusion between adjacent views, *e.g.* frontal-left and frontal. This can be explained by the fact that the pose annotations in ICARO are not as precise as in the 3D object categories database, where the images were taken from well controlled viewpoints.



(a) 8 viewpoints (b) 16 viewpoints



(c) 8 viewpoints (d) 16 viewpoints

Figure 6. Precision/recall curves for object localization (a)-(b), and for pose estimation (c)-(d) for DPM and mDPM trained on EPFL.

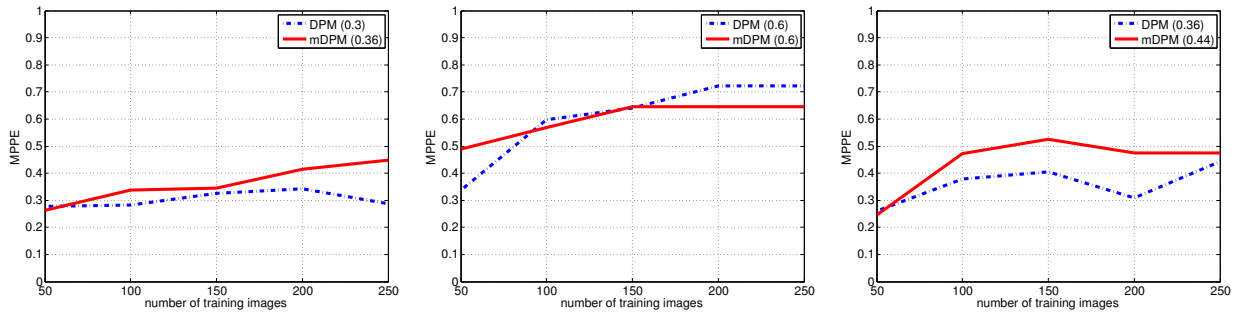
5.3. EPFL Multi-view car

We want to explore how DPMs perform in pose estimation when the number of discrete views increases. In this experiment we train 3 different DPM and mDPM models: for 16, 8 and 4 poses (collapsing the 16 viewpoints into 8 and 4). Figure 5 shows the confusion matrices. Even though the number of viewpoints has doubled, the MPPE for 16 viewpoints is not much lower than for 8 viewpoints, showing that accurate pose estimation is possible. Furthermore, our mDPM, for 16 viewpoints, obtains a MPPE of 66.1%, while the method in [16] resulted in a MPPE of 41.6%[3]. We also benchmark the 2D object detection in this dataset. Figures 6(a) and 6(b) show the precision/recall curves for object detection. DPM (AP=1) and mDPM (AP=0.97) obtain better results than the method in [16] (AP=0.85). In addition, we also benchmark the pose estimation in this dataset measuring the PEAP. Figures 6(c) and 6(d) show the precision/recall curves and the PEAP for each method. The PEAP of mDPM is higher than of DPM.

5.4. PASCAL VOC datasets

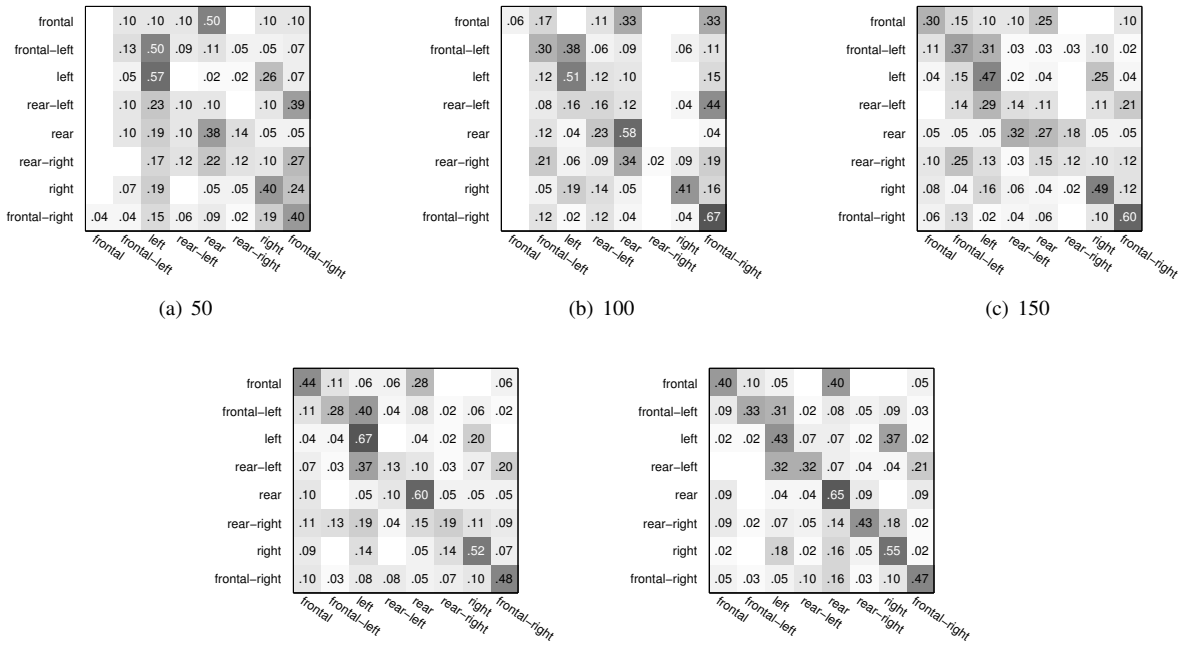
Finally, we show results on the PASCAL VOC 2006, 2007 and 2010 datasets for classes bicycle and car. Table 3 shows a comparison of DPM and mDPM in the PASCAL VOC 2006 with state-of-the-art results reported by Gu and Ren [7] and Sun *et al.* [21]. Again, mDPM gives better results than DPM. Our models only outperform the results of [21]. This is due to a large confusion between frontal and rear views for class car. In addition to the large visual similarity between these views, the low accuracy of the

[3]The authors of [16] were contacted and they sent us the actual values.



(a) car (b) laptop (c) motorbike

Figure 3. ICARO: a comparison of performance (MPPE) when the number of training images increases. The figure presents results for DPM and mDPM trained to estimate 8 poses.



(a) 50 (b) 100 (c) 150

(d) 200 (e) 250

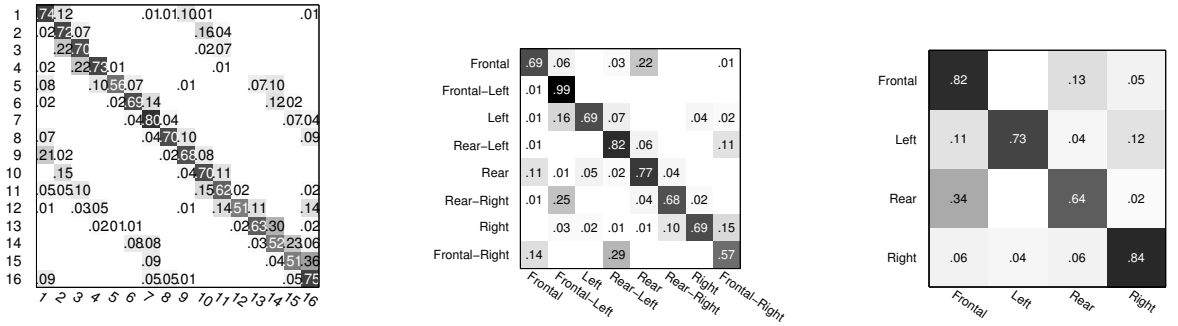
Figure 4. Confusion matrices for the class car and method mDPM trained on ICARO for different number of training images. This figure is best viewed with magnification.

models proposed can be attributed to the unbalancedness of the training set. Moreover, within the same viewpoint, the objects present changes in scale, azimuth and zenith, and some of them are occluded. All these factors have a negative impact on the performance of DPM and mDPM in pose estimation, which prefer a balanced and precise training set, such as in the case for the 3D Object categories.

We now use the PASCAL VOC Augmented sets. Figures 7(a) and 7(b) show the results obtained by mDPM in the standard and the augmented sets for classes bicycle and car respectively. While for class bike there are no differences between the results obtained with the standard and the augmented dataset, for the class car in the VOC 2006 and 2010 datasets, the MPPE increases when the augmented set is used. This may be due to the fact that these 2 augmented

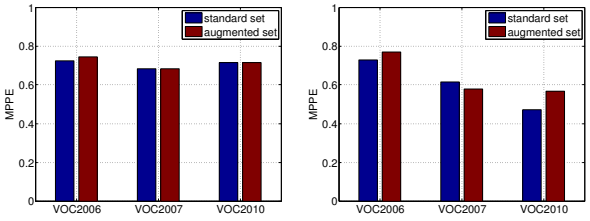
Class	MPPE DPM	mDPM	[7]	[21]
bicycle	.57	**.72**	n/a	n/a
car	.69	.73	**.86**	.57

Table 3. MPPE for classes car and bicycle in the PASCAL VOC 2006 for different methods.



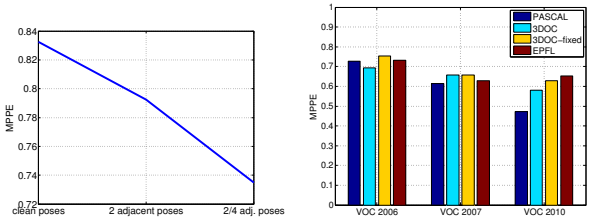
(a) 16 viewpoints, MPPE 66% (b) 8 viewpoints, MPPE 73.75% (c) 4 viewpoints, MPPE 75.75%

Figure 5. Confusion matrices for mDPM vs. the number of poses in the EPFL dataset.



(a) bicycle (b) car

Figure 7. MPPE for the mDPM using the standard and the augmented PASCAL VOC datasets.



(a) car (b) car

Figure 8. (a) MPPE vs mixing up adjacent poses for a mDPM. (b) MPPE of mDPMs trained on different datasets and tested in the PASCAL VOC.

sets are more balanced. Moreover, the higher the year of PASCAL VOC, the larger the variability on the dataset and the lower the performances of DPM and mDPM.

5.5. Discussion

Experimentally, we have detected that the performance in pose estimation of DPMs decreases for datasets where the intra-view variability is high (*e.g.* PASCAL VOC and ICARO), and where the training set is unbalanced. In order to investigate to what degree the intra-view variability affects the performance of DPM and mDPM, we set up the following experiments.

Using the EPFL dataset, we propose an experiment where the intra-view variability is controlled by mixing up adjacent views. We start training and testing a mDPM of 4 viewpoints, only using the images belonging to the 4 clean poses frontal, left, rear and right. Note that this is different from collapsing the 16 viewpoints into 4, as we did for the experiment in Figure 5. In the next step we merge 2 adjacent views (*e.g.* we label as frontal not just view number 1, but view numbers 1, 2 and 16). Finally, we merge 4 adjacent views with left and right, and 2 adjacent views with frontal and rear. Results in Figure 8(a) show that the MPPE monotonically decreases while we mix up adjacent views. Following this strategy, we can simulate the conditions we find in the PASCAL VOC and ICARO datasets.

Additionally, we test in the PASCAL VOC datasets, models that have been trained in the EPFL and the 3D Object categories datasets. With the latter, we train 2 different models. For the first one (3DOC-fixed), we only use images with a fixed zenith of 0 degrees. For the second one (3DOC), we use images with all the zenith angles provided in the 3D Object categories dataset. Figure 8(b) shows the obtained results. All the models, except the 3DOC trained with VOC 2006, obtain better results than the mDPM trained directly on the PASCAL train sets. This is remarkable, since usually best results are obtained when training on the same dataset. Especially for the VOC 2010 dataset, where the intra-view variability is very high, the mDPM trained on EPFL obtains very good results (MPPE 65.25%). So, we conclude that DPMs improve state-of-the-art results in category pose estimation if one ensures that the ground truth annotations are sufficiently precise.

As far as the 2D object detection performance of the DPM-based methods proposed is concerned, Table 4 shows the AP of the class car for each dataset. In general, a decrease of performance for mDPM is observed, although it depends on the dataset used. For instance, results for the EPFL are very significant. For an almost perfect and identical 2D detection (see Figure 6(a)) for both DPM and

	AP					
	3DOC	EPFL	ICARO	VOC2006	VOC2007	VOC2010
DPM	1	1	.33	.57	.44	.34
mDPM	.96	.97	.29	.52	.38	.13

Table 4. 2D Object detection AP for the class car.

mDPM, the PEAP increases from 61% (for DPM) to 71% (for mDPM). In conclusion, even with identical detection rates, the pose is estimated better by the mDPM approach.

6. Conclusion

We have presented a thorough performance evaluation of DPMs for category pose estimation. We also have proposed adapted training and novel learning strategies that improve the performance in viewpoint estimation of DPMs. We have extensively studied how the training data affects the performance of the proposed models. The high intra-view variability of some datasets negatively affects the performance of the pose estimators. In these cases, better results are obtained if the models are trained on external, clean data such as the 3D object categories. Provided that the training data is sufficiently balanced and clean, our training strategies significantly outperform the state-of-the-art ($+16\%$ and $+24\%$ classification accuracy for the 3D object categories dataset and the EPFL Multi-view cars dataset, respectively). In future work we aim to extend our models to the continuous viewpoint estimation problem.

Acknowledgements

This work was partially supported by projects TIN2010-20845-C03-03 and CCG10-UAH/TIC-5965, ERC grant 240359 (COGNIMUND), ARO grant W911NF-09-1-0310 and NSF CAREER #1054127. We wish to thank Carolina Redondo and Isabel Ventura for their help with the annotation of the PASCAL VOC datasets.

References

[1] M. Arie-Nachimson and R. Basri. Constructing implicit 3d shape models for pose estimation. In *ICCV*, 2009.

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

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC)challenge. *IJCV*, 88(2):303–338, 2010.

[4] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ pff/latent-release4/.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32:1627–1645, 2010.

[6] V. Ferrari, T. Tuytelaars, and L. Van Gool. Simultaneous object recognition and segmentation from single or multiple model views. *IJCV*, 67(2):159–188, 2006.

[7] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, 2010.

[8] D. Hoiem, C. Rother, and J. Winn. 3D LayoutCRF for multi-view object class recognition and segmentation. In *CVPR*, 2007.

[9] W. Hu and S.-C. Zhu. Learning a probabilistic model mixing 3d and 2d primitives for view invariant object recognition. In *CVPR*, 2010.

[10] A. Kushal, C. Schmid, and J. Ponce. Flexible object models for category-level 3d object recognition. In *CVPR*, 2007.

[11] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV*, pages 17–32, 2004.

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

[13] R. J. López-Sastre, C. Redondo-Cabrera, P. Gil-Jimenez, and S. Maldonado-Bascon. ICARO: Image Collection of Annotated Real-world Objects. http://agamenon.tsc.uah.es/Personales/rlopez/data/icaro, 2010.

[14] D. Lowe. Local feature view clustering for 3D object recognition. In *CVPR*, volume 1, pages 682–688, 2001.

[15] L. Mei, M. Sun, K. M.Carter, A. O. Hero III, and S. Savarese. Unsupervised object pose classification from short video sequences. In *BMVC*, 2009.

[16] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *CVPR*, 2009.

[17] S. Savarese and L. Fei-Fei. 3D generic object categorization, localization and pose estimation. In *ICCV*, pages 1 – 8, 2007.

[18] S. Savarese and L. Fei-Fei. View synthesis for recognizing unseen poses of object classes. In *ECCV*, pages 602–615, 2008.

[19] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *CVPR*, volume 1, pages 746–75, 2000.

[20] H. Su, M. Sun, L. Fei-Fei, and S. Savarese. Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In *ICCV*, 2009.

[21] M. Sun, H. Su, S. Savarese, and L. Fei-Fei. A multi-view probabilistic model for 3D object classes. In *CVPR*, 2009.

[22] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. Van Gool. Towards multi-view object class detection. In *CVPR*, volume 2, pages 1589–1596, 2006.

[23] M. Weber, W. Einhäuser, M. Welling, and P. Perona. Viewpoint-invariant learning and detection of human heads. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[24] P. Yan, S. Khan, and M. Shah. 3d model based object class detection in an arbitrary view. In *ICCV*, pages 1–6, 2007.

[25] L. L. Zhu, Y. Chen, A. Torralba, W. Freeman, and A. Yuille. Part and appearance sharing: Recursive compositional models for multi-view multi-object detection. In *CVPR*, 2010.

[26] L. L. Zhu, Y. Chen, A. Yuille, and W. Freeman. Latent hierarchical structural learning for object detection. In *CVPR*, 2010.