A Compositional Approach to Learning Part-based Models of Objects

Roozbeh Mottaghi[1], Ananth Ranganathan[2], and Alan Yuille[3,1,4]
[1]Department of Computer Science, University of California, Los Angeles, USA, {roozbehm}@cs.ucla.edu
[2]Honda Research Institute USA, Inc., Mountain View, California, USA {aranganathan}@honda-ri.com
[3]Department of Statistics, University of California, Los Angeles, USA, {yuille}@stat.ucla.edu
[4]Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea

Abstract

We propose a new method for learning probabilistic part-based models of objects using only a limited number of positive examples. The parts correspond to HOG bundles*, which are groupings of HOG features. Each part model is supplemented by an appearance model, which captures the global appearance of the object by using bags of words of PHOW features. The learning is invariant to scaling and in-plane rotations of the object, the number of parts is learnt automatically, and multiple models can be learnt to allow for variations of 3D viewpoint or appearance. Through an experiment, we show that 3D multi-view object recognition can be performed by a series of learnt 2D models. The method is supervised but can learn models for multiple object viewpoints without these viewpoints being labeled in the training data. We evaluate our method on three benchmark datasets: (i) the ETHZ shape dataset, (ii) the INRIA horse dataset, and (iii) a multiple viewpoint car dataset. Our results on these datasets show proof of concept for our approach since they are superior or close to the state-of-the-art on all three datasets while we do not use any negative examples.*

1. Introduction

In this paper, we propose a new method for learning part-based models for object categories. These models are augmented by appearance-based models and then tested on single and multi-view detection tasks. To better model the variations in the appearance or 3D viewpoint of the objects, in contrast to the current part-based methods (such as [4] and [7]), we do not specify the number of object models or their constituent parts. These numbers and the parameters corresponding to the spatial relationship of the object parts are determined automatically by the learning method. We tackle the difficult problem of simultaneous structure and parameter learning by introducing a new compositional learning approach.

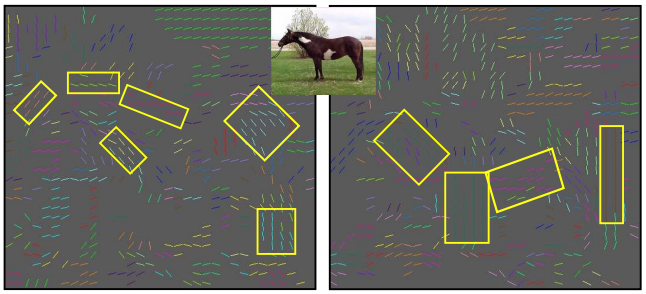

Figure 1. The part-model of the horse is illustrated. Parts are shown by yellow rectangles, which correspond to HOG-bundles. The HOG-bundles are computed by grouping neighboring HOG cells (indicated by similar colors) and shown in the left and right images that correspond to the gradients with left-to-right and right-to-left orientations. These gradients are described by two different parts of the HOG feature vector. The colored lines show the dominant orientation of each cell. **(Better viewed if zoomed in)**

Our compositional learning method proceeds by building part-based models by combining elementary components incrementally. This can be seen as a type of breadth first search in the space of object models. These part-based models are generative which enables our learning process to use model selection criteria. The learning only requires positive examples of each object category and is supervised in the sense that we know the label and bounding box of the objects during learning. However, we show the method has the capability of learning multiple 2D models to describe an object seen from different viewpoints without using any 3D viewpoint labels. The learning is also invariant to scale and in-plane rotation.

The object parts correspond to *HOG-bundles*, which are grouping of HOG features [5] and are computed from images in a pre-processing step. The HOG-bundles typically correspond to parts of the object (Figure 1), and so our object models can be used to parse the object into parts (though this is not the main goal of this paper). We also

use histograms of vector quantized PHOW features [1] for modeling the global appearance of the object. Overall, the HOG-bundles, supplemented by PHOW features, provide a rich and intuitive representation of the object.

Related work. One of the popular approaches in object detection is to learn a codebook of object features and use them later for recognition of new instances. Fergus et al. [8] propose an unsupervised generative model for configurations of the codebook words of objects. Leibe et al. [15] learn a shape model to specify where on an object a codebook entry may appear. Zhang et al. [22] investigate combining different detectors and descriptors with a classifier that can effectively use the information provided by the features. These features are sparse, so even if a generative model is learned, only a sparse set of object points can be generated.

Part-based models have proved successful on difficult object detection datasets recently. The Latent SVM work by Felzenszwalb et al. [7] or the work by Crandall et al. [4] are examples, where they describe a method of object detection using deformable part models. Kumar et al. [13] propose another approach to discriminative learning for part-based models. The advantage of our method to these methods is that we learn the number of parts (i.e. the graph structure) automatically while they use a specified number of parts. They also learn a fixed number of models for each object while our approach learns the number automatically, which allows better description of intra-class variability and is important for 3D recognition. Moreover, they use EM-type algorithms for learning, which requires good initialization, while our approach exploits compositionality to perform learning and requires no initialization.

This paper is most closely related to work which learns generative models of objects such as the compositional learning work by Zhu et al. [23] which is based on edge maps of objects or related work based on interest points [24]. But the method described here is based on different image features and uses a different (but related) search strategy. Work by Su et al. [18] describes a generative model for recognizing object classes and their 3D viewpoints but uses motion cues from video sequences (instead of static images). Lee et al. [14] also propose an unsupervised generative approach based on Deep Belief Networks, but this approach has not been tested on difficult datasets including significant scale and viewpoint changes.

This paper is organized as follows. Section 2 describes the image features that we use and, in particular, defines the HOG-bundles. Section 3 describes the part-based model and its inference algorithm. Section 4 describes how the part-based models are learnt. Section 5 introduces the appearance-based models and describes how these are combined with the part-based models. Section 6 describes the implementation and gives the results on the datasets that include single and multi-view images of objects.

2. Image Features – HOG Bundles and Bags of Words

Our object models use two types of image features which are extracted from the image in a pre-processing stage: Histograms of Oriented Gradients (HOGs) [5] and PHOW [1], which are used by the part-based and by the appearance-based models respectively.

An image $\mathbf{I}$ is represented by $\mathbf{I} = \{\mathbf{z}_i : i = 1, ..., N\}$ in terms of HOG-bundles described by $\mathbf{z} = (\mathbf{r}, \theta, \mathbf{f})$, where $\mathbf{r}$ is the position, θ is the orientation, and $\mathbf{f} = (f_h, f_w)$ is the height f_h and width f_w of the bundle. The number of HOG-bundles in an image is denoted by N. This can be supplemented – for the appearance-based models – by PHOW features $Ph(\mathbf{r})$ as a function of position $\mathbf{r}$.

The part-based models use HOG-bundles because they are robust to local variations in shape and image intensity. They also provide a richer description of the object than interest points, which are often used to capture the salient structures of the object. Moreover, compared to features such as the edge features used in [23], there are advantages to using HOG-bundles for learning because: (i) each HOG-bundle has distinguishing attributes (e.g., size, height and width), and (ii) there are only a few hundred of them in each image. The HOG features are computed following [7] where the orientations are quantized into 18 bins, resulting in a 31-dimensional feature vector for each HOG cell.

We compute *HOG-bundles*, illustrated in Figure 1, by grouping neighboring HOG cells which share similar properties using the following grouping rules: Two HOG cells are grouped if they are neighbors in the grid image and satisfy the following criteria:

- The difference between the feature vectors of two cells are small, as computed by the χ^2 distance function over the feature vectors.
- The orientation with the maximum magnitude should be similar for two HOG cells. Usually, the cells that belong to a part have a similar orientation.
- HOG cells with orientation in many directions will not be grouped, since they usually correspond to randomly textured areas such as grass. This is quantified by the squared difference $|\Omega - \omega|^2$, where Ω is the maximum magnitude of the orientation part of the feature vector, and ω is the mean of the magnitudes. However, we group the cells that correspond to uniform intensity regions (low-magnitude gradient cells).

To build the HOG-bundles, we start from an arbitrary cell in the image and check if its neighbors satisfy the grouping criteria. If they do so, we group them and check

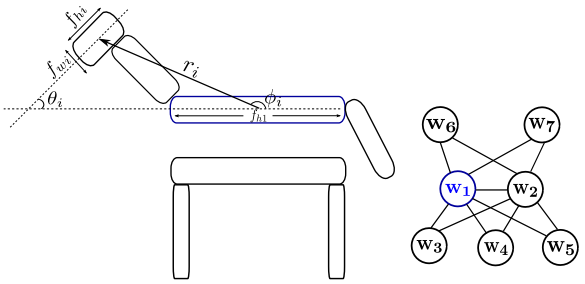

Figure 2. Left: the geometry of the part-based models illustrating $\theta, \phi, f_h, f_w, r$. The first reference part (corresponding to $\mathbf{w}_1$) is shown in blue. Right: the graphical model. All parts are connected to the two reference parts $\mathbf{w}_1, \mathbf{w}_2$. These reference parts are chosen by the learning algorithm.

for the neighbors of neighbors until all of the cells in the image are processed. Each HOG-bundle is approximated by a rectangle and has the following attributes: position of the center ($\mathbf{r}$), width (f_w), height (f_h), and orientation (θ), which is the mean of the orientations with the maximum magnitude in the histograms of the constituent cells.

The HOG-bundles can overlap because the grouping is performed based on two different parts of the HOG feature vector representing the gradients with left-to-right and right-to-left orientations as shown in Figure 1 (i.e. we consider dimensions 1 to 9 and 10 to 18, separately).

In addition to HOG-bundles that mainly capture the object contours and uniform gradient regions on the object, we use the PHOW features [1] (a variant of SIFT computed at multiple scales) in our model to capture regional appearance properties of the objects. We have adopted the implementation by [20]. These features are relatively invariant to local spatial and intensity changes. We denote these features as $Ph(\mathbf{r})$, and they are computed densely as a function of position $\mathbf{r}$. Within any image window, we can compute the histogram $\mathcal{H}(Ph(.))$ of the PHOW features using the standard clustering techniques.

3. Part-based object models and inference

This section describes the part-based object models and how we perform inference using them (the learning is described in the next section). Each object O will have several different models indexed by $\tau = 1, \ldots, T_O$. In multi-view recognition tasks, these models typically correspond to different views of the object. The number of these models is unknown and will be learnt automatically.

3.1. The Object Models

An Object O will be represented by a set (e.g. a mixture) of models $P(\mathbf{W}|O, \tau)$, where τ indexes the mixture component and $\mathbf{W}$ denotes the state variables defined below. For notational simplicity we ignore the indices O and τ in Section 3.1 and reintroduce them in Section 3.2. Each part-based model is a graphical model with state variables $\mathbf{W} = \{\mathbf{w}_i : i = 1, ..., M_O\}$, where $\mathbf{w}_i = (\mathbf{r}_i, \theta_i, \mathbf{f}_i)$ represents the position $\mathbf{r}_i$, the orientation θ_i, and the feature properties $\mathbf{f}_i$ of the i^{th} part. The feature properties can be decomposed into $\mathbf{f} = (f_w, f_h)$ where f_w and f_h describe the width and height of a HOG-bundle. Figure 2 visualizes these terms for an example bundle and also represents an example graphical model for the object. Our graphical model is similar to the 2-fan model of Crandall et al. [3] but as described in Section 4, our learning and inference procedures are quite different since we learn the graph structure and the number of models as well.

Probabilistic modeling of these part-based models requires the specification of a prior distribution on them and also a likelihood function. The prior probability is of form, see Figure 2 (right):

$$P(\mathbf{W}|\mathbf{\Lambda}) = P(\mathbf{w}_1)P(\mathbf{w}_2|\mathbf{w}_1, \boldsymbol{\lambda}_2)\prod_{i=3}^{M_O} P(\mathbf{w}_i|\mathbf{w}_1, \mathbf{w}_2, \boldsymbol{\lambda}_i), \tag{1}$$

where the model parameters $\mathbf{\Lambda} = (\boldsymbol{\lambda}_2, ..., \boldsymbol{\lambda}_{M_O})$, and the number of parts M_O will be learnt from the data, as described in Section 4. The form of the model enables efficient inference, invariant to scale and in-plane rotation, as discussed in Section 3.2.

We specify the probability distributions of Equation 1 as follows. First, we define a coordinate change $(\mathbf{r}_i - \mathbf{r}_1) = r_i(\cos\phi_i, \sin\phi_i)$ in radial coordinates based on the position $\mathbf{r}_1$ of the first part (the first reference part). We define $P(\mathbf{w}_1)$ to be the uniform distribution $U(\mathbf{w}_1)$. We also assume that the spatial and feature terms are independent:

$$\begin{aligned} P(\mathbf{w}_i|\mathbf{w}_1, \mathbf{w}_2) &= P(\mathbf{f}_i|f_{h1})P(\phi_i|\phi_1)P(\theta_i|\phi_1)P(r_i|\mathbf{r}_1, \mathbf{r}_2) \\ P(\mathbf{w}_2|\mathbf{w}_1) &= P(\mathbf{f}_2|f_{h1})P(\phi_2|\phi_1)P(\theta_2|\theta_1)P(r_2|\mathbf{r}_1), \end{aligned} \tag{2}$$

where f_{h1} represents the height of the bundle corresponding to the first reference part. It should be noted that the two reference parts are chosen by the learning algorithm automatically.

We specify these distributions in terms of Gaussian and uniform distributions, using the notation that $\mathcal{N}(\mu, \sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2. These distributions are chosen to ensure invariance to the scale of the features, the orientation of the object, and the scale of the object (features sizes and orientations are defined relative to those of the first part, and the relative positions are scaled by the distances between the first two parts).

$$\begin{aligned} &P(\mathbf{f}_i|f_{h1}) = \mathcal{N}(f_{h1}\boldsymbol{\mu}_i^{\mathbf{f}}, f_{h1}^2\sigma_{\mathbf{f}}^2),\ P(\phi_i|\phi_1) = \mathcal{N}(\mu_i^{\phi} + \phi_1, \sigma_{\phi}^2), \\ &P(\theta_i|\phi_1) = \mathcal{N}(\mu_i^{\theta} + \phi_1, \sigma_{\theta}^2),\ P(r_i|\mathbf{r}_1, \mathbf{r}_2) = \mathcal{N}(r_2\mu_i^r, r_2^2\sigma_r^2), \\ &P(r_2|\mathbf{r}_1) = U(r_2). \end{aligned} \tag{3}$$

The model parameters are the mean feature properties and angles $\{(\boldsymbol{\mu}_i^{\mathbf{f}}, \mu_i^{\phi}, \mu_i^{\theta}) : i = 2, ..., M_O\}$ and positions $\{\mu_i^r : i = 3, ..., M_O\}$. These are learnt from the training data. There are an additional four parameters which are fixed $\sigma_{\mathbf{f}}^2, \sigma_{\phi}^2, \sigma_{\theta}^2, \sigma_r^2$ (which will be learnt in future work).

The likelihood function assumes that the HOG-bundles $\{\mathbf{z}_i : i = 1, ..., N\}$ in the image are generated either from the object model $P(\mathbf{I}|\mathbf{W})$, or from a *background model* $P_B(\mathbf{z})$ which generates HOG-bundles independently. There is a default *background model* which assumes that $P_B(\mathbf{I}) = \prod_{i=1}^{N} P_B(\mathbf{z}_i)$, i.e. each HOG-bundle is generated independently by $P_B(.)$, where N is the number of HOG-bundles in the image. We define $P(\mathbf{I}|\mathbf{W}) = \prod_{i=1}^{M_O} \delta(\mathbf{z}_i, \mathbf{w}_i)$, where $\delta(\mathbf{z}, \mathbf{w}) = 1$ if $\mathbf{z} = \mathbf{w}$, and equal to zero otherwise i.e. the HOG-bundles generated by the object model have the same size, positions, and orientations as the state variables $\mathbf{w}$ of the object parts. For simplicity we set $P_B(.) = U(.)$ the uniform distribution. Hence the likelihood function for an image assumes that:

$$\begin{aligned} &\mathbf{z}_i : i = 1, ..., M_O \text{ sampled from } P(\mathbf{I}|\mathbf{W})P(\mathbf{W}) \\ &\mathbf{z}_i : i = M_O + 1, ..., N \text{ sampled from } P_B(\mathbf{z}). \end{aligned} \tag{4}$$

3.2. Inference

The inference task determines if there is an object in the image and, if so, where it is. Recall that each object O can have several different part-based models indexed by τ. These are expressed as $P(\mathbf{W}|O, \tau)$, which is of the form given by Equation 1 with parameters $\Lambda_{O,\tau}$ which depend on the object O and the object model τ. All part-based models for the same object have the same number of parts and the likelihood functions are the same for all part-based models.

There are two types of inference tasks. Firstly, to find whether the image contains an object, and if so, to determine the object and type. Secondly, to determine where the object is in the image, i.e. to determine its configuration $\mathbf{W}$.

Detecting the optimal configuration for each object and type O, τ requires solving:

$$\hat{\mathbf{W}}_{O,\tau} = \arg\max_{\mathbf{W}} P(\mathbf{I}|\mathbf{W})P(\mathbf{W}|O, \tau). \tag{5}$$

The form of the likelihood term (4) means that the state variables $\{\mathbf{w}_i : i = 1, ..., M_O\}$ can only take values $\{\mathbf{z}_i : i = 1, ..., N\}$ from the HOG-bundles computed from the image. The form of the part-based models in Equation 1, means that we can express Equation 5 as minimizing a function of form $E(\mathbf{w}_1, \mathbf{w}_2) + E(\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3) + ... + E(\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_{M_O})$. $E(.)$ is the negative logarithm of the probability, and so is a sum of quadratic terms if the distributions are Gaussian.

Inference can be performed in polynomial time using dynamic programming (DP). Our models are graphical models with a limited number of closed loops and so DP can be applied using a variant of the junction trees algorithm. Inference starts by evaluating all combinations for the first two (reference) parts. Because of the form of the potential terms, DP is very efficient for the remaining parts. For example, to find the fourth part, for each configuration $(\mathbf{w}_1, \mathbf{w}_2)$ of the two reference parts, we need to find only $\mathbf{w}_4$ minimizing $E(\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_4)$, which involves simple calculations. This enables us to rapidly find the lowest energy configurations of each model. Candidate object configurations are the ones whose overall energy is below a threshold.

We perform model selection to determine which object, if any, is present in the image, and its type. The object and its type (O_i, τ_i) are calculated such that:

$$P(\mathbf{I}|\hat{\mathbf{W}}_{O,\tau})P(\hat{\mathbf{W}}_{O,\tau}|O, \tau) > \prod_i P_B(\mathbf{z}_i), \tag{6}$$

Strictly speaking, model selection should sum over all possible configurations of the models but, in practice, the locations are strongly peaked so we replace the sum by the dominant term. In this paper, $P_B(.)$ is a constant so this model selection reduces to keeping model configurations for which the probability $P(\mathbf{I}|\mathbf{W})P(\mathbf{W}|O, \tau)$ lies above a threshold.

For each image, this gives a set of n models and types (O_i, τ_i) together with their configurations $\hat{\mathbf{W}}_{O_1,\tau_1}, ..., \hat{\mathbf{W}}_{O_n,\tau_n}$. These denote possible classifications of the object and possible detection locations for it (if $n = 0$, then no object is detected). These candidate models and configurations are then combined with the results of the appearance-based model, as described in Section 5. Note that the part-based models described here need to be supplemented by the additional appearance cues specified by the appearance-based models.

Our inference is robust against occlusions to some extent. For example, if a model has 8 parts and 6 parts are detected with low partial energy we report this as a detection of the object (despite the two missing parts). Alternatively, we could explicitly incorporate occlusion into our likelihood term similar to Fergus et al.'s [8].

4. Compositional Learning

The learning algorithm involves estimating the unknown (hidden) variables of the problem, namely, M_O (number of parts), T_O (number of models/viewpoints), the correspondence between the features $\mathbf{z}$ and the variables $\mathbf{w}$, and $\boldsymbol{\Lambda}_{O,\tau}$ (the parameters of the model). Our situation is more complex compared to the approaches that assume M_O and T_O are known and can apply the standard EM (e.g. [4]). Due to the complexity of the problem, EM will not work for our case, hence, we propose a novel compositional learning algorithm to tackle the problem.

Our strategy is to build models by searching through the space of possible models and applying a set of composition

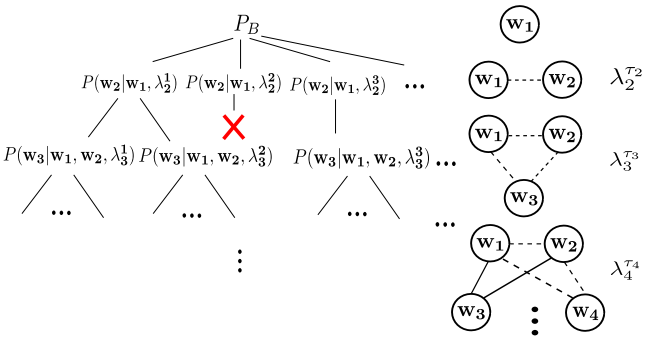

Figure 3. The compositional learning proceeds by adding an extra node to the model and estimating the parameters for the newly added part. $\boldsymbol{\lambda}_i^{\tau_i}$ denotes the parameters corresponding to the i^{th} part of the τ_i^{th} model. An example model that cannot be expanded, because it fails model selection or is pruned by the frequency criteria, is shown by the red cross.

rules to a *root* model. These compositional rules take n-part models and add an extra part to form $n+1$-part models. Model selection is used to check that the new models give better descriptions of the data. The procedure terminates automatically when this is not the case. This can be thought of as a breadth-first search over the space of all object models. This procedure is done for each object separately so we ignore the object label O, keeping only the type τ. The procedure is shown in Figure 3. The idea of learning using compositions is similar to [24] but they use a depth-first greedy search strategy.

The input is a set of images indexed by t where each image is represented by its HOG-bundles $\{\mathbf{z}_i^t : i = 1, ..., N_t\}$. We assume a default distribution for the images is specified by the background distribution: $P_B(\mathbf{I}^t) = \prod_{i=1}^{N_t} P_B(\mathbf{z}_i^t)$.

We first construct probability distributions defined for two-parts of form $P(\mathbf{w}_1, \mathbf{w}_2|\boldsymbol{\lambda}_2^{\tau_2})$, where τ_2 indexes these models. The parameters $\boldsymbol{\lambda}_2$ are determined by a clustering algorithm, described later in this section, which ensures that these models provide a better fit to the data. In a sufficient fraction η of images, this two-part model is matched (by the inference algorithm) to HOG-bundles $\mathbf{z}_i^t, \mathbf{z}_j^t$. The model selection requires that the two-part model, filling in the remaining features by the background model, gives a better fit than the pure background model – i.e.:

$$P(\mathbf{z}_j^t|\mathbf{z}_i^t, \boldsymbol{\lambda}_2^{\tau_2}) \prod_{k \neq j} P_B(\mathbf{z}_k^t) > \prod_i P_B(\mathbf{z}_i^t).$$
$$P(\mathbf{z}_j^t|\mathbf{z}_i^t, \boldsymbol{\lambda}_2^{\tau_2}) > P_B(\mathbf{z}_j^t). \quad (7)$$

This requirement gives us a family of two-part models indexed by τ_2 each of which is better for describing a sufficient fraction η of the images (as described in Equation 7) than the pure background model. We store these models and then grow them by adding extra parts. Each $n-1$ part model is specified by a set of parameters $\boldsymbol{\lambda}_2^{\tau_2}, ..., \boldsymbol{\lambda}_{n-1}^{\tau_{n-1}}$ and can be extended to an n-part model by specifying new parameters $\boldsymbol{\lambda}_n^{\tau_n}$, which determine the probability $P(\mathbf{w}_n|\mathbf{w}_1, \mathbf{w}_2)$ for the state of the n^{th} part in relation to the states of the first two parts $\mathbf{w}_1, \mathbf{w}_2$. We extend the model selection criterion from Equation 7 in the natural way to obtain the requirement that $P(\mathbf{z}_n^t|\mathbf{z}_i^t, \mathbf{z}_j^t, \boldsymbol{\lambda}_n^{\tau_n}) > P_B(\mathbf{z}_n^t)$ for a sufficient fraction η of images. Or in other words, the requirement that the n-part model gives a better description of a sufficient fraction of the images than the previous $n-1$-part model. We repeat until we fail to generate models with more parts. The learning procedure also stops in the case that at least one of the training images is not described by the newly created models. It should be noted that each n-part model is a child of an $n-1$-part model, and by construction the n-part model describes the data better than its parent (but does not necessarily give a better description than other $n-1$ part models).

Several grouping criteria can be used to estimate the parameters $\boldsymbol{\lambda}$. Equation 7 cannot be applied directly. Therefore, we have experimented with two clustering methods, the DBSCAN [6] and Affinity Propagation algorithms [11], which give roughly equal success and neither of them requires a pre-determined number of clusters. For example, DBSCAN performs a sequential search in feature space using clustering procedures equivalent to thresholding the left hand side of Equation 7 (provided they are Gaussian with fixed variance). The $\boldsymbol{\lambda}$ parameters correspond to the mean and variance of the clusters. After performing clustering, we can evaluate model selection in a validation step. The reported results in the experiments section are based on the Affinity Propagation method.

The learning procedure results in a set of part-based models for each object indexed by τ. For objects viewed from different viewpoints this includes models capturing each viewpoint. We prune the set of models based on two additional criteria: (i) remove those whose bounding box are significantly smaller than the bounding boxes of the training data, and (ii) eliminate those models which occur least frequently in the training images.

5. The appearance-based model

The part-based model described in Sections 3 and 4 is limited because it only uses appearance cues that can be represented by HOG-bundles. These correspond to dominant edges of the objects and, sometimes, regions with uniform gradients. Hence the models are generally poor at dealing with regional appearance cues.

In this section, we augment the part-based model with additional cues which are sensitive to regional properties of the objects. This corresponds to supplementing the HOG-bundles with the additional PHOW features, so that $\mathbf{I} = (\{\mathbf{z}_i\}, Ph(\mathbf{r}))$ where $Ph(\mathbf{r})$ are the PHOW features (refer to Section 2). This introduces a new *appearance*

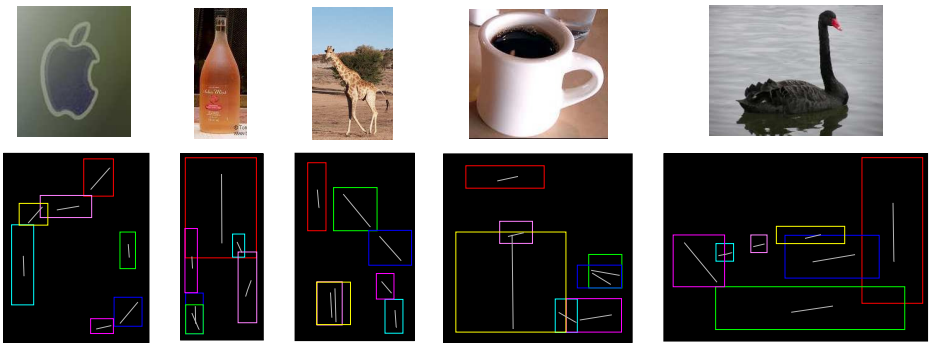

Figure 4. One of the learned models for each category of the ETHZ dataset is shown. The rectangles represent the HOG bundles and the line shows the dominant orientation of the HOG bundle. The number of parts and their relative position and orientation is determined by the learning procedure.

variable $\mathbf{w}_A$ for the model, which corresponds to the region occupied by the object. In addition, we add a new likelihood term, which couples the appearance variable to the histograms of PHOWs $\mathcal{H}(Ph(.))$ computed in the corresponding image region:

$$P(\mathcal{H}(Ph(.))|\mathbf{w}_A, O, \tau) = \frac{1}{Z} e^{-\min_a \mathcal{M}(\mathcal{H}(Ph(.)), \mathcal{H}_a^{O,\tau})}. \quad (8)$$

where $\mathcal{M}(.,.)$ is a measure of similarity between the histogram $\mathcal{H}(Ph(.))$ computed in the image region $\mathbf{w}_A$ and the histogram of one of several 'prototype histograms' $\mathcal{H}_a^{O,\tau}$ indexed by a for object and type O, τ. These prototypes are the histograms of the regions in the training images surrounded by object bounding boxes. The model chooses the nearest prototype using the min operation. We assume a default distribution $P(\mathcal{H}(Ph(.))$ to be uniform in regions where the object is not present. In this paper, we specify $\mathbf{w}_A(\mathbf{W})$ to be a deterministic function, e.g. bounding box, of the state variables $\mathbf{W}$ estimated from the part model.

During inference, we estimate $\hat{\mathbf{W}}_{O,\tau}$ for each object type O, τ by Equation 5. Then we compute $\mathbf{w}_A(\hat{\mathbf{W}}_{O,\tau})$ to obtain the position of the bounding box, followed by computing the overall fitness score for the object type by combining the contributions from the parts model and the appearance model.

This procedure for combining the appearance-model with the part-model is suboptimal: (i) the generative model is hand specified and not learnt from the data, (ii) the appearance variable $\mathbf{w}_A$ is a deterministic function of the part-based variables $\mathbf{W}$ but its relationship to them should be learnt, (iii) inference should estimate $\mathbf{W}$ and $\mathbf{w}_A$ together, instead of estimating $\mathbf{W}$ from the parts model and then computing $\mathbf{w}_A$, (iv) when combining the part and the appearance cues we should compute the normalization term.

Nevertheless, we obtain reasonable results using our current procedure. A method that addresses these issues (following [2], for e.g.) would presumably perform even better, and will be developed in future work.

6. Implementation and Results

In order to validate our method, we learned object models for the categories of the ETHZ dataset [10] and the INRIA horses [9]. Additionally, to show the performance of our method for 3D multi-view recognition, we applied our learning method to a multi-view car dataset [18], where the viewpoint labels were unknown to the learning method. After learning, the inference is performed in two stages. First, a set of candidate part configurations are obtained by thresholding the distributions of the part-based models (Equations 5 and 6). Then, we score each candidate bounding box with the appearance-based model (Equation 8). We plot the result curves by varying a threshold over these scores.

Initially, we show the performance of the method on single view datasets and then we show how the same learning procedure can be adopted to learn models for different 3D viewpoints of objects. ETHZ dataset consists of five object categories: Apple logos, Bottles, Giraffes, Mugs and Swans (255 images in total). The evaluation protocol for the ETHZ dataset is as follows. Half of the images of one category are used for learning and the other half and all of the images of the other categories are used for testing. Our part-based models are learned using only 6 training images. The appearance-based models use half of the images of one category (including those six used for part learning). We could use half of the training images for learning the part-based models as well but the learning becomes inefficient as the number of compositions of HOG-bundles grows. We do not use any negative examples and the reported results are obtained from 5 trials of random selection of training images. We also pruned the models that appeared in less than two-third of the training images (the parameter η in Section 4). One of the learned models for the ETHZ categories is shown in Figure 4.

We have compared our inference results quantitatively with the results of some of the recent methods [9, 10, 21, 16] in Figure 6. Also, we compare our results with the result of

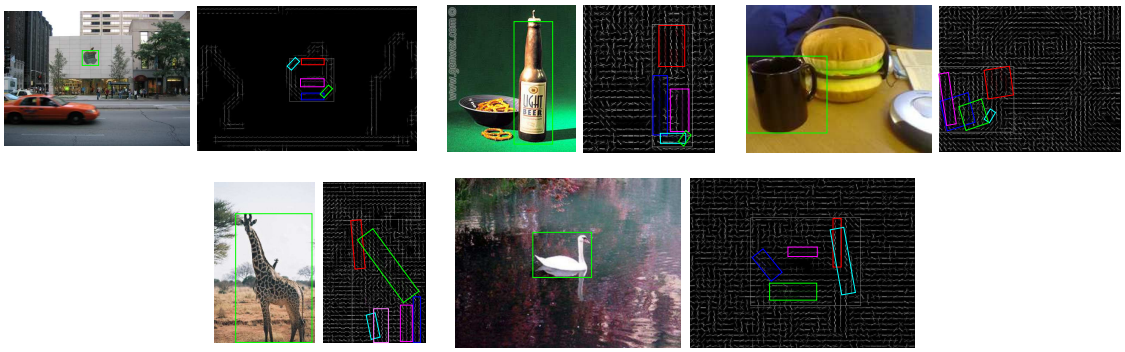

Figure 5. **(Better viewed if zoomed in)** Parsing and detection results of our inference method. Parts of the objects are shown with different colors. The number of detected parts are not necessarily the same as the number of parts in the full models. We have zoomed in the apple and swan inference images for better quality but the inference was performed on the full size image.

Felzenszwalb et al.'s code [7], which is obtained by [16]. Detection Rate versus False Positive per Image (FPPI) is used as the evaluation criteria. Our result is close to the others and we obtain the state-of-the-art results on Giraffe and Horse categories at 0.3/0.4 FPPI. The state-of-the-art for the horse category at 1.0 FPPI is obtained by [19] but their result is much worse than ours in the low FPPI regions. It should be noted that the dataset is small and an error in one image results in a significant drop in the curves. The reason that our method does not perform well on Mugs and Bottles categories is that their part-based models are usually rectangular and the rectangular structures are common in man-made environments. Also, the appearance-based model is not strong since the textures on the mugs and bottle labels vary greatly. Some examples of bounding box detection and object parsing are shown in Figure 5.

We now describe the multi-view recognition result that is obtained by learning multiple 2D models to describe different 3D viewpoints. The models are learned jointly and the viewpoint labels of the training images are not provided to the learning method. We provided our learning method with images of cars used by [18] which include images of 8 viewing angles, 3 scales, and 2 heights. Similar to [18] and [12], we use the first 5 instances for training (240 images, 24 of which were used for the part-based model), and the remaining 5 for testing (240 images). Unlike [18] which does not use the smallest scale, we perform the detection task on all of the scales. Our method outperforms [18] and [12] that use 3D cues in addition to the 2D cues. The state-of-the-art for this dataset is obtained by [17] but it is not fair to compare their method with ours as they use an approach based on 3D CAD models, which is highly specialized and is not purely image based. Their reported average precision is 89.9%. Our result together with two of the learned models representing two different viewpoints are shown in Figure 7. The η parameter for this experiment is 0.125.

Our method is not too sensitive to the scale of HOG cells and we only used 6×6 windows to compute the HOG features in all of our experiments. Our inference takes about a second on average for 10 models with 6 parts on a 200×150 image on a desktop with a 2.66 GHz CPU. So our method is faster than the state-of-the-art method on the ETHZ dataset ([16]) that takes about a few minutes for each image.

7. Conclusion

We proposed a novel approach for learning part-based models of objects. We also introduced *HOG-Bundles* as a novel representation for object parts and used them as the building blocks of our part-based model. The advantage of using HOG-bundles is that they are robust against local deformations of objects, and each image contains only a small number of them. We augmented our part-based models by a global appearance model based on PHOW features to model the regional properties of objects. The obtained results using these models are superior or close to the state-of-the-art even though we do not use any negative examples.

Our learning method learns the structure of the graphical model and its parameters simultaneously. The reason for estimating the number of object models and the number of parts from the training data is to provide a better description for intra-class variability and the changes in the viewpoint. The standard EM algorithm seems impractical in this situation, where the number of graph nodes and the number of models are hidden variables as well.

Our learning method is invariant to scale and in-plane rotations but to capture general 3D rotations, we need to provide enough training images of different viewpoints. We performed our experiments in a supervised setting, where we knew the object bounding box and object label during learning. However, by applying our method on a multi-view car dataset, we showed that the labeling constraint can be relaxed, and also the variation in the 3D viewpoint can be modeled by a set of 2D models.

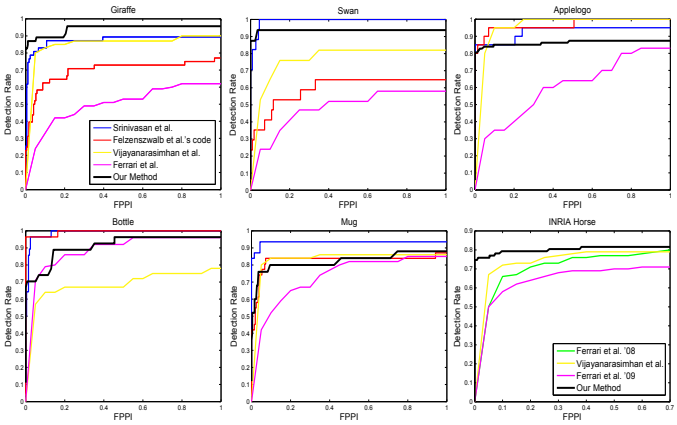

Figure 6. Quantitative comparison of our results with [16], [21],[7], [10] and [9]. The PASCAL criterion is used for the evaluation. Detection Rate versus False Positives per Image (FPPI) is shown. Unlike the other approaches, we do not use negative examples.

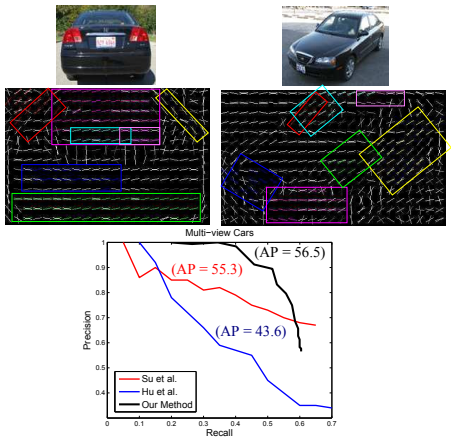

Figure 7. The curves show the result of our method (black line) compared to [12] and [18] in terms of precision-recall. Two of the automatically learned models representing two different viewpoints are also shown.

Acknowledgments

This work was supported by Honda Research Institute, the U.S. Office of Naval Research under the MURI grant N000141010933, and the Korean Ministry of Education, Science, and Technology, under the National Research Foundation WCU program R31-10008. We would also like to thank the National Science Foundation for support from grant NSF0917141.

References

[1] A. Bosch, A. Zisserman, and X. Munoz. Image classifcation using random forests and ferns. In *Proc. of the Intl. Conf. on Computer Vision (ICCV)*, 2007. 2, 3

[2] Y. Chen, L. Zhu, A. Yuille, and H. Zhang. Unsupervised learning of probabilistic object models (poms) for object classification, segmentation and recognition using knowledge propagation. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 31(10), 2009. 6

[3] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005. 3

[4] D. Crandall and D. Huttenlocher. Composite models of objects and scenes for category recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007. 1, 2, 4

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005. 1, 2

[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD)*, 1996. 5

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 32(9), 2010. 1, 2, 7, 8

[8] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2003. 2, 4

[9] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of adjacent contour segments for object detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 30(1):36–51, 2008. 6, 8

[10] V. Ferrari, F. Jurie, and C. Schmid. From images to shape models for object detection. *Intl. J. of Comp. Vision (IJCV)*, 87(3), 2009. 6, 8

[11] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(972–976), 2007. 5

[12] W. Hu and S.-C. Zhu. Learning a probabilistic model mixing 3d and 2d primitives for view invariant object recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010. 7, 8

[13] M. P. Kumar, A. Zisserman, and P. H. Torr. Efficient discriminative learning of parts-based models. In *Proc. of the Intl. Conf. on Computer Vision (ICCV)*, 2009. 2

[14] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proc. of the 26th Annual Intl. Conference on Machine Learning (ICML)*, pages 609–616, 2009. 2

[15] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *Intl. J. of Comp. Vision (IJCV)*, 77(1):259–289, 2008. 2

[16] P. Srinivasan, Q. Zhu, and J. Shi. Many-to-one contour matching for describing and discriminating object shape. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010. 6, 7, 8

[17] M. Stark, M. Goesele, and B. Schiele. Back to the future: Learning shape models from 3d cad data. In *British Machine Vision Conf. (BMVC)*, 2010. 7

[18] 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 *Proc. of the Intl. Conf. on Computer Vision (ICCV)*, 2009. 2, 6, 7, 8

[19] A. Toshev, B. Taskar, and K. Daniilidis. Object detection via boundary structure segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010. 7

[20] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algs. `http://www.vlfeat.org/`, 2008. 3

[21] S. Vijayanarasimhan and A. Kapoor. Visual recognition and detection under bounded computational resources. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010. 6, 8

[22] J. Zhang, M. Marszaek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories. *Intl. J. of Comp. Vision (IJCV)*, 73(2), 2007. 2

[23] L. Zhu, Y. Chen, A. Torralba, W. Freeman, and A. Yuille. Part and appearance sharing: Recursive compositional models for multi-view multi-object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010. 2

[24] L. Zhu, Y. Chen, and A. Yuille. Unsupervised learning of probabilistic grammar-markov models for object categories. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 31(1), 2009. 2, 5