3D Pose-by-Detection of Vehicles via Discriminatively Reduced Ensembles of Correlation Filters

Yair Movshovitz-Attias[1]
www.cs.cmu.edu/~ymovshov

Vishnu Naresh Boddeti[2]
vishnu.boddeti.net

Zijun Wei[2]
hzwzijun@gmail.com

Yaser Sheikh[2]
www.cs.cmu.edu/~yaser/

[1] Computer Science Department
Carnegie Mellon University
Pennsylvania, USA

[2] Robotics Institute
Carnegie Mellon University
Pennsylvania, USA

Abstract

Estimating the precise pose of a 3D model in an image is challenging; explicitly identifying correspondences is difficult, particularly at smaller scales and in the presence of occlusion. Exemplar classifiers have demonstrated the potential of detection-based approaches to problems where precision is required. In particular, correlation filters explicitly suppress classifier response caused by slight shifts in the bounding box. This property makes them ideal exemplar classifiers for viewpoint discrimination, as small translational shifts can often be confounded with small rotational shifts. However, exemplar based pose-by-detection is not scalable because, as the desired precision of viewpoint estimation increases, the number of exemplars needed increases as well. We present a training framework to reduce an ensemble of exemplar correlation filters for viewpoint estimation by directly optimizing a discriminative objective. We show that the discriminatively reduced ensemble outperforms the state-of-the-art on three publicly available datasets and we introduce a new dataset for continuous car pose estimation in street scene images.

1 Introduction

Accurate estimation of the pose of a 3D model in an image is a fundamental operation in many computer vision and graphics applications, such as 3D scene understanding [24], inserting new objects into images [16], and manipulating current ones [6]. One class of approaches to pose estimation is correspondence-based [27, 28]: individual parts of the object are detected, and a pose estimation algorithm (e.g., perspective-N-point) can be used to find the pose of the 3D object in the image. When the parts are visible, these methods produce accurate continuous estimates of pose. However, if the size of the object in the image is small or if the individual parts are not detectable (e.g., due to occlusion, specularities, or other imaging artifacts), the performance of such methods degrades precipitously. In contrast to

correspondence-based approaches, pose-by-detection methods use a set of view-specific detectors to classify the correct pose; these methods have appeared in various forms such as filter banks, visual sub-categories, and exemplar classifier ensembles [11, 19, 22]. While such approaches have been shown to be robust to many of the short-comings of correspondence-based methods, their primary limitation is that they provide discrete estimates of pose and as finer estimates of pose are required, larger and larger sets of detectors are needed.

To maintain scalability, dimensionality reduction has been explored in prior work [10, 21, 31]. Reduced representations are attractive because of their statistical and computational efficiency. Most approaches reduce the set of classifiers via the classic notion of minimizing the reconstruction error of the original filter set. Such a reduction does not directly guarantee optimal preservation of *detection* performance. This is particularly problematic in the case of viewpoint discrimination, as filters of proximal pose angles are similar. Reduction designed to minimize reconstruction error often results in a loss of view-point precision as the distinctive differences in proximal detectors are averaged out by the reduction.

Correlation filters [30] are designed to explicitly suppress side lobes (false classifier response caused by small translational shifts in the bounding box). As small translational shifts confound small rotational shifts, this property makes correlation filters ideally suited for viewpoint discrimination. In this paper, we present a pose-by-detection approach that uses an ensemble of correlation filters for precise viewpoint discrimination, by using a 3D CAD model of the vehicle to generate renders from viewpoints at the desired precision. A key contribution of this paper is a training framework that generates a discriminatively reduced ensemble of exemplar correlation filters [4] by explicitly optimizing the detection objective. As the ensemble is estimated jointly, this approach intrinsically calibrates the ensemble of exemplar classifiers during construction, precluding the need for an after-the-fact calibration of the ensemble. The result is a scalable approach for pose-by-detection at the desired level of pose precision.

While our method can be applied to any object, we focus on 3D pose estimation of vehicles since cheap, high quality, 3D CAD models are readily available. We demonstrate results that outperform the state-of-the-art on the Weizmann Car View Point (WCVP) dataset [14], the EPFL car multi-view car dataset [22], and the VOC2007 car viewpoint dataset [10]. We also report results on a new data-set based on the CMU-car dataset [4]) for precise viewpoint estimation and detection of cars. These results demonstrate that pose-by-detection based on ensemble of exemplar correlation filters can achieve and exceed the level of precision of correspondence based methods in real datasets; and that discriminative reduction of an ensemble of exemplar classifiers allows scalable performance at higher precision levels.

2 Related Work

Contemporary approaches to pose estimation can be categorized into approaches that use local part correspondences and approaches that use a codebook of view specific detectors. The correspondence based approaches use various forms of local correspondences from points [8], patches [9, 14], and parts [5, 13, 25, 26, 29]. Recently, structure from motion was applied by Glasner et al. [14] on a set of car images to build a 3D object representation, and a discriminative refinement process, comprised of eight viewpoint-aware Support Vector Machines (SVMs), was used to produce the final predictions. Stark et al. [27] used 3D models to learn viewpoint specific car detectors with a parts based representation using rendered non-photo realistic 2D projections of the 3D car models. In a similar vein, Stark et al. [28] trained
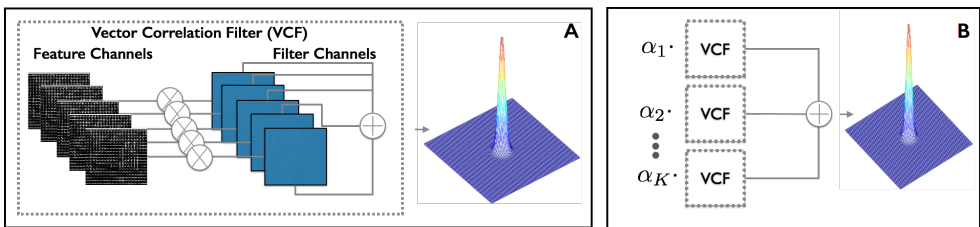
Figure 1 Overview of learning the Exemplar Correlation Filter Basis (EECF). (a) The Vector Correlation Filter (VCF) design aggregates the responses of all feature channels to produce a correlation output which is constrained to have a sharp peak only at the target location. We use $\otimes$ for convolution and $\oplus$ for element-wise sum. (b) Our method (EECF) jointly learns a set of Vector Correlation Filters such that their linear combination produces the sharp peak.

a modified Deformable Parts Model [12] detector using car images retrieved from Google Image Search, classifying cars into one of a discrete set of eight views. When noisy point correspondences are available, a perspective-N-point method such as SoftPosit [8] or EPnP [17] can be used to estimate 3D pose precisely from such local correspondences. While these methods are highly accurate, they are susceptible to failure when these correspondences are compromised, e.g., due to resolution or occlusion.

In contrast to the correspondence-based methods, detector-based approaches implicitly infer object view-point via view-specific detectors. Murase et al. [21] pioneered the use of reduced representation of view-specific detectors based on object appearance. In [22], SIFT histograms were used within a naive Bayes formulation. Liebelt et al. [19] used codebooks of SURF features for matching. In [6], an image sequence of known view-point angles was used for training, and given a test image, distances from each training image were computed and an SVM classifier applied to decide the closest view point that the test image belongs to. In [11], a 3D model was constructed using 2D blueprints, and pose was recovered by optimizing a matching score of quantized-HOGs from 2D images and the 3D model. Zhang et al. [31] noted that multi-view images of 3D objects lie on intrinsically low-dimensional manifolds. They used that intuition to decompose a view matrix $\mathbf{C}$ into $\mathbf{C} = \mathbf{USV}^T$ where $\mathbf{US}$ is a viewpoint basis, and $\mathbf{V}$ is an instance/category basis.

Several approaches to 3D pose estimation have extended the deformable parts model framework to 3D. These approaches [13, 15, 23, 32] augment real 2D images either with synthetically generated renders of 2D projections of 3D object models or introduce additional 3D meta-data to the 2D images. However, the main focus of these methods is precise object localization in 3D, i.e., to predict 3D bounding boxes for objects and estimate object pose from these bounding boxes. Thus, these methods usually only estimate coarse object viewpoints, whereas predicting fine-grained viewpoint is the main focus of this paper.

In this paper, we perform 3D pose-estimation via a set of exemplar classifiers, one for each pose, while addressing the computational and statistical efficiency in using these exemplar classifiers via dimensionality reduction. Unlike previous approaches, we directly learn a ensemble of detectors by optimizing for the *discriminability* of the reduced set, rather than its ability to reconstruct the complete detector ensemble. This encourages the distinctions that allow discrimination of precise pose differences to be preserved.

3   Method

Our approach learns a discriminatively reduced ensemble of exemplar classifiers that spans the vehicle's appearance as it changes with respect to viewpoint. Given a 3D model and a desired precision of $d°$, we densely sample $V = \lceil 360/d \rceil$ viewpoints of the object (along one axis of rotation) and create renders using an empty background in a graphics rendering package (Maya). Exemplar classifiers [20] are trained using a single positive instance and a large set of negative instances. This procedure creates a classifier that is tuned to the characteristics of the single positive instance. We use the vector correlation filter formulation[1] introduced in [4] with Histogram of Oriented Gradients (HOG) features [7].

3.1   Ensemble of Exemplar Classifiers for Pose-by-Detection

Exemplar classifiers are suited to the task of pose-by-detection. For each one of the V viewpoint renders we train an Exemplar Correlation Filter (ECF) using the rendered image as the single positive, and $N - 1$ image patches selected randomly from a background set of images that do not contain the object instance. Each ECF is trained to detect the object from a specific viewpoint.

Let $\{\mathbf{x}_i\}_{i=1}^N$ be a set of Histogram of Oriented Gradients (HOG) representations of the training examples, consisting of one positive exemplar rendering of the v-th view and $N - 1$ negative bounding boxes. Also, define $\{\mathbf{g}_v^1, \cdots, \mathbf{g}_v^C\}$ as the ECF for a viewpoint v, where C is the number of channels of the HOG feature representation (commonly 32). The response of an image $\mathbf{x}_i$ to the filter is defined as

$$\sum_{c=1}^C \mathbf{x}_i^c \otimes \mathbf{g}_v^c = \text{Correlation Output}, \tag{1}$$

where $\otimes$ denotes the 2D convolution operator. The ECF design is posed as:

$$\min_{\mathbf{g}_v^1, \cdots, \mathbf{g}_v^C} \sum_{i=1}^N \left\| \sum_{c=1}^C \mathbf{x}_i^c \otimes \mathbf{g}_v^c - \mathbf{r}_i \right\|_2^2 + \lambda \sum_{c=1}^C \left\| \mathbf{g}_v^c \right\|_2^2, \tag{2}$$

where $\mathbf{r}_i$ is the matrix holding the desired correlation output of the i-th training image, and λ moderates the degree of regularization. The desired correlation output $\mathbf{r}_i$ is set to a positively scaled Gaussian for the positive exemplar and to a negatively scaled Gaussian for the negative patches. This choice of the desired output correlation shape also implicitly calibrates the different exemplar classifiers. The minimization problem can be equivalently posed in the frequency domain to derive a closed form expression, which in turn lends itself to an efficient solution [4]. It should be noted that, as a complete set, each view $v \in V$ is trained independently, and that increase in the desired precision d increases the size of the ensemble (linearly for one axis of rotation, quadratically for two, and cubically for all three).

---

[1]Correlation Filters [30] are a type of classifier that explicitly controls the shape of the entire cross-correlation output between the image and the filter. They are designed to give a sharp peak at the location of the object in the image and no such peak elsewhere. In contrast to SVMs, which treat the HOG feature channels as independent of each other, the vector CF design jointly optimizes all the feature channels to produce the desired output via interactions between multiple channels. The Correlation Filter optimization has an analytical solution, which can be solved efficiently, significantly faster than traditional classifiers (such as SVMs).

3.2 Discriminative Reduction of Ensembles of Correlation Filters

The procedure described in Section 3.1 produces a large set of exemplar classifiers, one per view that needs to be resolved. Let $\mathbf{G} \in \mathbb{R}^{D \times V}$ be the matrix of all V filters arranged as column vectors, where D is the dimensionality of the feature. This set is an exhaustive representation of the object's appearance from many views, but applying all the filters during test time is computationally expensive. It is also highly redundant as many views of the object are similar in appearance. Our reduced Ensemble of Exemplar Correlation Filter (EECF) approach is designed to jointly learn a set of K exemplar correlation filters $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_K]$ (each with C channels) and a set of V sparse coefficient vectors $\mathbf{A} = [\alpha_1, \ldots, \alpha_V]$ such that a detector $\mathbf{g}_v$ for any viewpoint v of the object is defined by

$$\mathbf{g}_v = \mathbf{F}\alpha_v. \tag{3}$$

As before, there are V positive training images, one corresponding to each view that is expected to be resolved. Define B to be a set of randomly selected negative background patches. To learn a reduced EECF, we define the following discriminative objective:

$$\arg\min_{\mathbf{F},\mathbf{A}} \underbrace{\sum_{i:\mathbf{x}_i \in V} \left\| \sum_{k=1}^{K} \alpha_k^i \left(\sum_{c=1}^{C} \mathbf{f}_k^c \otimes \mathbf{x}_i^c \right) - \mathbf{r}^{\text{pos}} \right\|_2^2}_{\text{Controls EECF behavior for positive images}} + \underbrace{\sum_{j:\mathbf{x}_j \in B} \sum_{i:\mathbf{x}_i \in V} \left\| \sum_{k=1}^{K} \alpha_k^i \left(\sum_{c=1}^{C} \mathbf{f}_k^c \otimes \mathbf{x}_j^c \right) - \mathbf{r}^{\text{neg}} \right\|_2^2}_{\text{Controls EECF behavior for negative images}}$$
$$+ \underbrace{\lambda_1 \|\mathbf{F}\|_2^2 + \lambda_2 \|\mathbf{A}\|_1}_{\text{Regularization and sparsity}}, \tag{4}$$

where $\mathbf{x}_i$ and $\mathbf{r}_i$ are as defined for Eq. (2) and $\mathbf{f}_k^c$ is the c-th channel of the k-th reduced filter. α^i is the sparse mixing coefficient for the i-th training image, and λ_1, λ_2 control regularization and enforce sparseness. The need for sparsity will be explained presently.

The first part of the equation guides the optimization to find a reduced set of correlation filters F and a matrix A of coefficients such that Eq. (3) holds. That is, that a detector for any viewpoint can be estimated by a linear combination of the columns of F, weighted by α_i. The second part of the equation controls the discriminability of the ensemble. The key idea is that, as there is no value of α that can be defined for a negative instance, we enforce a negative response $\mathbf{r}_j$ for each negative instance, with any of the learned α. This optimization can be solved efficiently by posing the problem in the Fourier-domain. Details of the derivation are included in the supplementary material.

The mental picture one should have in mind when learning the $\mathbf{F}$ matrix, is that shown in Figure 1 (b). The full basis of K filters is convolved with the image and the convolution with $\mathbf{f}_k$ are weighted by α_k.

Figure 2 shows a number of filters from learned Ensemble of Exemplar Correlation Filter (EECF). Note that the ensemble members capture the shape of the car from different viewpoints. This is interesting as the optimization does not restrict the learned filters to be specific viewpoints.

3.3 Predicting the Viewpoint

The reduced EECF is used to reconstruct the filter responses of the complete ensemble set $\mathbf{G}$. As the learned coefficient matrix $\mathbf{A}$ is sparse this procedure will be much faster than applying
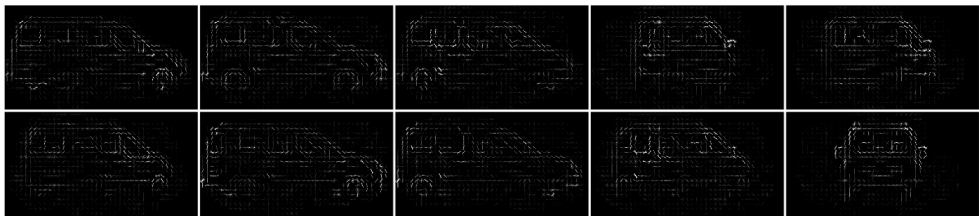
Figure 2 Some filters from a learned Ensemble of Exemplar Correlation Filter (EECF) of size 40. The filters are learned in HOG space. We use the standard HOG visualization in which light pixels show orientations with a large positive weight.

all ECFs on the image. The reduced EECF is applied on an input image at varying scales and locations. The response of all views is estimated using the sparse coefficients. Peaks in these responses are used to predict the location and viewpoint of the car in the image.

Let $\mathbf{r}_v^g \in \mathbb{R}^M$ be the response of evaluating ECF $\mathbf{g}_v$ on a test image I of M HOG cells; $\mathbf{r}_v^g$ can be expressed as $\mathbf{r}_v^g = \mathbf{g}_v \otimes I$ where $\mathbf{g}_v$ is the v-th column of $\mathbf{G}$. With Eq. (3),

$$\tilde{\mathbf{r}}_v^g = \Big(\sum_{k=1}^{K} \alpha_k^v \mathbf{f}_k \Big) \otimes I = \sum_{k=1}^{K} \alpha_k^v (\mathbf{f}_k \otimes I) = \sum_{k=1}^{K} \alpha_k^v \mathbf{r}_k^f, \tag{5}$$

where $\tilde{\mathbf{r}}_v^g$ is an estimator of $\mathbf{r}_v^g$. That is, the response of the ECF corresponding to the v-th view, can be estimated as a weighted sum of the responses of the K ensemble elements. We can reshape $\mathbf{r}_v^g$ and $\mathbf{r}_k^f$ as vectors, and arrange them as the columns of the matrices $\mathbf{R}^g$ and $\mathbf{R}^f$ respectively. An estimator $\tilde{\mathbf{R}}^g$ for the response of all the exemplar filters on the image is

$$\tilde{\mathbf{R}}^g = \mathbf{R}^f \mathbf{A}. \tag{6}$$

Note that as $\mathbf{A}$ is sparse this multiplication is efficient even though $\mathbf{R}^f$ is large.

3.4 Context Rescoring

To reduce the effect of false positive detections we pool information from nearby boxes. Each box is associated with a specific angle and we can use that information to make a better decision. For a box b with a detection score $\hat{s}(b)$, we assign the following score

$$s(b) = \frac{\sum_{b_n \in \mathbb{B}} \mathrm{OS}(b, b_n) \cdot \mathcal{K}(b, b_n) \cdot \hat{s}(b_n)}{\sum_{b_n \in \mathbb{B}} \mathrm{OS}(b, b_n) \cdot \mathcal{K}(b, b_n)}, \tag{7}$$

where $\mathbb{B}$ is the set of all boxes in the image, $\mathrm{OS}(b, b_n)$ is a function that measures the overlap of two boxes, and $\mathcal{K}(b, b_n)$ is a Gaussian Kernel in angle space (circular normal distribution), centered at the angle of box b. The rescoring function reduces the score of false positives as they are unlikely to overlap with other boxes that predict the same angle.

4 Results

The objective of this paper is to estimate the viewpoint of a 3D model using a pose-by-detection approach. Therefore, we focus our evaluation on viewpoint precision for two main

use-case scenarios on four datasets: WCVP, CMU-Car, VOC2007 car viewpoint, and EPFL multi-view cars. The first case (**KNOWN**) is where the image contains a car for which we have a 3D CAD model corresponding to the particular make and model. In the second case (**UNKNOWN**), the image contains a car for which we do not have the exact 3D CAD model. When this is the case, we need to fall back to a generic strategy. We create views for four representative car models: a sedan, a van, a compact car, and a hatchback.

For each car, we create $V = 360$ renders by rotating a synthetic camera around the 3D model in increments of 2 degrees in azimuth and increments of 10 degrees for elevation values (for elevation of 0 and 10 degrees). We train 360 exemplar correlation filters for each one of the views and learn a reduced ensemble with $K = 20$, and $K = 40$ ensemble elements. For a given test image, we apply all the ensemble elements at varying scales and locations and use their output to estimate the cross correlation response of all 360 ECFs. Finally, we apply non-maxima suppression on the predicted bounding boxes. Each bounding box prediction is associated with a specific exemplar and we use that exemplar's angle as our prediction. For the *UNKNOWN* case we follow the protocol described above, but use filters learned from the four representative 3D models.

WCVP. The Weizmann Car Viewpoint (WCVP) dataset [14] contains 1530 images of cars, split into 22 sets. Each set has approximately 70 images and shows a different car model. On this dataset we evaluate both the *KNOWN* and the *UNKNOWN* scenarios. For the *KNOWN* case, we obtained 10 3D CAD models of cars from this dataset from the on-line repositories Doschdesign and Trimble 3D Warehouse. There are 683 images in the data set for these 10 models and we evaluate on those. For the *UNKNOWN* case we evaluate on all the images in WCVP. Table 1(a) shows the median angular error for azimuth prediction over the 10 car models tested for the *KNOWN* use case, and the full dataset for the *UNKNOWN* use case. Using a known 3D CAD model and a full set of exemplar correlation filters as described in Section 3.1 produces an angular error of 6.9°, which is a reduction of 40% in error rate from the 12.25° reported by Glasner et al. [14]. Using a reduced set of 40 filters the error increases by less than 1° to 7.6°. When the model is unknown, a 40 ensemble filter produces an error of 8.4°. This quantifies the benefits of using a known 3D CAD model, compared to the harder problem of using a model trained on a holdout set as in [14] or in the *UNKNOWN* use case. Figure 4 shows a polar histograms of the predicted angles for two examples. The figure shows a distinctive ambiguity in prediction cause by the car's symmetric structure.

CMU-Car. The MIT street scene data set [2] was augmented by Boddeti et al. [4] with landmark annotations for 3,433 cars. To allow for evaluation of precise viewpoint estimation we further augment this data set by providing camera matrices for 3,240 cars. To get the camera viewpoint matrices, we manually annotated a 3D CAD car model with the same landmark locations as the images and used the POSIT algorithm to align the model to the images. To ensure a clean ground truth set, we then back projected the 3D points to the 2D plane and only used cars for which the sum of reprojection error over all landmark points was smaller than 8 pixels. The CAD model used was different from those used later in testing.

We use CMU-Car to evaluate the *UNKNOWN* scenario. For each car, we use the ground truth bounding box to crop a large image section around it (the area of the cropped image is 3 times the area of the bounding box and may conatin other cars). Figure 5 shows examples of detections and viewpoint prediction on this dataset. Table 1(b) shows pose estimation results on this dataset. Most images in this dataset were taken from standing height which explains the low elevation error. The median error in azimuth is 11° when using an EECF of 40 filters. Figure 3 (Right) shows the distribution of angular errors for an ensemble of size 40. The errors made by the algorithm can be split in two; small estimation errors, and

Method	Median Angular Error	
	KNOWN	*UNKNOWN*
EECF, $K = 20$	$10.2°$	9.4
EECF, $K = 40$	$7.6°$	8.4
ECF full (360)	$6.9°$	7.5
Glasner et al. [14]	-	$12.25°$

Method	Median Angular Error	
	Azimuth	Elevation
EECF, $K = 20$	$26.0°$	$3.8°$
EECF, $K = 40$	$11.48°$	$3.6°$
ECF Full (360)	$3.2°$	$3.0°$
Glasner et al. [14]	-	-

(a) Azimuth Estimation: WCVP   (b) Pose Estmation: CMU Car

Table 1 (a) Median angular error in azimuth estimation for WCVP dataset. When the car model in the image is known, using 40 filters, our ensemble exemplar correlation filter method achieves a median error of $7.6°$. When the car model is unknown a basis of 40 filters has a median error of $8.4°$. Previous results on this dataset have a median error of $12.25°$. (b) Median angular error in azimuth and elevation estimation for CMU Car dataset using unknown models.
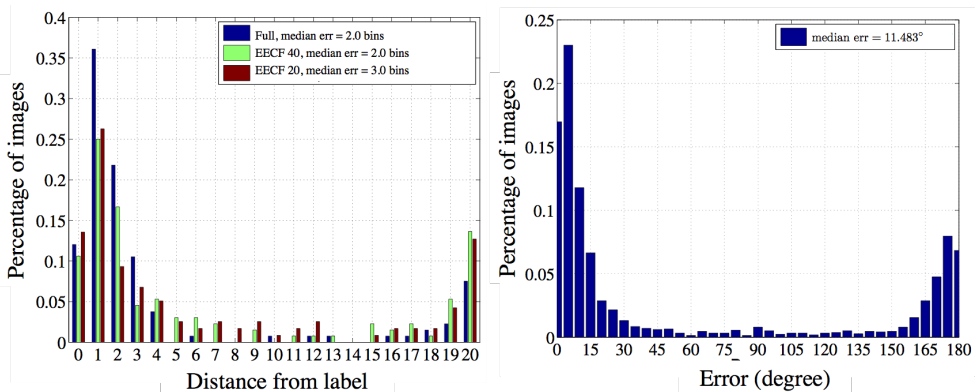

Figure 3 (Left) A histogram of view errors for the VOC2007 car viewpoint dataset. Most images were matched near their most similar pose and so there is a peak around a 1 bin error. (Right) Histogram of angular error on the CMU-Car dataset. The Median error is $11.48°$. In both cases, the smaller peak is due to the $180°$ symmetricity.

$180°$ambiguity errors. There are few errors in the $[30°, 165°]$ range.

VOC2007 car viewpoint. In [1], Arie-Nachimson and Basri provided viewpoint annotations for 200 cars from the PASCAL VOC 2007 test set car category. Each car is labeled with one of 40 viewpoint labels that correspond to reference images. Figure 3 (Left) shows a histogram of prediction distance from true labels. The majority of the predictions are within a distance of 2 to the ground truth label. This is an improvement over the results of [1] which had a median distance of 3 bins.

EPFL Multi-View Cars [22]. This dataset contains 2299 images of 20 different car models. Each car is imaged at intervals of $3°$ for a full turn. We apply the *UNKNOWN* use case on this dataset as outlined above. Using 40 ensemble filters we achieve a median error of $19°$ compared with $24.83°$ reported in [14].
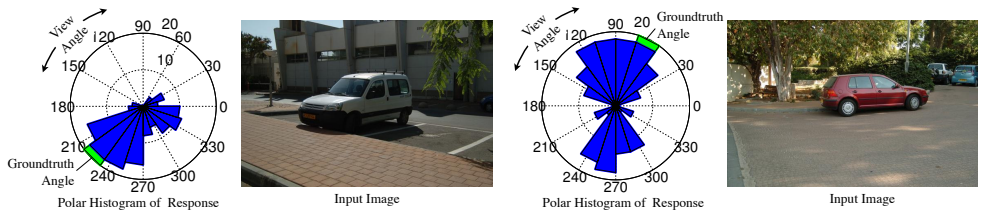
Figure 4 Polar histogram of scores. The left example shows a van at an oblique angle, with little ambiguity in the distribution of responses. The right example shows a side view with the distinctive symmetric ambiguity.

5 Discussion

Exemplar-based approaches have been gaining popularity as they can provide state-of-the-art detection for precise detection problems. However, these methods scale poorly; as the desired precision of viewpoint estimation increases, the number of exemplars needed increases as well. We present a pose-by-detection framework that considers both computational and statistical efficiency. Our approach *directly* optimizes discriminative power to efficiently detect the viewpoint of an object. The need for reducing the number of applied filters is especially prominent at the two extremes of scale: on a mobile platform where the available computation power is limited, and on the data-center scale where the number of images to be evaluated is vast. In lieu of reported computation times that depends on hardware specification and implementation, we analyze the computational complexity of our approach. For an image with a HOG representation of size $M \times C$, a filter of size $m \times C$, V exemplars, and a learned ensemble of K filters, when all exemplars are convolved, the complexity is $\mathcal{O}(CVM\log_2 m)$. With the reduced set the complexity is $\mathcal{O}(CKM\log_2 m + \gamma MKV)$, where γ is the fraction of non-zero elements in $\mathbf{A}$. Thus, the computational savings are $\mathcal{O}\left(K/V + \gamma K/C\log_2 m\right)$.

Our method produces state-of-the-art results on the WCVP dataset and the EPFL dataset, significantly reducing the error from previous results. Additionally, we have introduced a new dataset, CMU-Car, for viewpoint estimation that contains more than 3000 images. On this dataset, we achieve $11.5°$ azimuth error, and $3.6°$ elevation error by using a 40 element EECF basis. A fundamental limitation of pose-by-detection approaches, including the method presented in this paper, is that as the precision is increased beyond a point, it becomes increasingly harder to discriminate between nearby viewpoints, because the underlying features are designed to provide invariance to small spatial variations. This suggests an important direction of future work, tying feature selection into the detector optimization.

6 Acknowledgments

(a) Alignment results with known 3D Vehicle Model

(b) Alignment results with unknown 3D Vehicle Model

(c) Failure cases

Figure 5 Example results from all of the datasets used. Each row shows input images (top) and overlaid pose estimation results (bottom). (a) Results using a *Known* 3D model, (b) results using an *Unknown* 3D model, and (c) failure cases.

References

[1] M. Arie-Nachimson and R. Basri. Constructing implicit 3d shape models for pose estimation. In *International Conference on Computer Vision*. IEEE, 2009.

[2] S.M. Bileschi. *StreetScenes: Towards scene understanding in still images*. PhD thesis, Massachusetts Institute of Technology, 2006.

[3] V. Blanz, B. Schölkopf, HCBVV Bülthoff, C. Burges, V. Vapnik, and T. Vetter. Comparison of view-based object recognition algorithms using realistic 3d models. In *Artificial Neural Networks-ICANN*. Springer, 1996.

[4] V. N. Boddeti, T. Kanade, and B. V. K. Vijaya Kumar. Correlation filters for object alignment. In *Computer Vision and Pattern Recognition*. IEEE, 2013.

[5] R.J. Campbell and P.J. Flynn. A survey of free-form object representation and recognition techniques. *Computer Vision and Image Understanding*, 2001.

[6] T. Chen, Z. Zhu, A. Shamir, S. Hu, and D. Cohen-Or. 3-sweep: Extracting editable objects from a single photo. *ACM Transactions on Graphics*, 2013.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition*. IEEE, 2005.

[8] P. David, D. Dementhon, R. Duraiswami, and H. Samet. Softposit: Simultaneous pose and correspondence determination. *International Journal of Computer Vision*, 2004.

[9] Y. Deng, Q. Yang, X. Lin, and X. Tang. A symmetric patch-based correspondence model for occlusion handling. In *International Conference on Computer Vision*. IEEE, 2005.

[10] A. Elgammal and C. Lee. Homeomorphic manifold analysis (hma): Generalized separation of style and content on manifolds. *Image and Vision Computing*, 2013.

[11] E.Yörük and R. Vidal. Efficient object localization and pose estimation with 3d wireframe models. In *4th IEEE Workshop on 3D Representation and Recognition (ICCV)*. IEEE, 2013.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2010.

[13] S. Fidler, S.J. Dickinson, and R. Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. In *Neural Information Processing Systems*, 2012.

[14] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and continuous pose estimation. *Image and Vision Computing*, 2012.

[15] M. Hejrati and D. Ramanan. Analyzing 3d objects in cluttered images. In *Neural Information Processing Systems*, 2012.

[16] K. Karsch, C. Liu, and S.B. Kang. Depth extraction from video using non-parametric sampling. In *European Conference of Computer Vision*, 2012.

[17] V. Lepetit, F.Moreno-Noguer, and P.Fua. Epnp: An accurate o(n) solution to the pnp problem. *International Journal Computer Vision*, 2009.

[18] J. Liebelt and C. Schmid. Multi-view object class detection with a 3d geometric model. In *Computer Vision and Pattern Recognition*. IEEE, 2010.

[19] J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3d feature maps. In *Computer Vision and Pattern Recognition*. IEEE, 2008.

[20] T. Malisiewicz, A. Gupta, and A.A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *International Conference on Computer Vision*. IEEE, 2011.

[21] H. Murase and S.K. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision (IJCV)*, 1995.

[22] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *Computer Vision and Pattern Recognition*. IEEE, 2009.

[23] B. Pepik, P. Gehler, M. Stark, and B. Schiele. $3d^2pm - 3d$ deformable part models. In *Computer Vision–ECCV 2012*. Springer, 2012.

[24] S. Satkin, J. Lin, and M. Hebert. Data-driven scene understanding from 3d models. In *British Machine Vision Conference*, 2012.

[25] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *International Conference on Computer Vision*. IEEE, 2007.

[26] J. Schels, J. Liebelt, K. Schertler, and R. Lienhart. Synthetically trained multi-view object class and viewpoint detection for advanced image retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. ACM, 2011.

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

[28] M. Stark, J. Krause, B. Pepik, D. Meger, J.J. Little, B. Schiele, and D. Koller. Fine-grained categorization for 3d scene understanding. In *British Machine Vision Conference*, 2012.

[29] H. Su, M. Sun, L. Fei-Fei, and S. Savarese. Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In *International Conference on Computer Vision*. IEEE, 2009.

[30] B. V. K. Vijaya Kumar, A. Mahalanobis, and R. D. Juday. *Correlation Pattern Recog*. Cambridge Univ. Press, 2005. ISBN 0521571030.

[31] H. Zhang, T. El-Gaaly, A. Elgammal, and Z. Jiang. Joint object and pose recognition using homeomorphic manifold analysis. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[32] Z. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3d representations for object recognition and modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2013.