Discriminative Mixture-of-Templates for Viewpoint Classification

Chunhui Gu[1] and Xiaofeng Ren[2]

[1] University of California at Berkeley, Berkeley, CA 94720, USA
chunhui@eecs.berkeley.edu
[2] Intel Labs Seattle, 1100 NE 45th Street, Seattle, WA 98105, USA
xiaofeng.ren@intel.com

Abstract. Object viewpoint classification aims at predicting an approximate 3D pose of objects in a scene and is receiving increasing attention. State-of-the-art approaches to viewpoint classification use generative models to capture relations between object parts. In this work we propose to use a mixture of holistic templates (e.g. HOG) and discriminative learning for joint viewpoint classification and category detection. Inspired by the work of Felzenszwalb et al 2009, we discriminatively train multiple components simultaneously for each object category. A large number of components are learned in the mixture and they are associated with canonical viewpoints of the object through different levels of supervision, being fully supervised, semi-supervised, or unsupervised. We show that discriminative learning is capable of producing mixture components that directly provide robust viewpoint classification, significantly outperforming the state of the art: we improve the viewpoint accuracy on the Savarese et al 3D Object database from 57% to **74**%, and that on the VOC 2006 car database from 73% to **86**%. In addition, the mixture-of-templates approach to object viewpoint/pose has a natural extension to the continuous case by discriminatively learning a linear appearance model locally at each discrete view. We evaluate continuous viewpoint estimation on a dataset of everyday objects collected using IMUs for groundtruth annotation: our mixture model shows great promise comparing to a number of baselines including discrete nearest neighbor and linear regression.

1 Introduction

One fundamental property of visual sensing is that it is a projection process from a 3D world to a 2D image plane; much of the 3D information is lost in the projection. How to model and re-capture the 3D information from 2D views has been at the center of the computer vision research. One classical example is the *aspect graphs* of Koenderink and van Doorn [1], where a 3D object is modeled as a collection of inter-connected 2D views.

A complete understanding of objects in a visual scene comprises not only labeling the identities of objects but also knowing their poses in 3D. Most of the recent vision research has been devoted to the recognition problem, where

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part V, LNCS 6315, pp. 408–421, 2010.

huge progresses have been made: the SIFT matching framework [2] and the HOG models [3,4] are good representatives of how much object recognition capabilities have progressed over the years. The 3D object pose problem have received much less but still considerable attention. The series of work from Savarese and Fei-Fei [5,6,7] are good examples of how people approach the 3D pose problem in modern contexts, where large benchmarks are established and evaluated for discrete viewpoint classification [5,8].

There have been, however, divergent trends between object recognition and pose estimation. Latest progresses in object recognition employ discriminative templates directly trained from image gradients [4]; latest 3D pose models group features into parts and learn generative models of their relationships [6,7].

We believe the two problems should be one and identical, that a good framework of object detection should be able to handle both category and viewpoint classification. In particular, discriminative learning, which has seen great successes in category classification, should readily apply to viewpoint classification.

In this work we present strong empirical proof that it is indeed the case: a discriminatively learned mixture of templates, extending the latent HOG framework of Felzenszwalb et al [4], is capable of representing a large number of viewpoints (as components) and handling both category and viewpoint classification. A mixture-of-HOG model produces superior results for all the three cases of supervised (with groundtruth view labels), semi-supervised (with a subset of view labels) and unsupervised (no view labels) viewpoint learning (see Figure 1). Furthermore, the mixture-of-templates approach has a natural extension to the continuous case: we propose a continuous viewpoint model which linearly approximates local appearance variations at each discrete view. This model is discriminatively trained, just as in the discrete case, and outputs a continuous 3D viewpoint/pose.

We evaluate our approach on a number of 3D object databases, including the 3DObject Database of Savarese [5], the VOC2006 car database [8], and a dataset of our own for benchmarking continuous viewpoint estimation. We

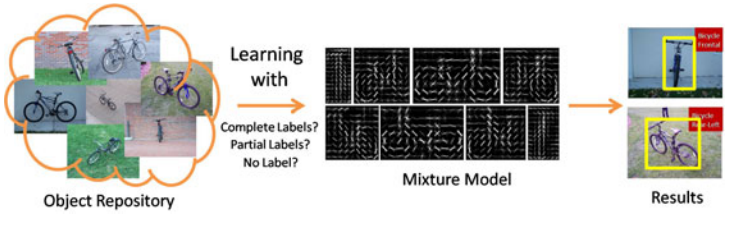


Fig. 1. We propose to use a discriminative mixture of templates for object viewpoint classification. We discriminatively learn a large mixture of templates using HOG [3,4] and show that the templates correspond well to the canonical views of an object, which are directly used for viewpoint classification and significantly outperform the state of the art. We show that the mixture model works well when trained with complete viewpoint labels (supervised), a subset of labels (semi-supervised), and no viewpoint labels (unsupervised). We then extend the mixture model for continuous pose prediction, again using a discriminative mixture of templates.

show that we significantly outperform the state-of-the-art results on all these challenging benchmarks: we improve the 8-way viewpoint classification accuracy on the 3DObject database from 57% to 74%, and that on the VOC2006 cars from 73% to 85%. For the continuous case, we show that our discriminative mixture model outperforms a number of baselines, including one using the closest discrete viewpoint and one using linear regression on top of the viewpoints.

2 Related Work

Understanding 3D objects and scenes from 2D views is the fundamental task of computer vision. In the early days vision researchers paid close attention to the 2D-to-3D correspondence, but many approaches were line-based and had many difficulties dealing with real-life images. The aspect graph of [1] presents a theory for modeling 3D objects with a set of inter-connected 2D views. This theory has a sound psychological foundation (e.g. [16]) and has been very influential and underlies most approaches to 3D object recognition.

Estimating the 3D pose of objects is a classical problem, and many solutions have been developed using either local features (e.g. [17]) or shape outlines (e.g. [18]), usually assuming perfect knowledge of the object. With the maturation of local feature detection (as in SIFT and its variants), latest progresses on pose estimation have mostly been local-feature based (e.g. [19,20]) and performed fairly well on instances of objects, preferably with texture.

There has been an increasing interest lately in 3D object pose classification, which aims at predicting a discrete set of viewpoints. A variety of approaches have been explored (e.g. silhouette matching [10] or implicit shape models [9] or virtual-training [13]). At the same time, many works on category-level classification also address the issue of multiple views (e.g. [21,14]).

The series of work from Savarese and Fei-Fei [5,6,7] directly address the problem of 3D viewpoint classification at the category and are the most relevant for us. They have developed a number of frameworks for 3D viewpoints, most adopting the strategy of grouping local features into parts and learning about their relations. Similar approaches have been adopted in a number of other works (e.g. [12,22]) that show promising results. The 3DObject dataset of Savarese et al [5] is a standard benchmark for viewpoint classification and has a systematic collection of object views. A number of categories from the PASCAL challenge [8], such as cars, are also annotated with viewpoints. We quantitatively evaluate our approach on these datasets.

The most recent progress in object recognition sees the use of discriminatively trained templates [3,4,23]. These techniques have been shown to perform very well on real-life cluttered images. In particular, the work of [4] presents a way to train mixture-of-components for object detection, and they illustrated the procedure with two components on cars and pedestrians. The context-based discriminative clustering work of [24] is similar in spirit. Our work is based on the mixture-of-HOG approach but focuses on viewpoints instead of categories. We explicitly handle viewpoints and train HOG models with a large number

of viewpoints/components. We also develop approaches for semi-supervised and unsupervised learning of viewpoints, and extend the discrete viewpoint model to the continuous case.

3 Discrete Viewpoint Models

In this scheme, given an example object, the models for each category return a confidence score of the object being in that category as well as a *discrete* viewpoint label associated with a canonical pose of that category. In many cases, such poses have semantic meanings, for instance, the frontal/side views of a car. We design each of these models as a mixture of HOG-based templates corresponding to multiple canonical poses of the category. We formulate the score function of example x as

$$S_{\mathbf{w}}(x) = \max_{v \in \mathcal{V}} \ \langle w_v, \ \psi_v(x) \rangle = \max_{v \in \mathcal{V}} \ w_v^T \psi_v(x) \tag{1}$$

where $\mathbf{w} = \{w_1, w_2, \ldots, w_V\}$ are the learned mixture of templates, $\mathcal{V} = \{1, 2, \ldots, V\}$, V is the number of canonical viewpoints in the model, and $\psi_v(x)$ is the feature representation of x under viewpoint label v. Since the dimensions of templates can be different, $\psi_v(x)$ is designed to match the dimension of w_v.

Accordingly, the predicted viewpoint label of x is

$$\tilde{v}_d(x) = \arg\max_{v \in \mathcal{V}} \ w_v^T \psi_v(x) \tag{2}$$

where the subscript d indicates a *discrete* label.

Features: We are in favor of HOG-based features because they encode spatial layout of object shape and handle well with intra-class and intra-viewpoint variations. We use the implementation of [4] for feature construction and normalization.

Detection: We adopt the standard framework of multi-scale window scanning for localizing objects in the image. The windows whose scores are higher than a learned threshold are picked as candidate detections, and non-max suppression is applied as postprocessing to remove redundant window detections.

3.1 Training

We extend the training algorithm of [4] to cope with viewpoint classification. For each category, we learn a mixture of V-component templates $\mathbf{w} = \{w_1, w_2, \ldots, w_V\}$ from a set of positive and negative training examples denoted by $\{x_1, x_2, \ldots, x_P\}$ and $\{z_1, z_2, \ldots, z_N\}$. Our learning framework attempts to "match" every positive example with at least one of these templates, and every negative example with none of the templates. Mathematically, the large margin optimization of this scheme is formulated as

$$(\mathbf{w}^*, \ \lambda^*) = \arg\min_{\mathbf{w}, \lambda} \sum_{v=1}^{V} \left\{ \frac{1}{2} \|w_v\|^2 + C_{Neg} \sum_{n=1}^{N} l\left(-w_v^T \psi_v(z_n)\right) + C_{Pos} \sum_{p=1}^{P} \lambda_v^p \cdot l\left(w_v^T \psi_v(x_p)\right) \right\} \tag{3}$$

subject to $\lambda_v^p \in \{0,1\}$ and $\sum_{v=1}^{V} \lambda_v^p = 1,\ \forall\ p = 1,\ldots,P$. Here, λ are binary component labels. $l(s) = \max(0, 1-s)$ is the hinge-loss function. C_{Pos} and C_{Neg} control the relative weights of the regularization term.

Our training procedure is directly based on [4]: each template w_v is initialized through a set of positive examples initially labeled as viewpoint v. In each iteration, all templates are updated simultaneously through data-mining hard negative examples and updating viewpoint labels λ of positive examples.

In [4], λ are considered as latent variables and thus the cost function does not enforce λ to match their true values. Here, we solve a more general problem which includes the scenarios when λ are partially or completely unknown. Furthermore, model initialization in [4] is solely based on aspect ratio; it is not designed for general viewpoint modeling and thus far from optimal for our problem. We will show that a carefully designed initialization is necessary to learn reasonable templates for canonical viewpoints.

Denote $\{v_d(x_1), v_d(x_2), \ldots, v_d(x_P)\}$ as the groundtruth viewpoint labels of the positive examples. In the following, we consider three scenarios, where these labels are completely known, partially known, and unknown. We name them supervised, semi-supervised, and unsupervised cases, respectively.

3.2 Supervised Case

In the supervised case, each $\lambda_v^p = \mathbf{1}\left[v = v_d(x_p)\right]$ is fixed. The model is initialized by partitioning the positive examples into groups based on the viewpoint labels and learn one viewpoint template from each group. In the model update step, the optimization is reduced to a linear SVM formulation.

We note that although we do not change component labels during the training process, this is different from training each component independently, as the training process uses a single regularization constraint and enforces the margin on all the clusters simultaneously. This has proved to be critical in learning mixture models that are balanced and accurate for viewpoint prediction.

3.3 Semi-supervised Case

In the semi-supervised case, we first build a multi-viewpoint classifier using the positive examples that have known viewpoint labels. In practice, we use the libsvm multi-class classification toolbox[25] on the HOG features. Once the rest of the positive examples are classified, we initialize component templates based on either known or estimated labels. In the model update step, we fix the labels for those who have known viewpoint labels, and allow the others to change.

3.4 Unsupervised Case

In the unsupervised case, model initialization is crucial for accurate viewpoint classification, because no explicit constraint in the later stage of optimization is imposed on the viewpoint labels. [4] partitions positive examples into component groups based on a simple aspect ratio criterion. We use a Normalized Cut-based

clustering scheme for initialization. We define an appearance distance between two positive examples x_i and x_j as

$$d(x_i, x_j) = \alpha \cdot \chi^2(\psi_0(x_i), \psi_0(x_j)) + (1 - \alpha) \cdot \|\text{Asp}(x_i) - \text{Asp}(x_j)\|_2 \quad (4)$$

where $\psi_0(x_i)$ is the HOG descriptor of x_i under a standard template size, and $\text{Asp}(x_i)$ is the normalized aspect ratio of the bounding box of x_i. Next, we convert the distances into affinity measurements using the exponential function and obtain the component groups by applying the Normalized Cut[26] algorithm on the resulting affinity matrix. This provides us with relatively even partitionings on the positive examples, which is important for good unsupervised performance.

In the model update step, since Eqn. 3 describes an integer-based non-convex problem([24], [27]), one tractable solution is to iterate between optimizing $\mathbf{w}$ given fixed labels λ and optimizing λ given fixed template weights $\mathbf{w}$. The former is an SVM and the latter optimization step is simply

$$\lambda_v^p = \mathbf{1}[v = \arg\max_s \left(w_s^T x_p\right)] \; \forall p = 1, \ldots, P \quad (5)$$

4 Continuous Viewpoint Models

In the continuous viewpoint case, we are interested in estimating the real-valued *continuous* viewpoint angles of an example object in 3D, denoted by $\theta \in \mathbb{R}^3$, which uses the angle-axis representation. We assume that the camera projection of the object is orthographic so that given a fixed orientation θ, the appearance of the object only changes in scale.

To obtain θ for a test object x, we modify the mixture model in the discrete viewpoint case and reformulate the score function as

$$S_{\mathbf{w}}(x) = \max_{v \in \mathcal{V}, \Delta\theta} f(v, \Delta\theta) = \max_{v \in \mathcal{V}, \Delta\theta} \left(w_v + g_v \Delta\theta\right)^T \psi_v(x) - d(\Delta\theta) \quad (6)$$

$$\theta(x) = \theta_{v^*} + \Delta\theta^* \quad (7)$$

where $\mathbf{w} = \{w_v\}$ and $\psi_v(x)$ are the same as before. g_v are the "gradients" of the template w_v over θ at discrete viewpoint v. $\Delta\theta$ are the offset viewpoint angles of x with respect to the canonical viewpoint angles θ_v. $d(\cdot)$ is a quadratic loss function that confines $\theta(x)$ to be close to θ_v. Denote $\Delta\theta$ by their elements $[\Delta\theta_1, \Delta\theta_2, \Delta\theta_3]^T$, then $d(\Delta\theta) = \sum_{i=1}^3 d_{i1}\Delta\theta_i + d_{i2}\Delta\theta_i^2$. In Eqn. (7), v^* and $\Delta\theta^*$ are obtained when the score function reaches its maximum. The variables w_v, g_v, θ_v and d_{i1}, d_{i2} are learned from training data.

This continuous viewpoint model can be interpreted as follows: we partition the continuous viewpoint space into small chunks where each chunk has a canonical viewpoint. For every viewpoint in the same chunk, we approximate its template as a linear deformation of the canonical template with respect to the difference of viewpoint angles from the canonical angles. We show that in practice, this approximation is reasonable when the chunk size is relatively small, and the model produces viewpoint classification performance superior to a number of baseline methods.

Detection: The multi-scale window scanning is again applied for localizing objects in the image. To find optimal v and $\Delta\theta$ in Eqn. (6) at a given location, we first maximize $\Delta\theta$ over any fixed v

$$\frac{\partial f(v,\Delta\theta)}{\partial \Delta\theta_i} = g_v(i)^T \psi_v(x) - d_{i1} - 2d_{i2}\Delta\theta_i = 0 \tag{8}$$

Hence, we obtain

$$\Delta\theta_i(v) = \left(g_v(i)^T \psi_v(x) - d_{i1}\right)/2d_{i2} \tag{9}$$

where $g_v(i)$ is the i'th column of g_v. Next, we enumerate over the discrete variable v with $\Delta\theta_i(v)$ and pick the pair with maximal score $S_{\mathbf{w}}(x)$.

4.1 Training

In training, for positive examples $\{x_1, x_2, \ldots, x_P\}$, their continuous viewpoint groundtruth labels $\{\theta_1, \theta_2, \ldots, \theta_P\}$ are given. Therefore, we rewrite the score function in Eqn. (6) as

$$f(v,\Delta\theta) = \left(w_v + g_v\Delta\theta\right)^T \psi_v(x) - d(\Delta\theta) \tag{10}$$

$$= \tilde{w_v}^T \tilde{\psi_v}(x) \tag{11}$$

where

$$\tilde{w_v} = [w_v,\ g_v(1),\ g_v(2),\ g_v(3),\ d_{11},\ d_{12},\ d_{21},\ d_{22},\ d_{31},\ d_{32}]$$
$$\tilde{\psi_v}(x) = [\psi_v, \Delta\theta_1\psi_v, \Delta\theta_2\psi_v, \Delta\theta_3\psi_v, -\Delta\theta_1, -\Delta\theta_1^2, -\Delta\theta_2, -\Delta\theta_2^2, -\Delta\theta_3, -\Delta\theta_3^2]$$

If all canonical viewpoint templates θ_v are known, $\psi_v(x)$ are completely observable and we can substitute $\tilde{w_v}$ and $\tilde{\psi_v}(x)$ for w_v and $\psi_v(x)$ in the training framework of the discrete viewpoint case. Now, θ_v are unknown, but we can initialize them from initial partitions of positive data (clustering on θ) and update them in each training iteration based on maximizing the cost function.

5 Experimental Evaluation: Discrete Viewpoints

For discrete viewpoint classification, we evaluate our proposed models on two standard and challenging databases: the 3DObject[5] and the VOC2006 cars[8]. The 3DObject dataset consists of 10 categories and 8 discrete viewpoint annotations for each category. We exclude the head and the monitor categories as they are not evaluated in previous work. Quantitative results on viewpoint and category classification are evaluated by means of confusion matrix diagonals, and averaged by 5-fold training/test partitions. On the other hand, the VOC2006 car database consists of 469 car objects that have viewpoint labels (frontal, rear, left and right). In the experiments we only use these labeled images to train mixture viewpoint models, with the standard training/test partition. The detection performance is evaluated through precision-recall curve. For both databases, we try our best to compare with previous works that have the same complete set of evaluations.

In the following sub-sections, we analyze our results in three different levels of supervision on the training data: *supervised*, *semi-supervised*, *unsupervised*.

Table 1. Supervised Case: viewpoint and category classification results (quantified by averages of confusion matrix diagonals). For category detection performance on the VOC2006 cars, we compare the precision-recall curves with [7] in Figure 2(d).

Database	3DObject		VOC2006 cars		
Method	[5]	Ours	[6]	[7]	Ours
Viewpoint	57.2%	**74.2 ± 0.9**%	57.5%	73.0%	**85.7**%
Category	75.7%	**85.3 ± 0.8**%	-	-	-

5.1 Supervised Case

Table 1 summarizes the viewpoint and category classification results when the viewpoint labels of the positive training data are known. We significantly outperform [5], the state of the art on the 3DObject database, in both viewpoint and category classification. We also show a significantly higher (4-view) viewpoint classification rate on the VOC2006 car database compared to the earlier work of [6] and [7]. Figure 2 shows a close look of our results.

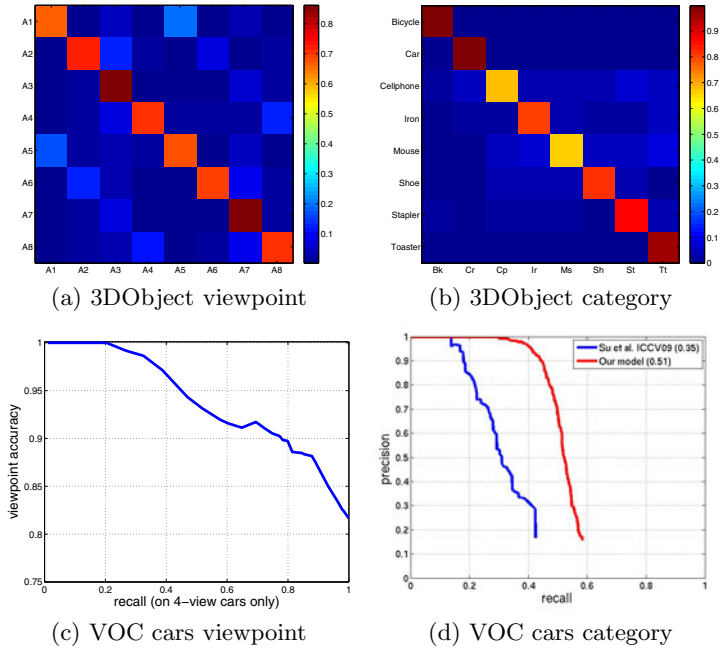
Note that in (a), the main viewpoint confusion pairs in 3DObject are those off by 180 degrees, for example, frontal vs. rear or left vs. right views. Category confusion matrix is shown in (b). (c) illustrates the change of viewpoint classification rate with object recall in VOC2006 cars. The curve suggests that the viewpoint classification accuracy increases with lower recall (and thus higher precision/category detection). (d) compares the precision-recall curves of [7] with ours. Note that even our car mixture model only covers 4 views, it still produces superior performance comparing to [7] in detection.

5.2 Semi-supervised Case

In the semi-supervised case, we are interested in knowing how much partial information from positive training data is "sufficient" to build a reasonable viewpoint model. Figure 3 (a, b) illustrate the viewpoint and category classification accuracies with changes in the proportion of training data having discrete viewpoint annotations. Zero proportion means no annotation which corresponds to the unsupervised case, whereas "proportion equals one" is the case of being totally supervised. Note that the accuracy numbers here are evaluated on the whole test set, not the set including only correct category prediction. We notice that even a small proportion (30% in the 3DObject) of annotated data significantly improves the viewpoint classification performance, while the category classification performance remains roughly constant with change of the number of annotated data. (We do not show the curve of category classification on the VOC2006 cars as it is a detection task.)

5.3 Unsupervised Case

Evaluation Methodology. The upper half of Table 2 compares three model initialization schemes in terms of the viewpoint and category accuracies. We



(a) 3DObject viewpoint (b) 3DObject category

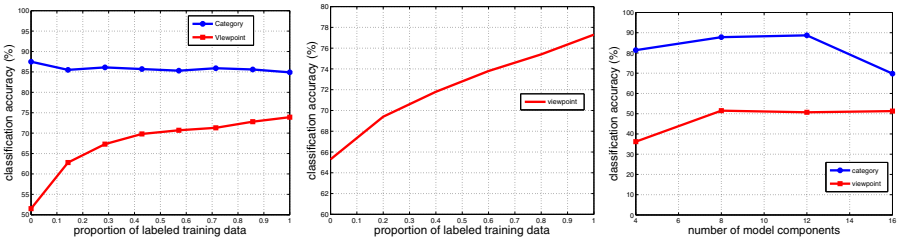
(c) VOC cars viewpoint (d) VOC cars category

Fig. 2. Supervised Case: viewpoint labels are all known for positive examples. **(a) (b)** Average confusion matrices of the viewpoint and category classifications in the 3DObject. **(c)** Viewpoint classification accuracy as a function of the object recall in the VOC cars. **(d)** Precision-recall curves of car detection. Note that our car model only trains on the 4-view cars and tests on the whole test images.

note that our proposed N-cut framework significantly outperformed the aspect ratio criterion by [4] for viewpoint classification. We also compute how far we can reach by computing an "upper bound" performance using the ground truth viewpoint labels of training data in initialization, shown in the third column of the first two databases. We see that the N-cut produces results close to and sometimes even better than the "upper bounds".

We quantitatively evaluate the quality of viewpoint clustering using the following statistics: *purity*, *normalized mutual information*, *rank index*, and *F measure*[28], shown in the bottom half of Table 2. All measurements of these statistics exhibit consistent behavior as the basic evaluation.

Number of Model Components. The number of components V in the unsupervised model is pre-determined. As a result, we are interested in knowing the impact of this parameter on the viewpoint and category classification performance. Figure 3(c) shows both accuracies with V on the 3DObject database. Note that for viewpoint classification, the accuracy undoubtedly breaks down when V is deficient (4) to explain the variety of data in viewpoint (8). It is, however, surprisingly insensitive to V when it gets large. On the other hand, for



(a) 3DObject Semi-supervised (b) VOC cars Semi-supervised (c) 3DObject Unsupervised

Fig. 3. Semi-supervised/Unsupervised Cases: viewpoint and category classification accuracies as a function of either the proportion of positive training data with viewpoint annotations (semi-supervised) or the number of model components/templates (unsupervised). **(a)**: Semi-supervised model on the 3DObject. **(b)**: Semi-supervised model on the VOC2006 cars. For these semi-supervised cases, the category detection performance is robust and largely independent of the availability of viewpoint labels. Viewpoint classification is robust up to about 30% of labeling. **(c)**: Unsupervised model on the 3DObject dataset.

Table 2. Unsupervised Case: viewpoint and category classification accuracies as well as four viewpoint clustering measurements[28] on two databases. We show comparison of 3 model initialization schemes ([4], N-cut, and Labels) on the 3DObject and VOC2006 cars. Note that [4] performs poorly in viewpoint classification. The "N-cut", proposed in this paper where the numbers are bolded, produces significantly better results than [4]. The "Labels" case uses the ground truth viewpoint labels to initialize models, which are considered to produce the "upper-bound" results.

Database	3DObject			VOC2006 cars		
Method	[4]	N-cut	Labels	[4]	N-cut	Labels
Viewpoint	40.2%	**51.5%**	63.4%	47.0%	**65.6%**	65.3%
Category	86.5%	**87.8%**	87.2%	-	-	-
Purity	0.42	**0.53**	0.65	0.58	**0.77**	0.76
NMI	0.43	**0.55**	0.61	0.41	**0.52**	0.50
Rank Index	0.77	**0.83**	0.86	0.71	**0.80**	0.80
F Measure	0.36	**0.45**	0.54	0.61	**0.68**	0.67

category classification, the accuracy breaks down when V is large, and insensitive with small V.

6 Experimental Evaluation: Continuous Viewpoints

For continuous viewpoint estimation, there is no standard benchmark database available, partly because it is considerably harder to establish groundtruth data to cover arbitrary 3D rotations. [19] uses a selected set of translations and rotations for (continuous) pose estimation. [29] does not use groundtruth but

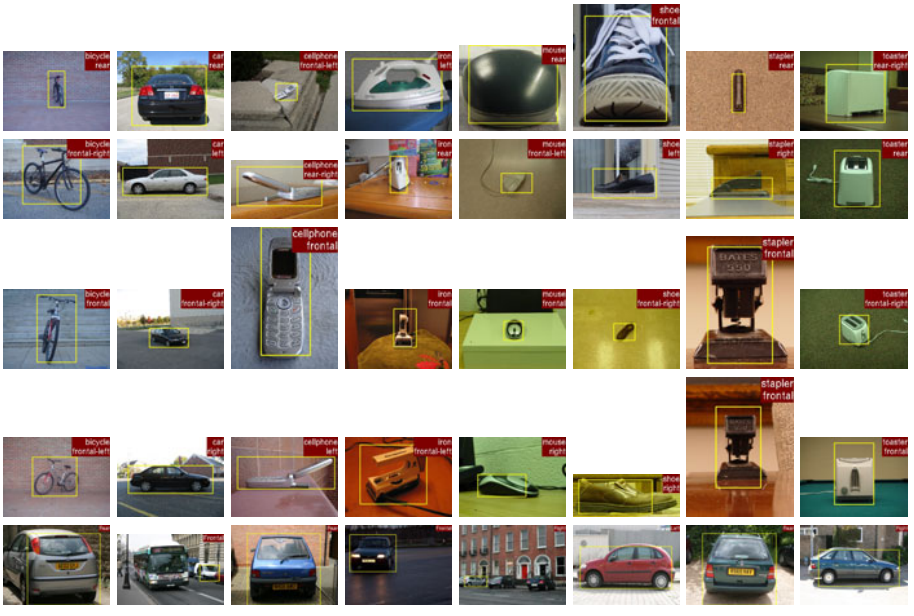


Fig. 4. Discrete viewpoint classification and category detection results. The yellow bounding boxes indicate object detection, and the labels on the upper-right corners show the predicted object category and viewpoint. The top 4 rows show results from the 3D object category database, and the bottom row shows results from the PASCAL VOC 2006 car database.

compare results using artificially distorted images. In the case of [30], rotation is limited to in-plane rotation on the ground.

We believe that a good database with full 3D pose groundtruth is crucial for the advances of pose estimation techniques, and we set to collect a 3D pose database using commercially available IMUs: we use the Microstrain 3DM-GX1 sensors and attach it to a PrimeSense video camera. The Microstrain provides gyro-stabilized full 3D orientation at about 80Hz, and the camera records 640x480 frames at 30Hz. The two streams are aligned manually.

We collect a continuous object pose database covering 17 daily objects with a variety of shape, appearance and scale (Fig 5(a)). We put each object on a turning table, let the object turn, while hand-holding the camera/IMU pair and moving it at varying heights and orientations. We typically let each object rotate for 4-5 circles and take about 2K video frames total. In our evaluation experiments, we use all 17 objects and about 1K frames for each object. Frames are evenly and randomly partitioned for training and testing. Object masks are computed from background subtraction and are used to find bounding boxes.

We compare our continuous viewpoint model with two baseline methods. The first one employs a nearest neighbor scheme. Each test example is assigned the same continuous viewpoint label as that of the example's closest mixture

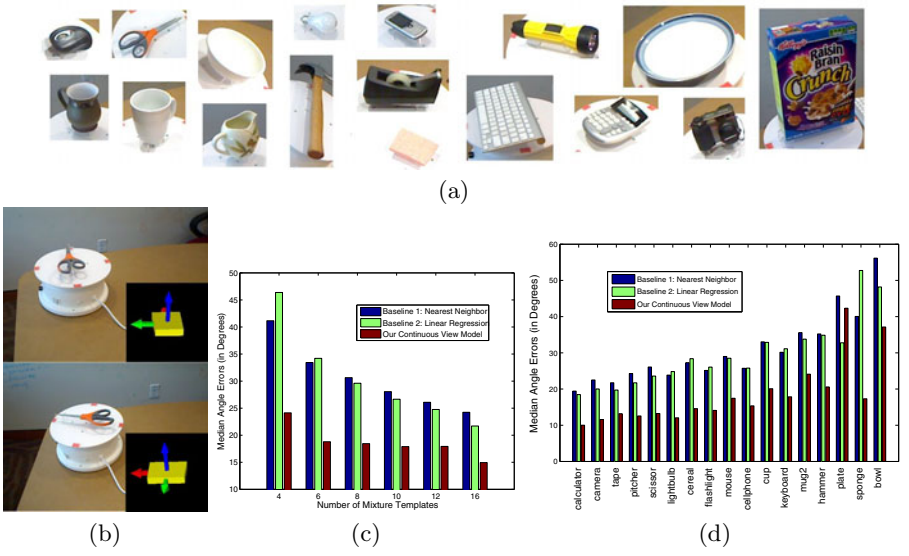


Fig. 5. Continuous viewpoint classification results on the new continuous viewpoint dataset consisting of 17 daily objects (a), covering a wide range of shape and scale. We place objects on a turning table and attach an IMU to a hand-held camera; groundtruth orientations of objects relative to the camera are estimated from both the IMU readings and turning table rotations (b). our discriminatively trained continuous pose model constantly outperforms two baseline methods (assigning to nearest discrete pose, and a linear regression on top of discrete poses). (c) shows the performance comparisons as the number of discrete viewpoints varies in the mixture model. (d) shows the results for each of the 17 objects, with the number of mixture components set to 8. We observe that viewpoint prediction is challenging (and ill-defined for some of the symmetric objects), and our discriminative approach consistently outperforms the baselines for most objects.

template. The second one learns a linear regression model on the responses of all mixture templates to infer viewpoint labels. The comparison of the results is shown in Figure 5(c) where prediction errors are measured by the amount of rotation it takes to go from the predicted pose to the groundtruth pose (in degrees). Because the errors can sometimes be very large due to the symmetry in the object shape and appearance, we use the median angular error as the evaluation metric.

Our proposed continuous viewpoint model constantly outperforms both baselines under different numbers of mixture templates. The errors are reduced as the numbers of templates increase which suggests that a sufficient number of canonical viewpoints is needed to cover the entire viewpoint hemisphere. A closer examination of the per-category performance is shown in Figure 5(d). The errors are in general large for symmetric categories(e.g. plate, bowl) and small for asymmetric ones which meets our intuition. As we see from the examples, the database is challenging: even though the background is simple and so far

instance-based, there is a lot of inherent ambiguity in inferring pose from shape, and the improvement in accuracy using our continuous model is substantial.

7 Conclusion

In this work we have applied the discriminative template learning framework for joint category and viewpoint classification. Our main contribution is to show that a mixture-of-templates model discriminatively learned in a detection framework capture the characteristics of different views and can be directly used for viewpoint classification. Our results significantly outperform the state-of-the-art on a number of standard 3D object databases. We have also shown that with a good initialization (e.g. Normalized Cuts and discriminative clustering), we are able to produce meaningful viewpoint clusters and promising classification accuracy with a small amount of training labels.

In addition, we have extended the mixture-of-templates approach to the continuous viewpoint case. We use a linear model to capture local appearance variations at each canonical view, and these models are discriminatively trained as in the discrete case. We have been building up a dataset with continuous viewpoint groundtruth, and our model has shown promising performance comparing to a number of baselines, including discrete nearest neighbor and linear regression.

Although our work is still in a preliminary stage, we believe that our results are very important in proving the use of discriminative learning for viewpoint classification. It is no coincidence that our results outperform the state of the art on 3D object databases. Just as in the category case, discriminative learning addresses the classification problem directly and is very powerful in exploring noisy image data. There are many future opportunities in exploring the synergies between object classification and viewpoint estimation.

References

1. Koenderink, J., van Doorn, A.: The internal representation of solid shape with respect to vision. Biological Cybernetics 32, 211–216 (1979)
2. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int'l. J. Comp. Vision 60, 91–110 (2004)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
4. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. TPAMI (2009)
5. Savarese, S., Fei-Fei, L.: 3d generic object categorization, localization and pose estimation. In: ICCV (2007)
6. Sun, M., Su, H., Savarese, S., Fei Fei, L.: A multi-view probabilistic model for 3d object classes. In: CVPR, pp. 1247–1254 (2009)
7. Su, H., Sun, M., Fei-Fei, L., Savarese, S.: Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In: ICCV (2009)

8. Everingham, M., Zisserman, A., Williams, C.K.I., Van Gool, L.: The PASCAL Visual Object Classes Challenge 2006, VOC 2006 Results (2006), http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf
9. Arie-Nachmison, M., Basri, R.: Constructing implicit 3d shape models for pose estimation. In: ICCV (2009)
10. Cyr, C., Kimia, B.: A similarity-based aspect-graph approach to 3d object recognition. Int'l. J. Comp. Vision 57, 5–22 (2004)
11. Hoiem, D., Rother, C., Winn, J.: 3d layoutcrf for multi-view object class recognition and segmentation. In: CVPR (2007)
12. Kushal, A., Schmid, C., Ponce, J.: Flexible object models for category-level 3d object recognition. In: CVPR (2004)
13. Chiu, H., Kaelbling, L., Lozano-Perez, T.: Virtual-training for multi-view object class recognition. In: CVPR (2007)
14. Rothganger, F., Lazebnik, S., Schmid, C., Ponce, J.: 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. Int'l. J. Comp. Vision 66, 231–259 (2006)
15. Berg, A., Berg, T., Malik, J.: Shape matching and object recognition using low distortion correspondence. In: CVPR, vol. 1, pp. 26–33 (2005)
16. Bulthoff, H., Edelman, S.: Psychophysical support for a two-dimensional view interpolation theory of object recognition. PNAS 89, 60–64 (1992)
17. DeMenthon, D., Davis, L.: Model-based object pose in 25 lines of code. Int'l. J. Comp. Vision 15, 123–141 (1995)
18. Lavallee, S., Szeliski, R.: Recovering the position and orientation of free-form objects from image contours using 3d distance maps. IEEE Trans. PAMI 17, 378–390 (1995)
19. Collet, A., Berenson, D., Srinivasa, S., Ferguson, D.: Object recognition and full pose registration from a single image for robotic manipulation. In: ICRA (2009)
20. Detry, R., Pugeault, N., Piater, J.: A probabilistic framework for 3D visual object representation. IEEE Trans. PAMI 31, 1790–1803 (2009)
21. Thomas, A., Ferrari, V., Leibe, B., Tuytelaars, T., Schiele, B., Gool, L.V.: Towards multi-view object class detection. In: CVPR (2006)
22. Liebelt, J., Schmid, C., Schertler, K.: Viewpoint-independent object class detection using 3d feature maps. In: CVPR (2008)
23. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: NIPS (2009)
24. Lampert, C.: Partitioning of image datasets using discriminative context information. In: CVPR, pp. 1–8 (2008)
25. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
26. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. PAMI 22, 888–905 (2000)
27. Aiolli, F., Sperduti, A.: Multiclass classification with multi-prototype support vector machines. Journal of Machine Learning Research (2005)
28. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
29. Rosenhahn, B., Brox, T., Weickert, J.: Three-dimensional shape knowledge for joint image segmentation and pose tracking. Int'l. J. Comp. Vision 73, 243–262 (2007)
30. Ozuysal, M., Lepetit, V., Fua, P.: Pose estimation for category specific multiview object localization. In: CVPR (2009)