# Synthetically trained multi-view object class and viewpoint detection for advanced image retrieval

**Johannes Schels, Jörg Liebelt, Klaus Schertler, Rainer Lienhart**

**Angaben zur Veröffentlichung / Publication details:**

Schels, Johannes, Jörg Liebelt, Klaus Schertler, and Rainer Lienhart. 2011. "Synthetically trained multi-view object class and viewpoint detection for advanced image retrieval." In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval - ICMR '11, April 18 - 20, 2011, Trento, Italy*, 3. New York, USA: ACM Press. https://doi.org/10.1145/1991996.1991999.



# Synthetically trained multi-view object class and viewpoint detection for advanced image retrieval

Johannes Schels
EADS Innovation Works
Munich, Germany
johannes.schels@eads.net

Jörg Liebelt
EADS Innovation Works
Munich, Germany
joerg.liebelt@eads.net

Klaus Schertler
EADS Innovation Works
Munich, Germany
klaus.schertler@eads.net

Rainer Lienhart
Multimedia Computing Lab
University of Augsburg
Augsburg, Germany
rainer.lienhart@informatik.uni-augsburg.de

## ABSTRACT

This paper proposes a novel approach to multi-view object class and viewpoint detection for the retrieval of images showing one or several objects from a given viewpoint, a viewpoint range or any viewpoint in image databases. All detectors are trained exclusively on a few synthetic 3D models without any manual bounding-box, viewpoint or part annotation, making object class and viewpoint detection a scalable learning task. Previous work on this topic relies on the detection of object parts for each individual viewpoint, ignoring the responses of part detectors specific to other viewpoints. Instead, we explicitly exploit appearance ambiguities caused by spurious detections of parts under more than one viewpoint by combining all detector responses in a joint spatial pyramid encoding. We achieve state-of-the-art results in multi-view object class detection and viewpoint determination on current benchmarking data sets and demonstrate increased robustness to partial occlusion.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; I.3.5 [**Computational Geometry and Object Modeling**]: Curve, surface, solid, and object representations; I.4.8 [**Scene Analysis**]: Object recognition

## General Terms

Algorithms, Experimentation

## Keywords

synthetic 3D models, part-based object class detection



*ICMR* '11, April 17-20, Trento, Italy
https://doi.org/10.1145/1991996.1991999

## 1. INTRODUCTION

There are various ways of representing the query in a content-based image retrieval system. In many systems the representation is done by one or more specific sample images of the object(s) to be queried. However, such a representation is limited to the retrieval of specific object instances in images with a sufficiently high similarity to the query in terms of appearance and viewpoint. Instead, a more flexible query formulation is desired where a user can choose either to retrieve a class of objects from a specific viewpoint, a range of viewpoints, or any viewpoint, or to retrieve a specific object from a specific viewpoint, a range of viewpoints, or any viewpoint (e.g. in [9]). In order to enable such image queries, a multi-view object class detection algorithm is required that can (a) detect objects of a given object class from multiple viewpoints and (b) recognize the approximate viewpoint under which the object was recorded. Moreover, the collection of data for the training of new object class detectors should require little manual intervention. One such source for training data could be 3D CAD models of objects, since 3D models allow for the automatic generation of a suitable amount of training images. An arbitrary large number of computer graphics renderings of the 3D models can be generated from arbitrary viewpoints with varying object backgrounds and lighting conditions. Knowing the exact viewpoints during training enables the estimation of the viewpoint during detection for queries such as "car AND front-view". The utilization of 3D object models therefore represents an attractive means for a flexible and information rich training of multi-view object class and viewpoint detectors for image retrieval.

In the present work we propose an approach to part-based multi-view object class detection, which does not require any bounding-box, viewpoint, or part annotation for training. The training process exclusively relies on synthetic 3D object models with an automatic identification of suitable part positions. Most part-based approaches to multi-view object class detection make use of the spatial consistency of the detections of specific parts typically visible from a given viewpoint. These approaches predominantly learn individual detectors for each part under each viewpoint. However, due to viewpoint symmetries, part similarities and ambigu-

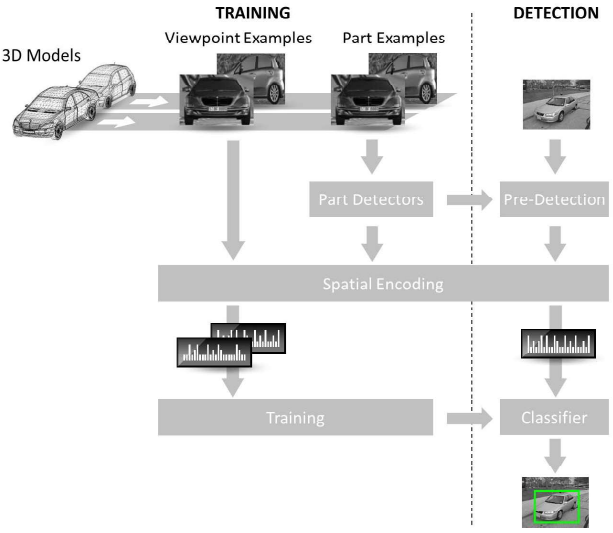


**Figure 1: Overview of our approach. Training (left side): viewpoint-specific part detectors are trained by means of synthetic 3D models. The spatial layout of all these part detectors is encoded in a spatial pyramid and discriminatively trained. Detection (right side): Regions of interest, generated by a part-based pre-detection step, are encoded and classified to obtain the final detection result.**

ities, the response behavior of such a specific detector on other parts and viewpoints can contain potentially valuable information which remains unused in previous approaches as they discard these "hallucinated" detector responses as nuisances. In this work we suggest combining all individual viewpoint detectors in a joint spatial pyramid encoding to fully exploit the information contained in the available set of detector responses. Initially, our approach uses synthetic images of 3D object models to automatically select the positions of relevant object parts from different views and train their appearances into part detectors. In a second step, the spatial layout of all detector responses is encoded in a spatial pyramid and discriminatively trained with non-linear SVMs on intersection kernels. As a result, we obtain a multi-view object class representation incorporating knowledge provided by all individual part detectors. The approach is unsupervised in the sense that no tedious bounding-box, viewpoint, or part annotations are required at training time. While it is purely trained on synthetic 3D object models, we show that our approach achieves state-of-the-art results in multi-view object class detection and viewpoint determination on current benchmarking data-sets with high robustness against partial occlusion. We use this multi-view object and viewpoint detector to annotate our image database, thus allowing us either retrieving objects of the desired viewpoint set or pre-filtering the image database for a standard query-by-image system.

The paper is structured as follows: Section 2 summarizes previous work on part models and multi-view object class detection. A system overview of the proposed approach is given in Section 3. Details for the training and detection procedure are presented in Section 4 and in Section 5. Experimental results and a comparison with state-of-the-art are given in Section 6.

**Figure 2: Examples of 3D object models of the three classes "car", "bike", and "iron".**

## 2. RELATED WORK

The approach to content-based image retrieval described in the present work relies on multi-view object class detection to identify the images relevant to the query. Most recent work on multi-view object class detection focuses on deriving geometric representations of an object class as a set of two-dimensional constellations of object parts for a few discrete viewpoints. The combination of 2D detectors to cover an entire object over a multi-view sphere has been the initial step towards a more comprehensive use of geometry for object class detection: Thomas et al. [21], for example, suggest linking Implicit Shape Models for specific viewpoints amongst each other, thereby achieving a detection over multiple viewpoints. In order to increase robustness towards pose changes, additional probabilistic layout models as well as local 2D geometric constraints have been introduced. Originally described in [8], the idea is taken up by [2] who introduce a simplified layout which assumes a set of mutually independent branch parts which only depend on a few root parts instead of modeling all pairwise interactions. The approach is further extended in [7] with discriminatively learnt part appearance, different heuristic layout models for the main viewpoints and a root part per viewpoint which covers the entire object, thereby increasing robustness. In [12], the approach of [7] is applied to building viewpoint-specific discriminative detectors with varying levels of supervision. Instead of modeling sparse sets of parts, a fixed grid-based subdivision of object views has been suggested in [10, 15] to detect regions of parts which conform to the training part layout. Savarese and Li [18] determine homographies of groups of local features in order to map large 2D image regions onto a collection of near-planar parts to form a viewpoint-independent 3D model; more recently, [20] introduced a probabilistic approach to learn affine constraints between sparse object patches. In [1], sparsely annotated 2D feature positions are factorized to obtain a 3D implicit shape model which extends the original implicit shape model to 3D transformations and occlusion issues.

More recently, the use of synthetic CAD models as training data source has been advocated to enable a more fine-grained viewpoint subdivision: in [16], local features are derived from synthetically rendered models to evaluate the global consistency of a 2D detection with respect to a 3D geometry; in [19], CAD models with semantic part annotations are used to learn a probabilistic spatial model of edge-based features.

In combining individual detectors in a common representation, our approach is similar in spirit to [14] who classify entire images into scene categories based on a set of object detector responses.

## 3. SYSTEM OVERVIEW

Figure 1 gives an overview of the multi-view object class detection approach presented in this paper. We rely on a database of 3D object models as training data source for

each object class that should be detected (see Figure 2 for some model examples of different classes). In contrast to other work [19], semantic 3D part labels which CAD designers sometimes assign to parts of the model geometry during the creation process (i.e. "wheel" or "car door") are not required in our approach, since we have frequently found these manual labels to be inconsistent. The 3D object models are used to generate two independent sets of training images by means of computer graphics rendering, the part examples and the viewpoint examples.

The part examples are used as training source for the discriminative learning of viewpoint-specific part detectors. The objective of this learning step is to automatically identify a layout of object parts which describes the characteristic appearance of an object class under a given viewpoint. For each viewpoint, individual part locations are automatically chosen as those regions which consistently possess dominant gradients across all object models of a class (c.f. Figure 3). While the selected part regions do not necessarily have a semantic meaning, this process ensures that the appearance of parts is sufficiently structured for detection. The part layout of our approach is then used in a second step to train viewpoint-specific part detectors from patches of the part examples at these chosen locations (c.f. Figure 4). More details are given in 4.2. Spatial layout models (c.f. Section 4.2.3), which are based on these part detectors, allow to determine regions of interest in an image which have a high likelihood of containing an object of the trained class. However, due to the differences in layout and appearance discriminativity of the different viewpoints, the scores of the spatial layout models do not yet allow for a comparison between viewpoints and classes which is necessary to rank the detection results with respect to their relevance for the query.

In order to establish a comparable ranking of the regions detected by the above described spatial layout models, we suggest here a joint spatial encoding of the responses of all part detectors in a detected region which is subsequently scored by a more powerful classifier, a nonlinear SVM with an intersection kernel. The training examples for this classifier are determined by applying the above described part detectors on a second set of example images, the viewpoint examples (c.f. Figure 6). On each of the viewpoint example images we apply all the part detectors resulting in a set of detector responses that include real detections (e.g. responses of front view part detectors on an actual front view example) as well as hallucinated detections (e.g. responses of front-view part detectors on a side-view example). The useful contribution of hallucinated detections for the overall detection of objects is illustrated in Figure 5, where the consistent response of a front-view detector (red) contributes to the evidence of a side view detection (green). Details for the spatial encoding are presented in 4.3.

During detection the spatial layout model generates on each test image a set of detection hypotheses for the viewpoint on which it was trained. The full set of the part detectors is then applied to these object hypotheses. The resulting spatial layouts, which encode all the individual part detector responses into a single spatial descriptor, are then classified by the nonlinear SVM. A non-maxima suppression discards all those object hypotheses which overlap by more than 50% with a higher-scoring object hypothesis. The remaining detections form the ranked query response on the test image.

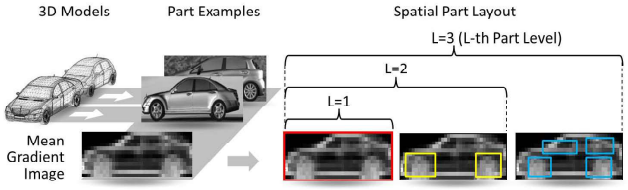


**Figure 3: Concept of the object class representation: for a specific viewpoint (here side-view) part examples are generated from renderings of 3D object models. We use these training examples to obtain a mean gradient image from which a spatial part layout for $L$ part levels is derived (see Section 4.2.1).**

More details of the detection process are given in Section 5.

# 4. TRAINING

This section outlines the necessary training steps. It starts with the use of 3D object models as training source, and is followed by the unsupervised training approach for the viewpoint-specific part detectors and the joint spatial encoding of these part detectors.

## 4.1 Training Examples

As shown in Figure 1, the approach presented in this paper is purely trained on synthetic 3D object models. Figure 2 gives some examples of our 3D object model database. The use of synthetic 3D object models as training source allows to generate training images (i.e. part examples and viewpoint examples) of an object from arbitrary viewpoints. For each training example, the projection of the 3D object model into the image allows to automatically determine the actual 2D bounding box of the object within the image and its viewpoint label. In addition, light conditions and background for each rendered image can be changed in order to account for the imaging conditions in real images.

## 4.2 Part Detectors

Learning the appearance of an object class must take into account large intra-class and viewpoint variations as well as partial occlusions and background. In addition, when dealing with part-based object class detection, object parts have to be chosen such that they are suitable for the training of discriminative classifiers. A manual annotation of these part positions is time consuming; moreover, there is no guarantee that the selected parts are suitable, i.e. sufficiently discriminative, for the training process. As a consequence, some authors propose a fixed part layout [10, 15] or suggest unsupervised approaches to localize suitable part positions. For example, in [7] the object is decomposed into six object parts, selecting the part positions such that the resulting patches capture a maximum of the object structure. However, a multi-view object class detection approach requires the choice of part positions on different viewpoints with different sizes, aspect ratios and appearance characteristics. The method of [7] results in a spatial part layout where each part has approximately the same size, and each viewpoint is subdivided into the same number of parts. However, this may not be a suitable approach for the multi-view representation of all classes, since for those viewpoints where the object covers a smaller area, the chosen patches could be too small and therefore may not contain sufficient structure to

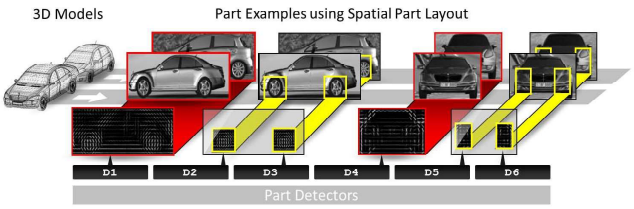


**Figure 4: The normalized part positions of the spatial part layouts and the corresponding part examples to train viewpoint-specific part detectors (here D1 to D6).**

be suitable for a discriminative classifier. The front-view of the bicycle class is a good example of a viewpoint for which a subdivision into one or two parts is adequate, whereas a bicycle side-view may require a more fine-grained part subdivision.

#### 4.2.1 *Spatial Part Layout*

In this work we extend the idea of [7] in order to determine suitable part positions while circumventing the above described problem of a spatial part layout with a fixed number of parts for different viewpoints. More specifically, we propose a method to derive a spatial part layout which decomposes the object model into $L$ part levels.
The concept of this object decomposition is shown in Figure 3. Starting with our database of 3D object models, we generate the training examples for the parts of a specific viewpoint (e.g. side-view). We scale all rendered training examples of a given viewpoint to their average size, apply a Laplacian filter mask and average over the filtered examples to obtain a mean gradient image for a specific viewpoint. On each part level $l$, we define an area $a$ for each part such that the object is decomposed into $2^{l-1}$ parts and $a \cdot 2^{l-1}$ equals about 70% of the area of the mean gradient image. For each part location, we sequentially choose a rectangle with area $a$ that captures the highest gradient over all training examples; the chosen area is masked out in the gradient image and the procedure is repeated until the $2^{l-1}$ parts are selected. Decomposing an object under a given viewpoint finally leads to a spatial part layout with $N$ viewpoint-specific part positions on $L$ part levels:

$$N = \sum_{l=1}^{L} 2^{l-1} = 2^L - 1. \quad (1)$$

Since the above described procedure is repeated for each of the $V$ viewpoints, we finally obtain $V \cdot N$ normalized part positions.

#### 4.2.2 *Training of Part Detectors*

Once the part layout for each viewpoint has been identified with the method described in 4.2.1 (see Figure 4), we resort to the HOG descriptor of [3] to encode the appearance of the parts. For each part position, a separate linear SVM classifier is learnt. The negative examples for the training of the linear classifier are initially chosen randomly from a background data set. After the initial training run, the classifier is refined by a bootstrapping procedure on an extended training set which has been augmented with the false positives and false negatives of the initial SVM classifier. Consequently, $V \cdot N$ discriminatively learnt object part detectors are obtained, each representing the appearance of an object part under different viewpoints; for an example see D1 to D6 in Figure 4.

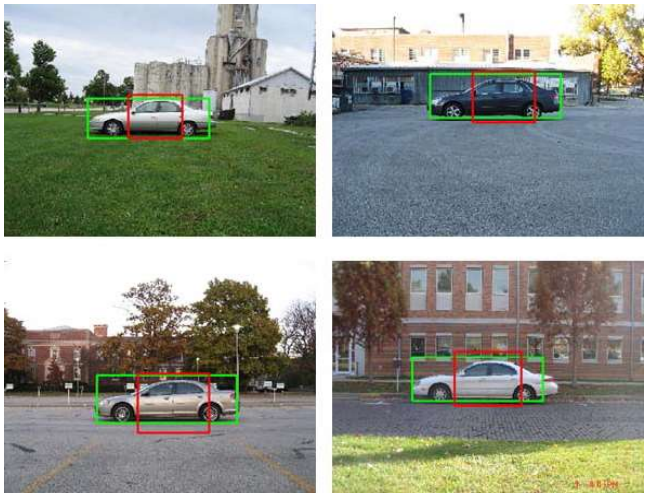

**Figure 5: The spatial encoding builds on the combination of all viewpoint-specific trained part detectors. For example, detector D4 (trained on front view images) also provides a consistent response on side view images; instead of discarding these responses, we exploit their information content in a joint encoding.**

#### 4.2.3 *Spatial Layout Model*

For the pre-detection step we establish for each viewpoint a spatial layout model as described in [6], based on the trained part detectors of Section 4.2.2. To this purpose, we reapply the part detectors to the part examples and model the spatial uncertainty for each detector with a Gaussian distribution. We follow the approach of [6] and use a tree structure and dynamic programming to compute the best locations for the parts within an image. We choose the tree root to be the center of the bounding box of the part detector defined on level $l = 1$. Different scales are covered by applying the method of [6] to an image pyramid representation.

### 4.3 Spatial Encoding

As described in Section 4.2, viewpoint-specific part detectors are derived in an unsupervised way. In the following, we introduce a spatial encoding to obtain a multi-view object class representation which jointly captures the spatial layout of all the responses of the individual part detectors on the viewpoint examples and allows to consistently rank detections for different viewpoints. Note that the scores of the spatial layout models described in Section 4.2.3 alone do not allow for such a ranking, mainly due to their differences in layout and appearance discriminativity of the different viewpoints.
Due to viewpoint symmetries, part co-occurrences and ambiguities the trained part detectors sometimes locate object parts at wrong positions or in viewpoints where these parts are not actually visible. Still, these "hallucinated" part detections often appear consistently within the object class. An example of such a hallucinated part detection is given in Figure 5: as expected, a part detector, which was trained on

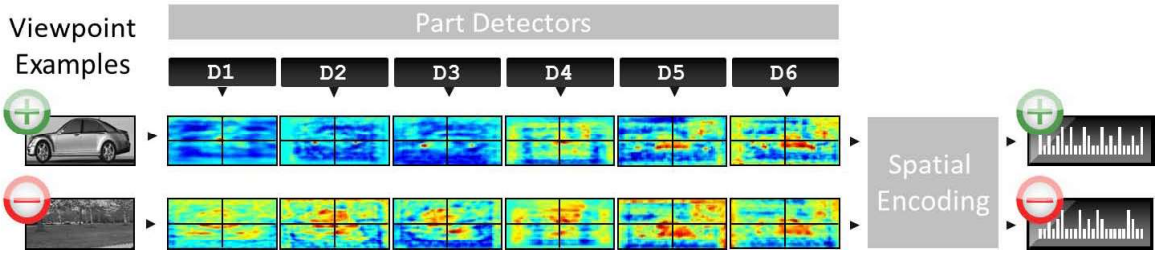


**Figure 6: We use all viewpoint-specific part detectors (D1 to D6) and encode the spatial layout of their responses on the training examples into a spatial pyramid.**

example images for cars from a side-view, provides consistent "true" responses on images showing the object from this viewpoint (green); however, another part detector, which was trained on example images for cars from a front-view, also provides consistent "hallucinated" (false positive) responses on these images (red). We suggest exploiting this kind of readily available additional information with a spatial encoding to combine the "true" as well as the "hallucinated" detector responses for a more discriminative multi-view object class representation.
The idea of the spatial encoding concept is shown in Figure 6. For each training instance from the set of viewpoint examples which show the object to be detected from different viewpoints, HOG features with the same layout as in Section 4.2 are computed densely and classified by all viewpoint-specific part detectors. Following the approach of [13], we rely on a spatial pyramid to encode the responses for each part detector within the area of a training instance. The spatial pyramid consists of $K$ levels of a fixed hierarchy of rectangular windows, each containing a histogram of the number of positive responses of each part detector in that window. We concatenate the resulting histograms of all windows and pyramid levels and obtain a histogram based object class representation with $d$ dimensions for $V \cdot N$ object part detectors:

$$d = VN \sum_{k=0}^{K} 4^k = \frac{1}{3} VN(4^{K+1} - 1). \quad (2)$$

Given the viewpoint examples and a set of negative examples, a nonlinear SVM classifier with an intersection kernel [11] is trained. In order to compensate for the initially random choice of negative training examples, a standard bootstrapping procedure (as in Section 4.2) selects the most difficult false positives and false negatives for the subsequent training iterations.

# 5. DETECTION

This section describes the two detection steps, the pre-detection based on a spatial layout model of part detectors for each viewpoint to obtain object hypotheses and the spatial encoding for a consistent and comparable reranking of those hypotheses.

## 5.1 Pre-Detection

In order to identify regions of interest which potentially contain an object of a given class, we rely on the viewpoint-specific spatial layout models of Section 4.2.3 to provide regions of interest. These models provide a detection score and a viewpoint label alongside each generated object hypothesis. However, since each spatial layout model for a given viewpoint is trained independently of all other viewpoints and relies on different layouts and appearance characteristics, the scores of the spatial layout model alone do not allow for a consistent ranking of the detections (see Section 6 for experimental results). Consequently, in the following step we build on the classification of the joint spatial encoding of the part detectors over all viewpoints in order to obtain a normalized and comparable detection score for each object hypothesis.

## 5.2 Reranking based on Spatial Encoding

The final detection result consists of a consistent and comparable scoring of the obtained object hypotheses based on the spatial encoding process in 4.3. Each object hypothesis is scaled to the size of its corresponding tree root (c.f. Section 4.2.3). The responses of all part detectors in the scaled hypothesis area are encoded in a spatial pyramid representation as described in 4.3 which is then classified by the nonlinear SVM to obtain the final detection score. Since the detection process can result in multiple overlapping hypotheses, a non-maximum suppression retains high scoring bounding boxes and discards those covered by a higher-scoring bounding box with an overlap of more than 50%.
As a result, each final detection is assigned both a detection score based on the spatial encoding as well as an approximate pose label based on the viewpoint-specific pre-detection with a part-based spatial layout model.

# 6. EXPERIMENTAL EVALUATION

This section outlines the experimental results we achieve with our approach on publicly available benchmark data sets.

## 6.1 Training Setup

For the part detectors and the spatial encoding the proposed approach relies exclusively on training data rendered from synthetic 3D models available from the distributors turbosquid.com, doschdesign.com and 3dvia.com. We use 24 car, 8 bicycle and 2 iron models for the respective classes (see Figure 2 for examples). We train viewpoint-specific part detectors and spatial encodings for five different viewpoints (i.e. $V = 5$), i.e. left, front-left, back-left, front and back, where the respective symmetric views are covered by applying the approach to the horizontally mirrored images. In our experiments, our approach performs best with part examples generated from a fixed azimuth angle for each discrete viewpoint and an elevation angle of $0°$ for the object classes car and bicycle and $40°$ for the object class iron, whereas multiple viewpoint examples are generated for each

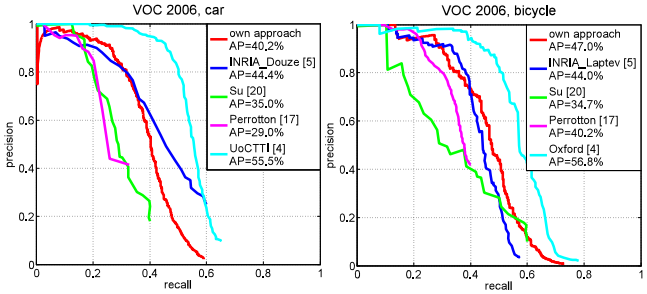


**Figure 7: Precsion/Recall curves for the PASCAL VOC 2006 car (left) and bicycle (right) data set of our approach (red curves) compared to state-of-the-art detectors.**

viewpoint by varying the azimuth angle by $\pm 9°$ and the elevation angle by $+9°$ for increased robustness towards small viewpoint variations. In order to train the linear SVMs for the part detectors as well as the nonlinear SVM classifiers for the spatial encoding, negative training examples are drawn from the PASCAL VOC 2006 training data set excluding the training images for the positive object classes. The part appearance is built on the HOG implementation of [7] with a HOG layout of 4 pixels per cell; we choose a spatial part layout consisting of $L = 4$ part levels, resulting in $N = 15$ part detectors per viewpoint. The spatial encoding is done with a spatial pyramid of $K = 2$ levels, resulting in a 1575-dimensional representation.

## 6.2 Evaluation Criteria and Data Sets

In order to evaluate the performance of our multi-view object class detector with respect to 2D ground truth bounding boxes, we use the detection quality criterion suggested by [5]: A predicted bounding box is considered correct if the overlap between the predicted bounding box and a ground truth bounding box exceeds 50%. If several bounding boxes are predicted in the same image area, only the highest scoring detection is considered as correct and the remaining detections are considered as false positives.

We evaluate our approach on the publicly available PASCAL VOC 2006 [5] data set for cars and bicycles and on the 3D Object Category [18] data set, the current state-of-the-art benchmark for multi-view object detection, for the classes car, bicycle and iron. The 3D Object Category data set has been explicitly designed as a multi-view detection benchmark, containing for each object class 10 different object instances, each shown in front of a varying background from 8 different 45°-spaced azimuth angles (left, front-left, front, front-right, right, back-right, back and back-left), 2 different elevation angles and 3 different distances.

## 6.3 VOC 2006

The precision/recall curves obtained with our proposed approach on the PASCAL VOC 2006 data sets for the classes car and bicycle are given in Figure 7 (red curves). For both data sets we provide the best performing approaches of the PASCAL challenge 2006 [5] (blue curves), the best performing approaches of the PASCAL challenge 2007 on the 2006 test set [4] (cyan curves) and the most recent multi-view approaches of [20] (green curves) and [17] (magenta curves). With 40.2% on the car data set and 47.0% on the bicycle data set, our detection approach achieves a higher average precision than these two multi-view approaches and can

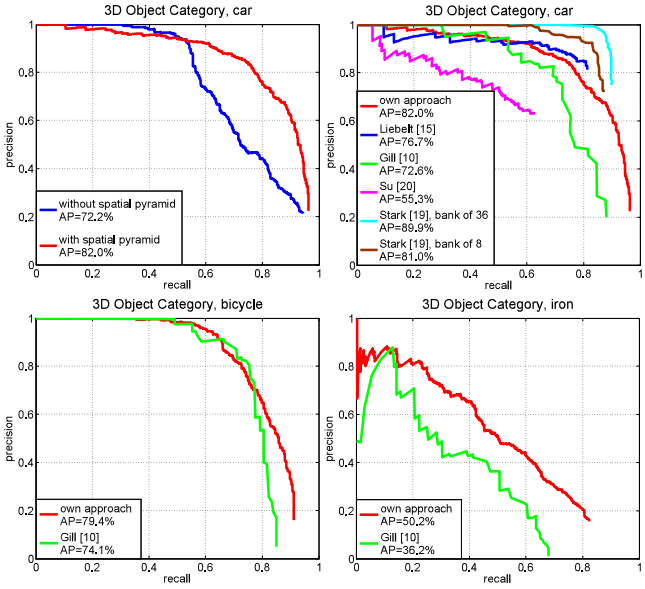


**Figure 8: The spatial encoding step increases the detection precision (upper left). Precision/Recall curves for the 3D Object Category data sets car (upper right), bicycle (lower left) and iron (lower right).**

compete with the best performing approaches of the PASCAL challenge 2007, despite being trained on a very different (because synthetically generated) data set. We observe that the viewpoint and appearance variations within the car test set are more pronounced than those within the bicycle class, which may be the reason for the observed performance difference of our approach: while the chosen synthetic bicycle models and the viewpoint variation of the viewpoint examples are sufficient to represent these variations of the bicycle test set, the synthetic car models and the corresponding viewpoint examples seem to be not representative enough.

## 6.4 3D Object Category

On the 3D Object Category test set we evaluate our approach on three different tasks: 2D localization, robustness to occlusions and 3D pose estimation.

### *6.4.1 2D Localization*

The precision/recall curves we obtain with our approach on the 3D Object Category data sets for the classes car, bicycle and iron are shown in Figure 8. To demonstrate the contribution of the proposed spatial encoding step (see Section 4.3), we evaluate our approach exemplarily on the car test set, once with and once without the encoding step as part of the detection process. Without the encoding step we exclusively rely on the detection score provided by the viewpoint-specific part models (see Section 5.1). The results are given in Figure 8 (upper left). Omitting the spatial pyramid encoding step (blue curve with 72.2%) results in an average precision which is significantly below the precision obtained with the proposed spatial encoding step (red curve with 82.0%). By including the spatial pyramid encoding into the detection process, the detection precision can thus be substantially increased. In Figure 8 (upper right), we compare our result on the car data set to the most recent approaches of [10, 15, 19, 20]. As can be seen, with

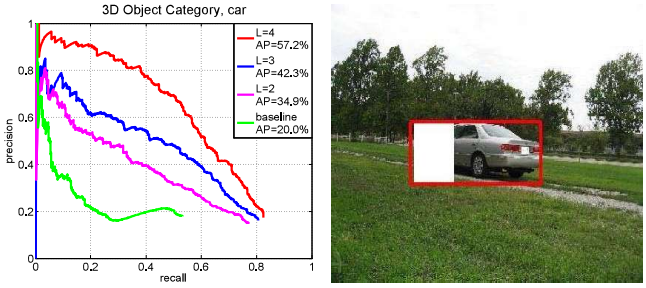


**Figure 9: The quality of our approach (red curve) under partial occlusion is shown in comparison to a baseline approach (green) which is based on [3]. Spatial part layouts with fewer part levels result in a lower average precision (blue and magenta).**

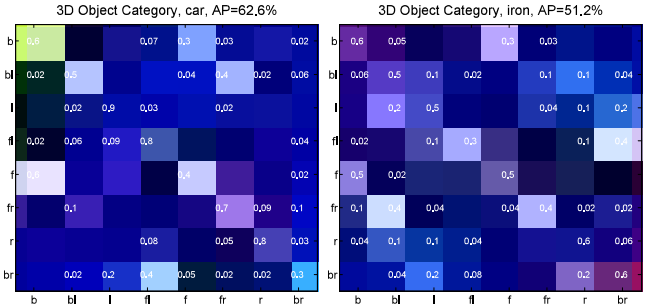


**Figure 10: Confusion matrices (rows: ground truth, columns: estimates) for the 3D Object Category data sets car and iron.**

82.0% our detection approach outperforms the approaches of [10, 15, 20] and is comparable to the results achieved by [19]. Note that the best result of [19] is based on a set of 36 viewpoint-specific shape models; when using only a comparable number of 8 viewpoint models, we outperform their average precision (brown curve with 81.0%). The precision/recall curves obtained with our approach on the bicycle (lower left) and iron data set (lower right) are given in Figure 8. On both data sets, we compare to the approach of [10], showing that with 79.4% on the bicycle test set and 50.2% on the iron test set, our approach achieves a higher average precision than the multi-view approach of [10], despite being trained on a different, i.e. synthetically generated, data set. Some successful detection results of our approach on the 3D Object Category test sets are shown in Figure 11.

### 6.4.2 *Occlusion*

In order to assess the quality of our approach in the presence of occlusion, we modify the 3D Object Category data set for cars and generate a test set with artificial partial occlusions. For this purpose, we replace 30% of the annotated ground truth for all images by a white area. An example for this modified test set is given in Figure 9. In this experiment, the object class detector with the same settings as in Section 6.4.1 is applied to the modified data set, i.e. without any retraining or adaptation. In order to compare the performance of our part-based approach to detection methods which do not use parts, we implemented a baseline approach based on [3]. This baseline approach is trained on the viewpoint examples described in 6.1; it consists of one HOG descriptor covering the entire object under each viewpoint, which is classified by a linear SVM. We use a sliding window approach and rely on the non-maximum suppression step of Section 5.2 to combine the viewpoint-specific classifier responses.

As can be seen in Figure 9, with 57.2% compared to 20.0%, our part-based approach (red curve) outperforms the baseline approach which relies on a global description of the object (green curve). Since our object representation is based on several parts of different sizes due to the chosen spatial part layout on $L = 4$ levels, the partial occlusions have less effect on the overall description of the object than on the baseline approach.

We also assess the influence of the chosen spatial part layout. We exemplarily evaluate a part layout with three part levels ($L = 3$) and a part layout with just two part levels ($L = 2$). The results for these part layouts are shown in Figure 9, too (blue and magenta curves). As expected, the average precision is reduced when decreasing the part levels due to the lack of information of the small object parts. Note that with two part levels, our approach still performs better than the baseline approach, although the reduction in performance is considerable when compared to a spatial part layout with $L = 4$ part levels. Consequently, a part layout with sufficiently many part levels is necessary to achieve increased robustness against partial occlusions.

### 6.4.3 *Pose Estimation*

In Section 5.2, we mention that our approach is able to provide an approximate pose label based on the viewpoint-specific pre-detection step. Figure 10 shows the resulting confusion matrices on the car and iron data sets for classifying all true positive detections into the 8 azimuth angles defined by the 3D category data set. For cars, we observe that neighboring viewpoints are rarely confused. Confusion is more pronounced for opposing views due to the symmetries inherent in the car class. For example, 30.0% of back views are classified as front views. Note that the pose estimation relies exclusively on the pose label provided by the best-scoring viewpoint-specific part layout model of the pre-detection step. Since the parts of these layout models are selected in an unsupervised way which does not take into account the inter-viewpoint discriminativity of the parts, such a behavior cannot be avoided in the present approach. Still, with an average precision of 62.6%, our approach is comparable to other reported results, notably [20] with approximately 67.0% or [15] with 70.0%. For the iron class, we observe the same behavior with respect to diagonal views; no published pose estimation results on the iron data set are currently available for comparison.

## 7. CONCLUSION

In the present work, an approach to viewpoint independent retrieval of images containing objects of a certain class is described. We extend the common part-based approaches, which usually exploit the spatial consistency of the detections of visible parts from a given viewpoint, with our concept of "hallucinated" part detections by simultaneously learning the spatial layout of "true" as well as "hallucinated" part detector responses in a joint spatial encoding. In contrast to other approaches, our proposed method is unsupervised in the sense that no tedious bounding-box, viewpoint, or part annotations are required. Although it is exclusively trained on a few synthetic 3D object models, we achieve state-of-the-art results in multi-view object class detection and viewpoint

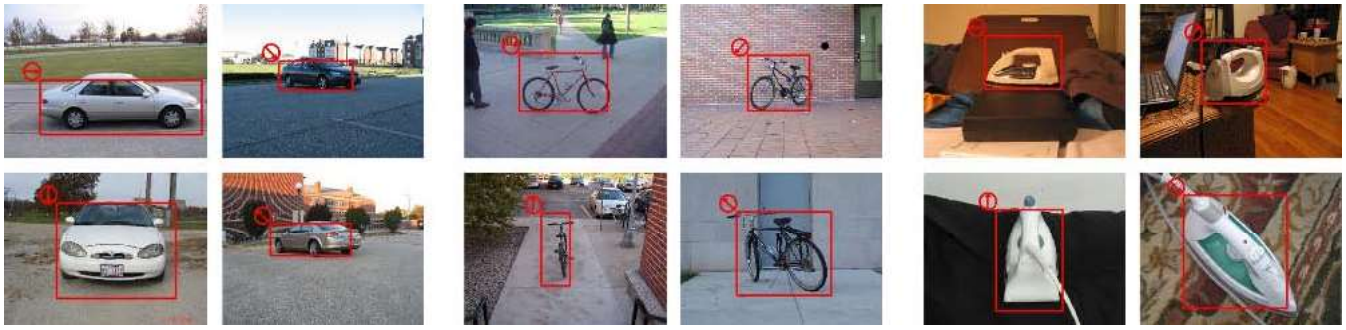

**Figure 11: Some successful detection results of our approach on the 3D Object Category data sets car (left), bicycle (center) and iron (right).**

estimation on current benchmark data sets and demonstrate increased robustness towards partial occlusion.

## 8. ACKNOWLEDGMENTS

J. Schels, J. Liebelt and K. Schertler acknowledge support by BMBF grant SiVe FKZ 13N10027.

## 9. REFERENCES

[1] M. Arie-Nachmison and R. Basri. Constructing implicit 3D shape models for pose estimation. In *International Conference on Computer Vision*, pages 1341–1348, 2009.

[2] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *Conference on Computer Vision and Pattern Recognition*, pages 10–17, 2005.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. Technical report, 2007.

[5] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. Technical report, 2006.

[6] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61:55–79, 2005.

[7] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[8] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *Transactions on Computers*, C-22(1):67–92, 1973.

[9] Y. Gao, M. Wang, J. Shen, Q. Dai, and N. Zhang. Intelligent query: Open another door to 3d object retrieval. In *International Conference on Multimedia*, 2010.

[10] G. Gill and M. Levine. Multi-view object detection based on spatial consistency in a low dimensional space. In *Symposium of the German Association for Pattern Recognition (DAGM)*, pages 211–220, 2009.

[11] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *International Conference on Computer Vision*, pages 1458–1465, 2005.

[12] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *European Conference on Computer Vision*, pages 408–421, 2010.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.

[14] L. Li, H. Su, Y. Lim, and L. Fei-Fei. Objects as attributes for scene classification. In *ECCV First International Workshop on Parts and Attributes*, 2010.

[15] J. Liebelt and C. Schmid. Multi-view object class detection with a 3d geometric model. In *Conference on Computer Vision and Pattern Recognition*, pages 1688–1695, 2010.

[16] J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3D feature maps. In *Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[17] X. Perrotton, M. Sturzel, and M. Roux. Implicit hierarchical boosting for multi-view object detection. In *Conference on Computer Vision and Pattern Recognition*, pages 958–965, 2010.

[18] S. Savarese and L. Fei-Fei. 3D generic object categorization, localization and pose estimation. In *International Conference on Computer Vision*, pages 1–8, 2007.

[19] M. Stark, M. Goesele, and B. Schiele. Back to the future: Learning shape models from 3d cad data. In *British Machine Vision Conference*, pages 106.1–11, 2010.

[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 *International Conference on Computer Vision*, pages 213–220, 2009.

[21] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. V. Gool. Towards multi-view object class detection. In *Conference on Computer Vision and Pattern Recognition*, pages 1589–1596, 2006.