A Multi-View Probabilistic Model for 3D Object Classes

Min Sun[1*] Hao Su[1,3*] Silvio Savarese[2] Li Fei-Fei[1]

[1]Dept. of Computer Science, Princeton University, USA, {msun,haosu,feifeili}@cs.princeton.edu

[2]Dept. of Electrical and Computer Engineering, University of Michigan at Ann Arbor, USA, silvio@eecs.umich.edu

[3]Dept. of Computer Science, Beihang University, China,

Abstract

We propose a novel probabilistic framework for learning visual models of 3D object categories by combining appearance information and geometric constraints. Objects are represented as a coherent ensemble of parts that are consistent under 3D viewpoint transformations. Each part is a collection of salient image features. A generative framework is used for learning a model that captures the relative position of parts within each of the discretized viewpoints. Contrary to most of the existing mixture of viewpoints models, our model establishes explicit correspondences of parts across different viewpoints of the object class. Given a new image, detection and classification are achieved by determining the position and viewpoint of the model that maximize recognition scores of the candidate objects. Our approach is among the first to propose a generative probabilistic framework for 3D object categorization. We test our algorithm on the detection task and the viewpoint classification task by using "car" category from both the Savarese et al. 2007 and PASCAL VOC 2006 datasets. We show promising results in both the detection and viewpoint classification tasks on these two challenging datasets.

1. Introduction

Imagine a busy street in downtown Rome or Beijing. There might be cars coming from all possible directions, pedestrians walking on the sidewalk, or scooters crossing the street. The ability to interpret the scene, recognize the objects within, estimate their locations and poses is crucial if one wants to avoid obstacles, interact with other people, and find a target location. However, what appears to us as natural may become tremendously difficult for an artificial vision system. How do we handle occlusions? How do we deal with intraclass and pose variability of objects? Most of the current researches in object categorization [7, 5, 9, 16, 1, 31] have focused on modeling object intraclass variability in single views (within a small range of planar rotations) or as a mixture of single view models [27, 33]. Very few methods have leveraged on the intrinsic geometric aspects of object categories.

In this paper, we introduce a generative probabilistic framework for learning visual models of 3D object categories. Our approach is inspired by a number of recent work for representing 3D object categories where appearance information and geometric constraints are combined (Sec. 2). We represent an object as a coherent ensemble of parts linked across different viewpoints (Sec. 3). Each part is represented by a distribution of characteristic appearance elements (codewords). A generative model is used for learning the relative position of parts within each viewpoint, as well as corresponding part locations across viewpoints (Sec. 4). Epipolar constraints are applied to ensure that the configuration of each part is consistent under viewpoint transformations. This is one of the first *generative probabilistic 3D object models that incorporates such geometric constraints in a principled way*. Unlike other methods that focus on discriminative recognition of characteristic patches, our model is able to generate a coherent distribution of parts that are representative of the object class and are consistent under viewpoint transformation (see Fig. 1). We have used this model to predict whether an object class is present in the image or not, estimate the location of the object as well as determine its viewpoint (Sec. 5). Our experiments show superior detection and viewpoint classification results on the 3D Objects dataset [25], and comparable results on PASCAL VOC 2006 datasets [6].

2. Previous Work

Researchers have done extensive work in single object recognition from multiple poses. Notably, [3, 32, 8, 24] have proposed methods for identifying and recognizing single objects such as a cereal box or a Teddy bear, under arbitrary viewing conditions. These methods are successful due to their reliance on the identification of strong geometrical constraints and highly discriminative object features. However, such constraints are not adequate in object categorization problems in which shape and appearance variability of each object class must be accounted for. Similar limitations

*indicates equal contributions

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

Figure 1. Viewpoint images generated from our learned 3D object model for the car category. Each image shows a mixture of randomly sampled parts as well as their geometric configurations from the learned model, super-imposed on an image of a car in this viewpoint. The fully trained car category model consists of 18 parts across all 32 discretized viewpoints on the full viewing sphere of the object class. Each part is represented by a color coded bounding box. Contrary to most of the mixture of viewpoints object models, the learned parts in our model are maintained across different viewpoints. This figure is best viewed in color under PDF magnification.

exist for representations based on aspect graphs [12].

Recently, a number of methods have brought new ideas into the problem of representing object categories from multiple views. Thomas et al. [30] have explored the idea of linking features across multiple views. Kushal et al. [14] propose to connect groups of features across views which retain pairwise local geometrical transformations in an MRF framework. Other methods [11, 34] propose to represent the object category by using a rough 3D model, on top of which the typical distribution of appearance element is learned. These methods have the advantage of yielding a compact view-point invariant representation of the object category, as opposed to [30, 14] but fail to accommodate intra-class 3D shape variability. Except for [4], none of the above models have the capability of generating spatial and appearance information of the object category under arbitrary viewing conditions. The framework presented in [25, 26] represents an object category as a collection of view-invariant regions linked by transformations that capture the relative change of pose among parts. The model has the ability to generate unseen views, but achieves limited accuracy in classification due to the lack of an explicit background model.

Our method tries to overcome some of the limitations of previous works. Unlike [30], but similar to [14, 25, 26] our model parts are linked across views under affine transformations. As opposed to [11, 34], we generalize on both appearance and geometry. Unlike [30, 25, 26], our method offers a principled way for learning a rigorous generative model for 3D objects[1].

3. Model Representation

Our proposed model learns a coherent part-based object representation across instances and viewpoints of an object class. In addition to being able to discover object parts that appear consistently across object instances, it can simultaneously establish part-level correspondences by exploiting the underlying geometric structure across nearby viewpoints (see Fig 1).

3.1. The Generative Model

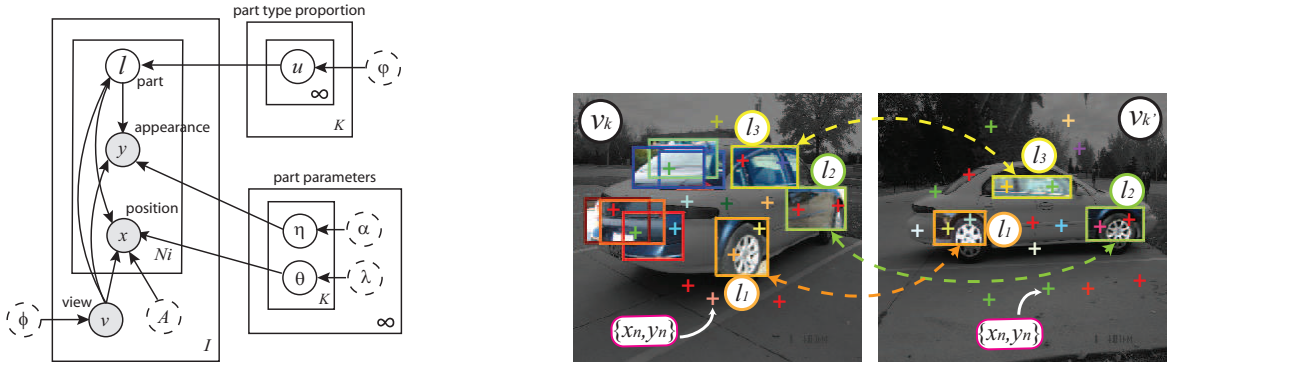
A generative graphical model (Fig.2 top left panel) illustrates the backbone of the model structure. We first introduce the main representation of this model. We then describe in details on how epipolar geometry is introduced to regularize the coherent representation of object parts across different viewpoints. The bottom panel of Fig.2 summarizes the important variables in the model.

Imagine a generative process for sampling an object image given a learned 3D model representation. We start with sampling a particular viewpoint v out of K possible 3D views of this object. Given v, we then sample an object part type assignment l for each image patch. For a robust 3D representation of an object class, it is important to point out that different viewpoints would render very different part distributions (e.g. wheel parts are more likely to be observed in a sideview of a car than a frontal view of a car). Given a part type assignment l for this patch, we sample the image patch based on the appearance (y) distribution as well as position (x) distribution.

We now introduce the generative process in a more mathematically rigorous way. In order to simplify the model description, we first assume that the different viewpoints of an object class are affine aligned. In other words, viewpoints, scales and orientations are first assumed to be matched. We will relax this assumption once the generative model is described. The overall model structure is a Dirichlet Process (DP) Gaussian mixture. The choice of using DP is particularly important to accommodate a variable number of object part types per viewpoint. We start with an image i.

1. **Generate viewpoint** v_i
   - Draw viewpoint $v_i \sim \text{Uniform}(\phi)$. For a total of K viewpoints, $\phi = 1/K$.

[1] [14] is the only other reported 3D object class model that is probabilistic in nature. The generative representation of our model differs from their discriminative framework.



Variables				Parameters	
	indexed by	value	distribution	distribution	variational parameters
appearance: y	$n = 1 \dots N_i$	$y = \omega\|_{1 \dots W}$	$y \sim Mult(\eta)$	$\eta \sim Dir(\alpha)$	τ
position: x	$n = 1 \dots N_i$	$x \in \mathbb{R}^2$	$x \sim N(\theta)$	$\theta \sim NW(\lambda)$	$\epsilon = \{\beta, m, W, \nu\}$
part type assignment: l	$n = 1 \dots N_i$	$l = t\|_{1 \dots \infty}$	$l \sim Mult(\mathbf{s})$ $s_t = u_t \prod_{j=1}^{t-1}(1 - u_j)$	$\mu_t \sim Beta(\varphi)$	$\gamma = \{\gamma_{kt};\ k = 1 \dots K,\ t = 1 \dots T\}$
viewpoint: v	1	$v = 1 \dots K$	$v \sim Mult(\phi)$	N/A	N/A

Figure 2. **Top left:** A graphical representation of our model. Nodes represent random variables, edges denote dependencies, and plates denote replicated structure. Note that visual word y, feature position x, and viewpoint v are observed during learning. The constraints described in Sec. 3.2 are not captured by the graphical model. **Top right:** Generation process of two particular viewpoints of a car. v is the viewpoint index of the object. Given a pair of images from nearby viewpoints, part types assignment l (color coded boxes) are sampled from the viewpoint specific spatial configuration and enforced to satisfy the part correspondences constraints (dashed arrows). Feature (colored crosses) positions and codewords are sampled given each part l and viewpoint v. **Bottom:** Summary of variables and their related parameters.

2. **For each of the $n = 1, ..., N_i$ feature patches of the image, generate part type assignment l_n**
   - Draw part type assignment $l_n \sim \text{Mult}(\mathbf{s}_k)$, where $\mathbf{s}_k = \{s_{kt} | t = 1 \dots \infty\}$, and s_{kt} is computed from $s_{kt} = u_{kt} \prod_{m=1}^{t-1}(1 - u_{km})$, the expected part proportion of part type t in viewpoint k. u_{kt} is a global parameter indicating the number of parts with type t in viewpoint k, where $u_{kt} \sim \text{Beta}(\varphi)$. Note that the Dirichlet Process model allows us to consider an infinite number of possible part types per viewpoint.

3. **For each of the $n = \{1, ..., N_i\}$ feature patches of the image, generate the patch appearance (y_n) and patch location (x_n)**
   - Given viewpoint v_i and part type assignment l_n, sample the patch location $x_n \sim N(\theta_{kt})$, where $N(\cdot)$ indicates a Gaussian distribution. θ_{kt} is the global parameter of the Gaussian distribution for part type t and viewpoint k. θ_{kt} is governed by a Gaussian-Wishart distribution $NW(\lambda_t)$, the conjugate prior of a Gaussian distribution, where λ_t consists of hyperparameters $\{\beta_0, m_0, W_0, \nu_0\}$.
   - Given viewpoint v_i and part type assignment l_n, sample the patch appearance $y_n \sim \text{Mult}(\eta_{kt})$, where $\text{Mult}(\cdot)$ indicates a Multinomial distribution. η_{kt} is the global parameter of the Multinomial distribution for part type t and viewpoint k. η_{kt} is governed by a Dirichlet distribution $Dir(\alpha_t)$, the conjugate prior of a Multinomial distribution, where α_t is the hyperparameter.

Putting all the observable variables (feature patches X and Y, viewpoints V) and latent variables (part type assignments L) together with their corresponding parameters, we can write down the joint probability of this model.

$$P(X, Y, L, V, \boldsymbol{u}, \boldsymbol{\eta}, \boldsymbol{\theta}) = \prod_i^I \{P(v_i|\phi) \prod_n^{N_i} P(l_{in}|\mathbf{s}_{v_i}) \cdot P(y_{in}|\eta_{\{v_i,l_{in}\}}) P(x_{in}|\theta_{\{v_i,l_{in}\}})\} \cdot \prod_k^K \prod_t^\infty \{P(u_{kt}|\varphi) P(\eta_{kt}|\alpha_t) P(\theta_{kt}|\lambda_t)\} \quad (1)$$

where I is the total number of images, and N_i is the number of feature patches in image i.

Affine transformation So far we have introduced the model by assuming complete affine alignment of viewpoints, scales and orientations. An important contribution of our model is the ability to automatically find an affine transformation across different viewpoints of object images such that object recognition can be accomplished under arbitrary view conditions. This is achieved by introducing the affine transformation variable A for each image, as shown in the top left panel of Fig.2. The transformation A operates on the patch location x to obtain the optimal patch alignment for the model corresponding to viewpoint v. The modified patch location probability becomes:

$$P(x|\theta, A) = N(x; \hat{A}\mu + b, \hat{A}\Lambda\hat{A}^T) \quad (2)$$

where $A = \left[\hat{A}\ \ b\right]$, $\theta = \{\mu, \Lambda\}$, μ and Λ are the means and the covariance matrices of the Gaussian distribution of the reference object coordinate respectively. Sudderth et al [29] have proposed a DP mixture model for 2D object classes by allowing a translational transformation for different patches of objects. But their model does not capture the 3D structure

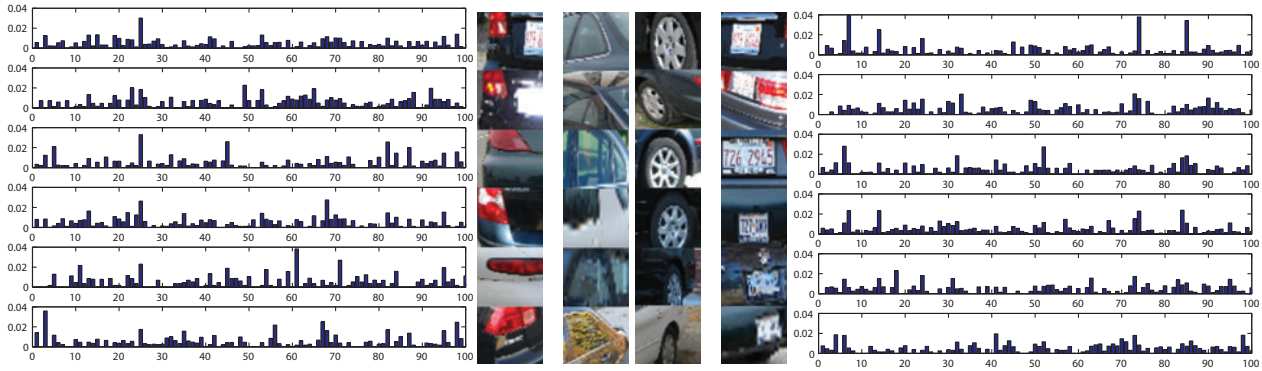


Figure 3. Example of learned parts from different viewpoints of the car model. Note that all parts are learned automatically. The first image column shows image regions of the rear light part of a car model sampled from different viewpoints. The corresponding patch appearance distributions of the rear light samples are shown to the left (only the top 100 most likely codewords are displayed out of a total of 1000 codewords.). We observe nearby viewpoints give rise to more similar appearance distributions. The second and third image columns show samples of the side window part and rear wheel part under different viewpoints. The fourth image column shows samples of the license plate part under different viewpoints. The corresponding patch appearance distribution of license plates are displayed to the right of this column.

of an object class, nor does it encode part correspondences across viewpoints.

3.2. Constraining Generative Model Across Viewpoints

Up to this point, for each viewpoint, object class are only governed by the hyper-parameters α_t, λ_t, and φ, where α_t and λ_t control the appearance and position variability of part type t, and φ regulates the part type proportions in different viewpoints. When learning and fitting 3D object class models, the number of possible transformations allowed for the object part locations and appearances can be orders of magnitude larger than those observed in 2D image based object models. This poses a potentially detrimental problem for a model regulated only by these hyper-parameters. In order to reduce the complexity of the parameter space, we propose to take advantage of the geometrical relationships regulating the appearance of 3D objects observed from multiple vantage points. The challenge here is to strike a good balance between enforcing such relationships and maintaining the ability to generalize well across instances within the same object class. We propose to enforce the following constraints:

- **Part appearance similarity constraints.** It is reasonable to assume that parts in nearby viewpoints share similar appearance. For example, the front wheels in the first row of images in Fig. 1 are likely to have similar distributions of visual words. When building the model, we enforce corresponding parts to look similar by penalizing the sum of the square errors of the difference of the visual word distributions in nearby viewpoints. A similar approach is proposed by [20]. Fig. 3 illustrates the appearance similarities of corresponding parts in different views.

- **Part configuration constraints via epipolar geometry.** The 2D configuration of object parts in the image plane changes as the viewpoint varies. We can use epipolar geometry [10] to constraint the 2D part configuration in images corresponding to nearby viewpoints. More specifically, given images of a particular object instance taken from nearby viewpoints, we obtain epipolar constraints via feature matching[2]. We then apply these constraints on the locations of the corresponding parts across nearby viewpoints during model learning (Fig 4). Each constraint is equivalent to a linear constraint in part locations (see Sec. 4.1 for details).

Such constraints enable our model to capture the underlying structure of a 3D object category across viewpoints. While simple in nature, these constraints play an important role in model learning, allowing much more accurate estimations of number of parts, their locations and alignment. In the next section, we describe how to incorporate these implicit constraints during model learning.

4. Model Learning

Given the model representation in Fig. 2, the goal of learning[3] is to infer the latent variables and estimate the hidden parameters by maximizing the log marginal probability

$$\ln P(X,Y) = \ln \sum_{V,L} \int P(X,Y,L,V,\boldsymbol{u},\boldsymbol{\eta},\boldsymbol{\theta})\, d\boldsymbol{u}\, d\boldsymbol{\eta}\, d\boldsymbol{\theta}. \quad (3)$$

In our current setting, we make the following assumptions during the learning process: i) object bounding boxes are provided; and ii) viewpoint labels are given to all images.

[2] In our current learning process, only one object instance (and its multiple examples across different viewpoints) is used for establishing this constraint. About 90% of the image pairs are automatically matched correspondences. But about a dozen missed pairs are hand pruned.

[3] Due to space limitation, more detailed derivations can be found in an accompanying technical report on the authors' website.

4.1. Updates of the Model Parameters

Computing the exact marginal distribution is intractable. We employ variational EM to learn the model in the stick-breaking representation [28]. Using the variational distribution, we maximize $L(X,Y)$, the lower bound of $\ln P(X,Y)$, and fit the variational parameters by coordinate ascent. The mean-field variational distribution equations are

$$q(V,L,\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\eta}) = q(V,L)q(\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\eta}) \quad (4)$$

$$q(V,L) = \prod_{i=1}^{I} q(v_i|\delta_i) \prod_{n=1}^{N_i} q(l_{in}|\rho_{inv_i}) \quad (5)$$

$$q(\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\eta}) = \prod_{k=1,t=1}^{K,T} q(u_{kt}|\gamma_{kt})q(\eta_{kt}|\tau_{kt})q(\theta_{kt}|\epsilon_{kt}) \quad (6)$$

where δ_i is the variational multinomial parameter over the K viewpoints, ρ_{inv_i} is the variational multinomial parameter over the feature part type assignment, γ_{kt}, τ_{kt}, and ϵ_{kt} are the variational beta, the variational Dirichlet, the variational Gaussian-Wishart parameters of part type t in viewpoint k, and finally T is the truncation number of part types in our model [2].

In the following, we summarize the variational update equations for each of the model parameters[4]. Particularly, we show how the 3D structure constraints introduced in Sec. 3.2 are applied to guide the variational distribution towards convergence to the true posterior.

Spatial Parameters Updates Feature patch location x is governed by the Gaussian distribution of parameter θ_{kt}, whose variational parameters are consisted of β_{kt}, m_{kt}, W_{kt} and ν_{kt}. Here, m_{kt} encodes the expected mean of part centers, and the other parameters model the degree of intra-class variation of the part center locations. The expected mean of part centers m_{kt} and $m_{k't}$ across nearby viewpoints are constrained by the epipolar geometry. In our model, we capture the epipolar geometry between image in viewpoint k and image in viewpoint k' by using the affine fundamental matrix $F_{kk'}$:$[m_{kt};1]^T F_{kk'}[m_{k't};1] = 0$, where $F_{kk'}$ is estimated from the epipolar geometry between viewpoint k and k' using the reference instance[5].

The variational parameters are then updated as follow:

$$\beta_{kt} = \beta_0 + N_{kt}, \qquad \nu_{kt} = \nu_0 + N_{kt} \quad (7)$$

$$\begin{aligned} W_{kt}^{-1} &= W_0^{-1} + N_{kt} \times S_{kt} + N_{kt}(\overline{x}_{kt} - m_0) \\ &\quad +\beta_0(m_{kt} - m_0)(m_{kt} - m_0)^T \end{aligned} \quad (8)$$

where the sufficient statistics of the spatial terms are:

$$N_{kt} = \sum_{i,n} \delta_{ik}\rho_{inkt}, \quad \overline{x}_{kt} = \tfrac{1}{N_{kt}} \textstyle\sum_{i,n} \delta_{ik}\rho_{inkt}x_{in} \quad (9)$$

$$S_{kt} = \frac{1}{N_{kt}} \sum_{i,n} \delta_{ik}\rho_{inkt}(x_{in} - \overline{x}_{kt})(x_{in} - \overline{x}_{kt})^T \quad (10)$$

There is no close form update rule for m_{kt} and A_i due to the geometric constraints. Hence, we formulate the update problem into a convex optimization problem with spare linear equality constraints, detailed in the technical report.

Appearance Parameters Updates As described in Sec. 3.2, our model enforces a feature patch appearance similarity in nearby viewpoints k and k'. We define a regularized marginal likelihood as:

$$O(X,Y,G) = (1-\zeta)L(X,Y) - \zeta R(G) \quad (11)$$

$$R(X,Y,G) = \frac{1}{2} \sum_t \sum_{(k,k')\in E} \sum_w^W (\tau_{kt}^w - \tau_{k't}^w)^2 \quad (12)$$

where W is the total number of visual codewords, G is the graph structure over the viewing sphere, E is the set of the edges defined in G[6], and ζ is the parameter that determines the significance of the regularization, whose value can range from 0 to 1. The variational parameter updates then become

$$N_{ktw} = \sum_i^I \sum_{n\in\{y_{in}=w\}}^{N_i} \delta_{ik}\rho_{inkt} \quad (13)$$

$$\hat{\tau}_{kt}^w = \alpha_t^w + N_{ktw}, \qquad \widetilde{\tau}_t^w = \frac{1}{K}\sum_k^K (\hat{\tau}_{kt}) \quad (14)$$

where $\hat{\tau}_{kt}$ is the updated τ_{kt} when ζ is set to 0, $\widetilde{\tau}$ is τ_{kt} when ζ is set to 1, and N_{ktw} is the sufficient statistics of the multinomial distribution. We further define $\tau_{kt} = (1-\pi)\widetilde{\tau} + \pi\hat{\tau}$. Similar to [20], we update τ in a greedy approach. Starting from $\pi = 0$, we use gradient descent algorithm to search for π such that $O(X,Y,G)$ decreases.

Part Type Proportion Parameters Update The update equations for the variational parameters of the part type are straightforward.

$$\gamma_{kt,1} = 1 + N_{kt}, \quad \gamma_{kt,2} = \varphi + \textstyle\sum_{f=t+1}^{T} N_{kf} \quad (15)$$

Part proportions in different viewpoints differ as the object turns in the 3D space. It is therefore important for our model to adopt the reordering approach [13] to adjust the proportion of parts.

Latent Variables Updates Since viewpoint v is provided in training, $\boldsymbol{\delta}$ is fixed during learning. $\boldsymbol{\rho}$ has a close form update rule. Please refer to the technical report for more details.

[4] All hyper-parameters $\{\phi, \alpha, \lambda, \varphi\}$, except A, are fixed.

[5] We use a sinlge object instance with all viewpoints captured as the reference instance.

[6] In this paper, G is a graph of eight nodes, indicating eight discretized viewpoints. Each viewpoint, or node, has four neighboring viewpoints, hence four Es.

4.2. Implementation Details

We have described a principled framework for learning each parameter of the generative model through variational inference. In principle, all training images can be applied simultaneously to jointly update these parameters. This requires a joint estimation of affine transformations in every EM iteration. The update is equivalent to solving a semidefinite programing problem whose complexity is quadratic to the number of images. We therefore adopt an incremental learning framework to curtail the amount of computation.

Neal and Hinton [22] provide a theoretical ground for incrementally learning of mixture models via sufficient statistics updates. At every iteration of the incremental learning, we fix the sufficient statistics of the parameters $N_k t$, $\overline{x}_{kt}$, $\overline{S}_{kt}$, and N_{ktw} associated with the previously learned images in all later updates.

Initialization We initialize our model using one single object instance (reference instance) across all viewpoints. This can be done by applying an existing feature matching algorithm in all pairs of the viewpoint images [35]. We then apply variational EM learning to all the reference instance training images.

Accurately learning part correspondences across viewpoints is critical for the rest of the learning. To demonstrate the importance of the additional constraints introduced in Sec. 3.2, we compare the initialized model with two simplified models – a basic DP Gaussian mixture model, and a DP Gaussian mixture model with epipolar constraints but no appearance constraints. Fig. 5 demonstrates that the full model shows the best part correspondence compared to the two simpler versions. But epipolar constraint alone can already boost the correspondences significantly. Fig. 6 shows examples of the resulting average images of the aligned parts using the full model.

Incremental Update Given a new training image, we estimate its affine transformation with respect to the reference instance in the corresponding viewpoint. The estimation is done by solving a semi-definite programming problem. Meanwhile, the rest of the model will be updated by adding the sufficient statistics of the new image. In this scheme, the model estimates the affine transformation for one image at a time.

Part Expansion Learned parts of an object class do not overlap with each other. But given the learned geometric configurations among parts, we can now further generate additional object parts by linearly interpolating and extrapolating between nearby parts. Empirical experiments confirm such expansion is helpful in object recognition tasks when objects are embedded in cluttered background.

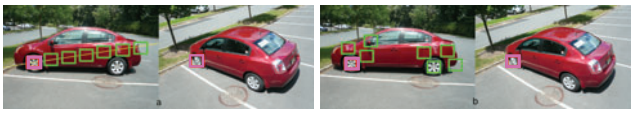


Figure 4. Image pairs show the candidate part correspondences (green boxes) and the correct one (red box) in the left figure across nearby viewpoints given a part (red box) in the right figure. In the **left** image pair, our full model enforces the candidate part correspondences in the left figure to belong to the epipolar line (the yellow line). In the **right** image pair, a model without the epipolar constraints yield arbitrary part correspondence locations in the left figure. This, in turn, increases the likelihood of obtaining erroneous part detections and correspondences, such as the back wheel of the car.

Figure 6. Average images of aligned parts across 15 instances.

5. Object Classification, Detection and Viewpoint Recognition

Given an unknown image, we can use the learned model to determine whether an instance of the object class is present, and to estimate the corresponding viewpoint. Inspired by the implicit shape model for object recognition [16], we accomplish the task in the following steps:

- **Extract features and propose candidate parts** We use Hessian-Affine feature detector [21] , Maximally stable extremal regions (MSER) [19] detector, and canny edge detector [23] to detect locally adapted elliptical regions. A feature codebook of size 1000 is obtained by vector quantizing the SIFT descriptors computed over these detected regions [18].

  Given a test image, we first detect a number of candidate parts for each of the learned parts in a particular viewpoint and scale. This is done by running a scanning window search method. The appearance similarities of the candidate parts are evaluated against the learned part model by using spatial pyramid matching [15] as the similarity score.

- **Object localization and viewpoint classification by mean-shift voting** Our object detection algorithm is similar to [16]. Given the spatial extent and the appearance similarity scores of the candidate parts, we look up the relative position of the object center to the object part in the learned spatial models. Each of the candidate part proposes an object center in the image. A mean-shift procedure is then applied to locate local maxima of the possible object centers based on each viewpoint model of the object class. Using this approach, we could handle multiple instance detection.

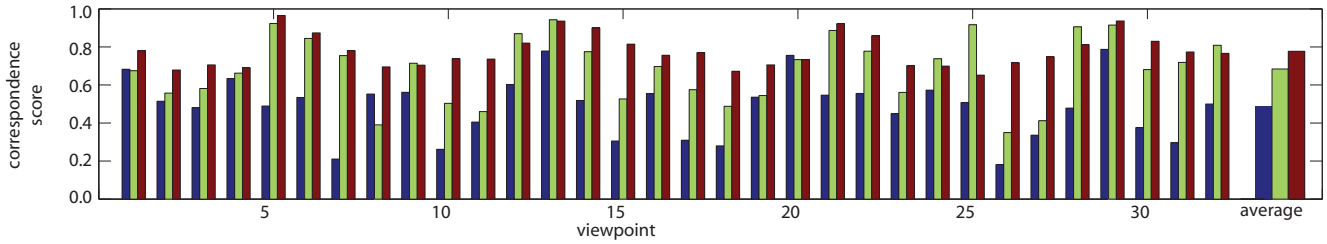


Figure 5. Correspondence score comparison in 32 different viewpoints among the DP Gaussian mixture model ("DP", blue), a DP Gaussian mixture with only the epipolar constraint ("DP-epipolar", green), and our full model (red). Average scores across all 32 viewpoints are shown on the far right. The full model consistently outperforms the other two. The score is defined as the percentage of the ground truth feature correspondences that are maintained in the part-level assignments.

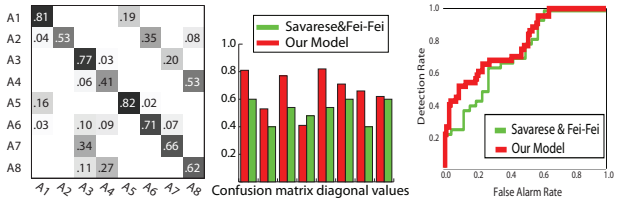
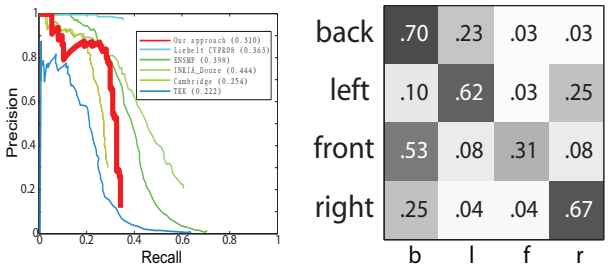


Figure 7. Detection and viewpoint classification using the Savarese et al. dataset [25]. **Left:** confusion matrix of the viewpoint classification. **Center:** diagonal element of our average confusion matrix (red) compared with the one from [25] (green). **Right:** binary object classification and detection result (ROC) (red) compared with the one from [25] (green).



(a) detection (b) viewpoint classification

Figure 8. Experimental results of PASCAL VOC 2006 data. **(a)** Detection results measured in precision-recall curve of our model (red) compared to [17] and the detection result of the PASCAL VOC 2006 challenge [6]-INRIA_Douze , [6]-INRIA_Laptev, [6]-TKK, [6]-Cambridge, and [6]-ENSMP. Average precision (AP) scores are shown in the legends. **(b)** Confusion table of the four viewpoint classification task. The average performance is 62% across the diagonal of the confusion.

We first test our model on the car dataset of the multi-view object dataset proposed by [25]. We assess the performances of our algorithm to localize cars and classify the viewpoints of the car. The car dataset comprises 320 images from up to 10 object instances. Each instance comprises 8 angles, 2 heights and 2 distances, a total of 32 viewpoints. We train our 3D object class model by using 160 images from 5 object instances with known viewpoint labels. The remaining 5 object instances are used as testing images. Binary classification result for car category is reported in Fig 7. Our model shows a superior performance over [25]. We also show consistently higher classification results across 8 viewpoints[7] compared to [25] (results in Fig 7, and examples in Fig.9 Top row).

We also conduct a car detection and viewpoint classification experiment by using the more challenging PASCAL VOC 2006 car dataset. We use the training data provided by PASCAL and the multi-view object datasets. The object class model is trained by assuming 8 different viewpoints for cars. Fig. 8(a) shows the car detection results of our model compared with the other state-of-the-art algorithms. Our algorithm performs on par with most of these discriminative methods. In addition, we show for the first time a quantitative viewpoint classification results on the PASCAL VOC 2006 car dataset. We test the performance of the model in the PASCAL test images by collapsing the 8 viewpoints into 4 views[8]. Fig.8(b) shows the confusion table result of the 4-way viewpoint classification task. Note that there is a relatively large confusion between the front and back view points. Besides the large visual similarity between these two views, we attribute the low discriminative ability of our model to the unbalanced training set (PASCAL VOC 2006 cars), which contains a much larger number of back-view car instances. Parts that are reliably detected provide strong evidence for object detection and viewpoint classification. For example, our model can robustly detect car wheels, which significantly contributes to the detection of left and right views. The bottom 2 rows of Fig.9 presents some sample recognition results.

6. Conclusions

We have proposed a rigorous generative model for learning the 3D structure of object classes. The model captures the geometric configurations of different parts of an object class linked across different viewpoints. Epipolar constraints are employed to ensure part consistencies when training object parts. We test the model in a car detection and car viewpoint classification experiment. We are espe-

[7]This is done on correctly detected cars.

[8]PASCAL data provides viewpoint labeling of only 4 views: front, right, left, back.

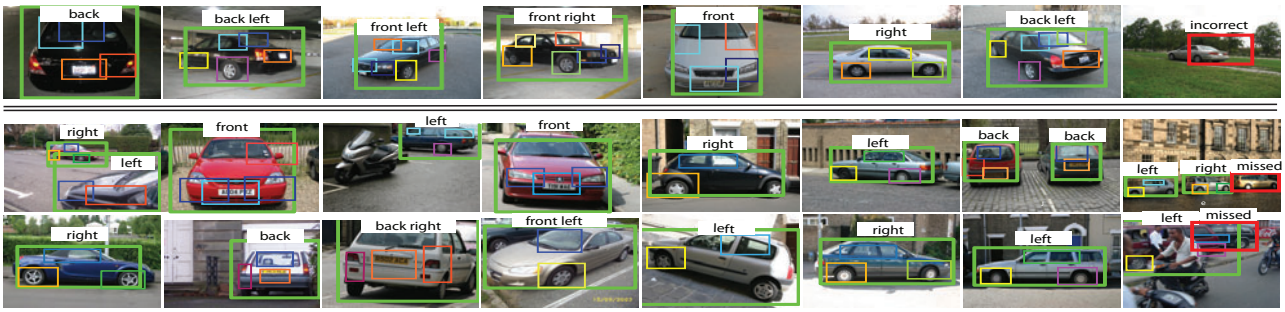


Figure 9. Sample viewpoint classification results. Proposed object detections are indicated by green bounding boxes. Text labels show the predicted viewpoints. Smaller rectangles illustrate the supporting parts learned before part expansion. The **top row** shows results from the [25] dataset; and the **bottom 2 rows** show results from the PASCAL VOC 2006 dataset. The last column shows examples of missed or incorrect detections.

cially encouraged to see the model's ability to perform satisfying viewpoint classification once the objects are detected. In the future, we plan to explore further the robustness and discriminability of the model.

Acknowledgments

The authors would like to thank Jia Deng, Barry Chai, and Bangpeng Yao for their help for data collection. The authors especially thank Juan Carlos Niebles for helpful comments and discussions. Hao Su is funded by China Scholarship Council. Li Fei-Fei is funded by a Microsoft Research fellowship and a Google award.

References

[1] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *CVPR*, 2005. 1

[2] D. M. Blei. Variational methods for the dirichlet process. In *In ICML*, 2004. 5

[3] M. Brown and D. Lowe. Unsupervised 3D object recognition and reconstruction in unordered datasets. In *3DIM05*, 2005. 1

[4] H. Chiu, L. Kaelbling, and T. Lozano-Perez. Virtual training for multi-view object class recognition. In *CVPR*, 2007. 2

[5] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *ECCV International Workshop on Statistical Learning in Computer Vision.*, Prague, 2004. 1

[6] M. Everingham, A. Zisserman, C. Williams, and L. V. Gool. The PASCAL Visual Object Classes Challenge 2006 Results. Technical Report, PASCAL Network, 2006. 1, 7

[7] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003. 1

[8] V. Ferrari, T. Tuytelaars, and L. Van Gool. Simultaneous object recognition and segmentation from single or multiple model views. *IJCV*, April 2006. 1

[9] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, Beijing, China, 2005. 1

[10] R. I, Hartley and A. Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge, UK, 2004. 4

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

[12] J. Koenderink and A. van Doorn. The singularities of the visual mappings. *Biological Cybernetics*, 24(1):51–59, 1976. 2

[13] K. Kurihara. Collapsed variational dirichlet process mixture models. In *IJCAI*, 2007. 5

[14] A. Kushal, C. Schmid, and J. Ponce. Flexible object models for category-level 3d object recognition. In *CVPR*, 2007. 2

[15] S. Lazebnik, C. Schmid, J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *CVPR*, June 2006. 6

[16] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *In ECCV workshop on statistical learning in computer vision*, 2004. 1, 6

[17] J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3d feature maps. *CVPR*, June 2008. 7

[18] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999. 6

[19] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002. 6

[20] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *WWW*, 2008. 4, 5

[21] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *IJCV*, 2002. 6

[22] R. M. Neal and G. E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998. 6

[23] P. Rockett. The accuracy of sub-pixel localisation in the canny edge detector. In *BMVC*, 1999. 6

[24] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *IJCV*, March 2006. 1

[25] S. Savarese and L. Fei-Fei. 3D generic object categorization, localization and pose estimation. ICCV, 2007. 1, 2, 7, 8

[26] S. Savarese and L. Fei-Fei. View synthesis for recognizing unseen poses of object classes. *ECCV*, 2008. 2

[27] H. Schneiderman and T. Kanade. A statistical approach to 3D object detection applied to faces and cars. In *CVPR*, 2000. 1

[28] J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994. 5

[29] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Describing visual scenes using transformed dirichlet processes. In *In NIPS*, pages 1297–1304. MIT Press, 2005. 3

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

[31] A. Torralba, K. Murphy, and W. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR*, 2004. 1

[32] S. Ullman and R. Basri. Recognition by linear combination of models. Technical report, Cambridge, MA, USA, 1989. 1

[33] M. Weber, W. Einhaeuser, M. Welling and P. Perona. Viewpoint-Invariant Learning and Detection of Human Heads. *ICAFGR*, 2000. 1

[34] P. Yan, D. Khan, and M. Shah. 3d model based object class detection in an arbitrary view. *ICCV*, 2007. 2

[35] M. Zuliani, C. S. Kenney, B. S. Manjunath. The multiRANSAC algorithm and its application to detect planar homographies. *ICIP*, Sept 2005. 6