Building a semantic part-based object class detector from synthetic 3D models

Johannes Schels, Jorg Liebelt, Klaus Schertler, Rainer Lienhart

Angaben zur Veröffentlichung / Publication details:

Schels, Johannes, Jorg Liebelt, Klaus Schertler, and Rainer Lienhart. 2011. "Building a semantic part-based object class detector from synthetic 3D models." In *IEEE International Conference on Multimedia and Expo, 11 - 15 July 2011, Barcelona, Spain*, 1–6. Piscataway, NJ: IEEE. https://doi.org/10.1109/icme.2011.6011850.

BUILDING A SEMANTIC PART-BASED OBJECT CLASS DETECTOR FROM SYNTHETIC 3D MODELS

Johannes Schels, Jörg Liebelt*, Klaus Schertler**

EADS Innovation Works
Munich, Germany
{johannes.schels, joerg.liebelt,
klaus.schertler}@eads.net

Rainer Lienhart

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

ABSTRACT

This paper presents a new approach for multi-view object class detection based on part models. While most existing approaches have in common that they use real images for training, our approach requires only a database of synthetic 3D models to represent both the appearance and the geometry of an object class. We use semantically equivalent object points on 3D models to build part models and encode the local appearance of the parts by a discriminative learning method that applies AdaBoost to histograms of gradients. The geometric configuration of the parts is represented by spatial distributions which are also directly derived from the 3D models. For recognizing an object in an image, our model provides object hypotheses which are re-ranked with global appearance models. The 2D localization is evaluated on the PASCAL 2006 data set for cars and bicycles, showing that its performance can compete with state-of-the-art detection results.

Index Terms— 3D models, multi-view object class detection

1. INTRODUCTION

Object class detection is one of the primary research topics in computer vision; it is of relevance to numerous applications, ranging from retrieval tasks to robotics. Fischler and Elschlager [1] originally introduced an approach to the problem consisting of a constellation of basic image parts connected by spring-like links. The idea has recently regained attention, notably in [2, 3, 4]. Other publications resort to the heuristic selection of parts as subregions without semantic meaning and model few sparse viewpoints [4]. Alternatively, brute-force regular part subdivisions are suggested in [5] which may introduce a large per-part variance into the training process. In the present work, we propose an approach to semantic part-based object class detection which relies on synthetic 3D CAD models as training data; part annotation is performed in 3D space once per model and allows generating arbitrary amounts of precisely labeled 2D part training annotations over the entire view sphere. Unlike previous work, both appearance and geometry of parts are learnt exclusively from synthetic data. We compare our approach to state-of-the-art detectors and show that we achieve comparable results using a single synthetic training procedure without requiring any data set specific retraining or adaptation.
The paper is structured as follows: Section 2 summarizes previous work on part models and multi-view object class detection. An overview of the training procedure is given in section 3. In section 4, details for the detection process are provided. Experimental results and a comparison with state-of-the-art on the PASCAL VOC 2006 [6] data set for cars and bicycles are outlined in section 5.

2. RELATED WORK

Recent work on multi-view object class detection can be divided into two groups which rely on different choices for the geometric representation of an object class, either by modeling two-dimensional constellations or by building 3D approximations. 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. [7] suggest linking several Implicit Shape Models to achieve 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 [1], the idea is taken up by [2] who introduce a simplified layout which assumes a set of mutually independent object parts. The approach is further extended in [4] with discriminatively learnt part appearance and different heuristic layout models for multiple viewpoints, thereby increasing robustness. Hoiem et al. [8] suggest a Layout Conditional Random Field to model part interactions for a set of discrete viewpoints from the pixel level upwards. Instead of modeling sparse sets of parts, a fixed grid-based subdivision of object views has been suggested in [5] who propose a greedy algo-

*acknowledge support by BMBF grant SiVe FKZ 13N10027

Fig. 1. Examples for 3D models from our training database.

rithm to detect regions of parts which conform to the training part layout. While being robust, these approaches mostly assume heuristic rules for choosing part regions, rely on a sparse set of viewpoint models and require tedious manual viewpoint and part annotations in 2D.

In contrast, viewpoint-annotated training data can be used to dynamically build 3D representations to better address the possible viewpoint variations of object classes. Savarese and Li [9] estimate 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, [10] introduced a probabilistic approach to learn affine constraints between sparse object patches. In [11], 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. Alternatively, 3D models have been suggested as training data. Lowe [12] resort to flexibly aligning groups of consistent edge segments from CAD models by probabilistic matching. Heisele et al. [13] generate training sets from synthetic 3D models for an active learning algorithm. Recently, Yan et al. [14] suggested to collect patches from 2D images with 3D viewpoint annotations and to map these patches onto an existing 3D CAD model. In [15], local features are derived from synthetically rendered models to evaluate the global consistency of a 2D detection with respect to a 3D geometry. Although these 3D representations may be closer to the actual object class geometry, they can be more difficult to train and may not be necessary for pure 2D detection tasks. In the following, we outline an approach that builds on 2D part representations learnt from semantically corresponding object regions and the exhaustive generation of automatically annotated 2D training data from 3D CAD models, thereby combining the main advantages of the two domains described above.

3. TRAINING

The training procedure of our approach makes use of standard 3D models as they are typically used in computer graphics applications. Some examples of such 3D models are shown in Figure 1.

A database of models of the same type (e.g. cars) is used to represent each of the object classes to be detected. Initially, every model in the database is manually annotated by specifying semantically equivalent object points (for example the left front wheel) in 3D space using a 3D labeling tool. As our approach is based on a small number of synthetic 3D models to represent a class, the labeling effort is rather small compared to a manual annotation of bounding boxes of an entire training database of real images. Figure 2a illustrates two car models of our database with annotated semantically equivalent object points. The annotated model database is then used to generate a large number of training images of the 3D models from arbitrary viewpoints in front of arbitrary backgrounds (Figure 2b). In addition, we also vary the light conditions for each image in order to cope with the imaging conditions in real images. The exact locations of the semantically equivalent object points as well as the exact bounding box of the object are determined in each training image by projecting the annotated 3D models into the 2D image space (Figure 2b). They are subsequently used to train a part model, consisting of per-part appearance and part geometry, as well as a global appearance model for each discrete viewpoint. Both the part model and the global appearance model together form an object detector for a specific viewpoint.

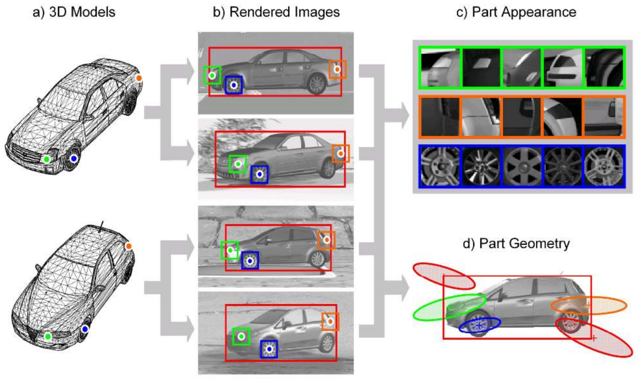

Fig. 2. Overview of the training steps using annotated synthetic 3D models (a). The part appearance (c) and part geometry (d) are learnt from the rendered training images using the projected positions of the 3D parts (b).

3.1. Learning the Part Appearance

From the known locations of the semantically equivalent object points within the training images, we generate for each part a collection of small patches representing the appearance of the given semantic part (Figure 2c) over all models in the training database, which are then fed into a learning method applying AdaBoost to histograms of gradients [16]. Discrete AdaBoost defines a strong binary classifier H as a linear combination of 'weak' classifiers h:

$$H(x) = sgn(\sum_{t=0}^{T} \alpha_t h_t(x)) \ . \qquad (1)$$

x is a fixed-size patch, as shown in Figure 2c. In this work

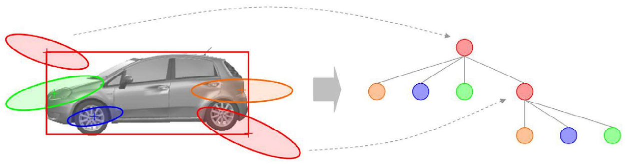

Fig. 3. Tree structure of our spatial model which allows predicting a bounding box by propagation from the root location (upper left corner) to the node (lower right corner). The Figure is best viewed in color.

a weak classifier consists of a feature f and threshold θ such that

$$h(x) = \begin{cases} +1 & \text{if } f(x) \geq \theta \\ -1 & \text{otherwise} \end{cases} . \quad (2)$$

We define our features f in terms of differences between two histogram bins of a HOG-descriptor g of [16]:

$$f(x) = g_i(x) - g_j(x) \quad i \neq j \, . \quad (3)$$

where $g_i(x)$ is the i-th and $g_j(x)$ is the j-th bin of the HOG-descriptor encoding a patch x. For our experiments (see section 5) we choose a quadratic HOG-layout with 5x5 cells and 18 orientation bins, resulting in a 450 dimensional HOG-descriptor g. As the result of the training a strong classifier H is assigned to each specific viewpoint of each semantically equivalent object point.

3.2. Learning the Part Geometry

To learn the geometry model of the parts of an object class, we make use of the known positions of the semantic parts and the upper left and lower right corner of the bounding box within a training image. We model the spatial distribution of these points as Gaussians (Figure 2d) and arrange them in a tree structure. Figure 3 shows such a tree structure of our spatial model, where the upper left corner is the root of the tree, the lower right corner is a node and the object points are the leaves. Consequently, the bounding box of an object can be predicted by propagation from the root location (i.e. the upper left corner) to the node (i.e. the lower right corner). The conditional spatial distributions $s_{ij}(c_i, c_j)$ between the location c_j of a parent p_j and the location c_i of a child p_i are modeled as

$$s_{ij}(c_i, c_j) = K e^{-((c_i - c_j) - m_{ij}) S_{ij}^{-1} ((c_i - c_j) - m_{ij})} \, . \quad (4)$$

where m_{ij} is the mean of the relative location to the parent location c_j, S_{ij} is the covariance matrix and K the normalization factor.

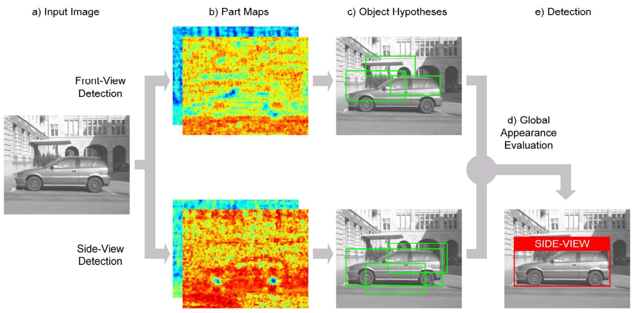

Fig. 4. Overview of the detection steps for recognizing an object in an image (a). We apply our strong classifiers to generate dense part maps (b). From the part maps we derive a final cost map for each viewpoint detector by using a dynamic programming approach; the local minima of those maps provide the object hypotheses (c). We score (d) and merge those hypotheses to obtain our final detection result (e). The Figure is best viewed in color.

3.3. Learning the Global Appearance

Detections based on our part models allow determining image regions which have a high probability of containing a set of object parts. However, the scores generated by the part model alone are not suitable for ranking detections for different viewpoints between each other, since the part models for different viewpoints vary in complexity and cannot easily be normalized. Some parts may be generally more difficult to detect, thus introducing a bias into the detection which influences the overall score; in addition, occluded objects may result in lower scores of the part model, although the occluded bounding box is predicted accurately. As a consequence, a different method for scoring the regions of interest (ROI) predicted by the part model has to be introduced. Subsequently, we use our part models to only provide ROIs (i.e. object hypotheses), which are then classified in a separate stage. This approach is similar to the scoring approach of the exemplar model in [17]. Instead of training an SVM classifier with real images as in [17], we rely on the global appearance of our rendered synthetic 3D models to train a linear SVM classifier. For this purpose, a HOG layout is chosen to cover the entire object in our training images (see Figure 2b). The negative examples for training are initially chosen randomly from a background data set. After the initial training, the classifier is refined with a bootstrapping procedure on an extended training set which has been augmented with the false positives and false negatives of the initial classifier. As a final result, each viewpoint detector has a linear SVM that classifies the global appearance of the provided ROIs.

4. DETECTION

Figure 4 illustrates the necessary steps for recognizing an object in a query image (Figure 4a). We apply the classifiers

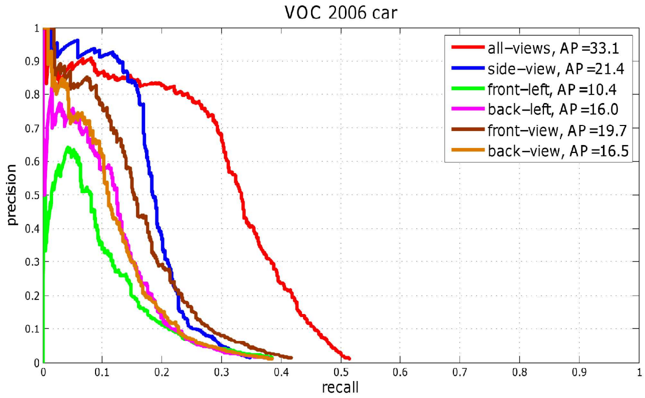

Fig. 5. Precision/Recall for the PASCAL VOC 2006 car data set for each individual viewpoint detector and the final combined multi-view detector (red).

of each viewpoint detector to obtain a cost map for each part (see part maps in Figure 4b). Dense classification of all possible patches in the image results in a dense cost map for each object point. Building on the efficient matching algorithm proposed by [3], an overall cost map is derived; its local minima indicate object hypotheses corresponding to a probable configuration of object parts (green boxes in Figure 4c). The location of a local minimum provides the upper left corner of the bounding box and by propagation to the node of the spatial model, we locate the position of its lower right corner. Eventually, we apply the global appearance SVM for the detected viewpoint to each object hypothesis in order to obtain a detection score (Figure 4d).

As our viewpoint detectors are defined on a fixed scale, the above described procedure is applied to each level of an image scale pyramid in order to generate scored object hypotheses on different scales. As shown in Figure 4c), the detection process can result in multiple overlapping detections. In order to determine the most promising single detections (Figure 4e), we rely on a non-maximum suppression where bounding boxes overlapping with higher-scoring boxes by more than 50% are discarded.

5. EXPERIMENTAL EVALUATION

This section presents the experimental results achieved with the proposed approach. We use the publicly available VOC 2006 data set to evaluate the performance of individual viewpoint detectors and the 2D localization of our approach compared to state-of-the-art results.

5.1. Evaluation Criteria and Data Sets

In order to evaluate the performance of our detector with respect to 2D ground truth bounding boxes, we use the detection quality criterion suggested by [6]. A predicted bounding box is considered correct if the overlap between this predicted bounding box and a ground truth bounding box exceeds 50%. Multiple detections are penalized. If a system predicts several bounding boxes, only one box is considered correct, the remaining detections are considered as false positives. The average precision scores a system.

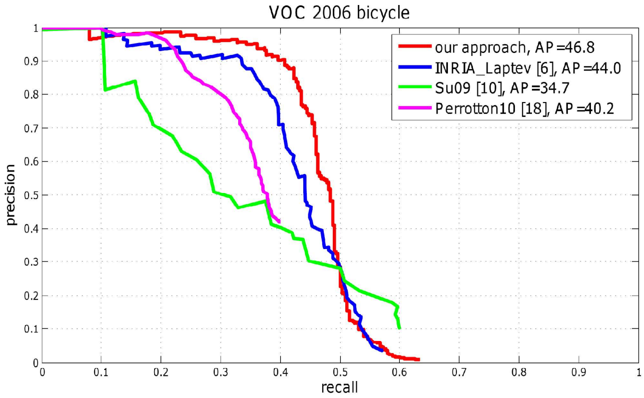

Fig. 6. Precision/Recall for the PASCAL VOC 2006 bicycle data set of our approach compared to state-of-the-art detectors.

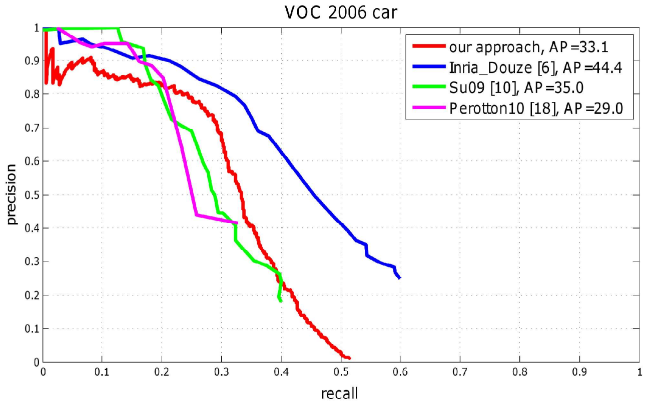

Fig. 7. Precision/Recall for the PASCAL VOC 2006 car data set of our approach compared to state-of-the-art detectors.

Our approach is evaluated on the PASCAL VOC 2006 [6] data set for cars and bicycles. In order to train the part and the global appearances, we rely on a background data set to provide negative training examples. For this purpose we use the PASCAL VOC 2006 training data set after excluding the training images for the respective positive object classes.

5.2. Training Setup for the Viewpoint Detectors

For training the object class car we use 25 synthetic 3D object models where 16 semantically equivalent object points (such as wheels or headlights) are specified for each model. The class bicycle is trained with 8 synthetic 3D models where 11 points (such as parts of the front and back wheel or handle bar) are annotated for each model. Note that the manual labeling has to be performed only once in 3D space, i.e. only 88 points have to be labeled for all 8 3D models of the object class bicycle. Subsequently, these 3D annotated points allow generating arbitrary amounts of precisely labeled 2D part training annotations over the entire view sphere. All our 3D object models are available from different free and commercial CAD model databases such as doschdesign.com or turbosquid.com.

In order to generate the training images for a viewpoint detector, we render 10.000 images with slight variations of the camera parameters; for symmetric views, only one classifier is trained and applied to both horizontally mirrored images. We vary the azimuth angle in a range of -8° to +8°, the elevation angle in a range of 0° to 8° and the scale in a range of -0.10% to +0.10% as this setting performed best in our experiments. We train five viewpoint detectors (i.e. side-view, front-left, back-left, front-view and back-view) for both object classes. A tree-structured spatial layout assumes that part appearances are independent which requires a non-overlapping part selection. For each viewpoint, patches of size 40x40 pixels represent the part appearance of those object points which are visible after perspective projection of the 3D model from that viewpoint. We choose a quadratic HOG-layout (5x5 cells) with 8 pixel per cell to encode the local and viewpoint-specific HOG-layouts with 4 pixel per cell to encode the global appearance.

5.3. Individual Viewpoint Detectors

The performance of each individual viewpoint detector on the VOC 2006 car test set is shown in Figure 5. Each detector is applied to all images of the test set, resulting in a lower recall because each detector is only responsible for a specific view. Note that some viewpoints, such as the profile views, can be detected more easily by our approach due to their more discriminative appearance. The combined multi-view detector achieves a significantly higher recall while retaining a consistently high precision which is due to the use of the re-scoring global appearance classifier on the detections provided by the part model.

5.4. Results on VOC 2006

Figure 6 shows precision/recall curves for the PASCAL VOC 2006 bicycle test set and Figure 7 shows the precision/recall curves for the car test set. On both data sets we provide the precision/recall curves of two multi-view approaches [18, 10] and the best-performing approaches of the PASCAL 2006 challenge. On the bicycle data set, our approach achieves a higher average precision (46.8%) than all other state-of-the-art detectors, although we train our bicycle detector on purely synthetic (i.e. non data set specific) training data of 8 models. On the car data set, our result (33.1%) can compete with the approaches of [18, 10] despite its being trained on a different

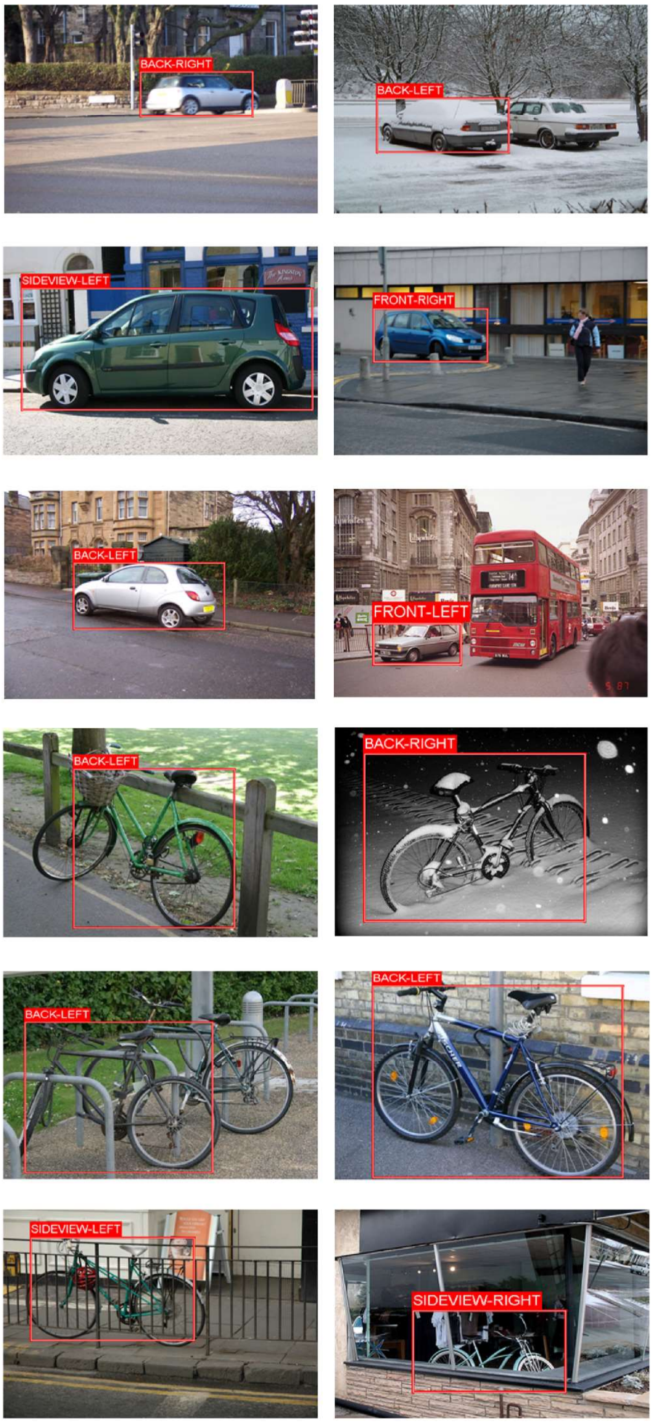

Fig. 8. Some successful detection results of our approach on the PASCAL VOC 2006 car (first three rows) and bicycle (last three rows) data sets. Each detection also provides an approximate viewpoint estimate.

(synthetic) data set. We observe that the appearance variations within the test set of the car class 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 are sufficient to represent these variations, the synthetic car models seem to be not representative enough. Figure 8 shows some examples for successful detections on the PASCAL 2006 test set. Note that our approach additionally provides approximate 3D orientation estimates for each detection.

6. CONCLUSION

In this work we present a new approach to part-based multi-view object class detection. In contrast to most existing work, our approach relies exclusively on synthetic 3D models to represent the object class to be trained. We use semantically equivalent object points on 3D models to build part models as well as global appearance models for arbitrary viewpoints. For recognizing an object, we derive object hypotheses from part models and score these hypotheses with global appearance classifiers. Even though only a single synthetic training procedure without any data set specific retraining or adaption is applied, we achieve results comparable to state-of-the-art on two different test sets. In contrast to other part-based approaches, our approach additionally provides approximate pose estimations for the detected objects. Currently, we use a labeling tool to manually annotate semantically equivalent object points; in future work, we will focus on establishing these points in an unsupervised way and extending the method to more, potentially non-rigid, object classes.

7. REFERENCES

[1] M. A. Fischler and R. A. Elschlager, “The representation and matching of pictorial structures,” *IEEE Transactions on Computers*, vol. 22, 1973.

[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*, 2005.

[3] P. Felzenszwalb and D. Huttenlocher, “Pictorial structures for object recognition,” *International Journal of Computer Vision*, vol. 61, 2005.

[4] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” in *Conference on Computer Vision and Pattern Recognition*, 2008.

[5] Gurman Gill and Martin Levine, “Multi-view object detection based on spatial consistency in a low dimensional space,” in *Symposium of the German Association for Pattern Recognition (DAGM)*, 2009.

[6] M. Everingham, A. Zisserman, C.K. Williams, and L. V. Gool, “The PASCAL Visual Object Classes Challenge 2006 (VOC2006) results,” Tech. Rep., University of Oxford, University of Edinburgh, KU Leuven, 2006.

[7] Alexander Thomas, Vittorio Ferrari, Bastian Leibe, Tinne Tuytelaars, Bernt Schiele, and L. V. Gool, “Towards multi-view object class detection,” in *Conf. on Computer Vision and Pattern Recognition*, 2006.

[8] D. Hoiem, C. Rother, and J. Winn, “3D LayoutCRF for multi-view object class recognition and segmentation,” in *Conference on Computer Vision and Pattern Recognition*, 2007.

[9] S. Savarese and L. Fei-Fei, “3D generic object categorization, localization and pose estimation,” in *IEEE International Conference on Computer Vision*, 2007.

[10] 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 *IEEE International Conference on Computer Vision*, 2009.

[11] Mica Arie-Nachmison and Ronen Basri, “Constructing implicit 3D shape models for pose estimation,” in *International Conference on Computer Vision*, 2009.

[12] David G. Lowe, “Three-dimensional object recognition from single two-dimensionalimages,” *Artificial Intelligence*, vol. 31, 1987.

[13] B. Heisele, G. Kim, and A. J. Meyer, “Object recognition with 3D models,” in *British Machine Vision Conference*, 2009.

[14] Pingkun Yan, Saad M. Khan, and Mubarak Shah, “3D model based object class detection in an arbitrary view,” in *International Conference on Computer Vision*, 2007.

[15] J. Liebelt, C. Schmid, and K. Schertler, “Viewpoint-independent object class detection using 3D feature maps,” in *Conference on Computer Vision and Pattern Recognition*, 2008.

[16] Navneet Dalal and Bill Triggs, “Histograms of oriented gradients for human detection,” in *International Conference on Computer Vision and Pattern Recognition*, 2005.

[17] O. Chum and A. Zisserman, “An exemplar model for learning object classes,” in *Conference on Computer Vision and Pattern Recognition*, 2007.

[18] X. Perrotton, M. Sturzel, and Michel Roux, “Implicit hierarchical boosting for multi-view object detection,” in *Conference on Computer Vision and Pattern Recognition*, 2010.