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

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

In this supplementary material we show a full derivation of our Reduced Ensemble of Correlation Filters and provide more quantitative results that are not shown in the paper due to space constraints.

1 Ensemble of Exemplar Classifiers for Pose-by-Detection

1.1 Exemplar Correlation Filters

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:

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

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 [1]. 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). Figure 1 (A) shows the training configuration for one exemplar correlation filter. For visualization clarity we do not show negative images.

1.2 Discriminative Reduction of Ensembles of Correlation Filters

The procedure described in Section 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:

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

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

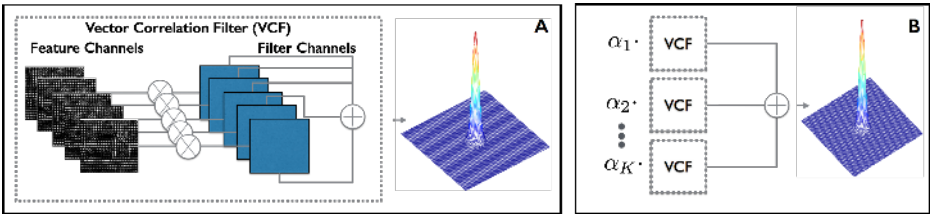

Figure 1: Overview of learning the Exemplar Correlation Filter Basis (EECF). (A) The Vector Correlation Filter (VCF) design aggregates the responses of all feature channels to produce a correlation output which is constrained to have a sharp peak only at the target location. We use $\otimes$ for convolution 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.

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.

2 Optimization

2.1 Learning Vector Correlation Filters

We first provide a short derivation of VCF here as it will be helpful for understanding our joint approach in 1.2. For an in depth tutorial, we direct the reader to [2].

We can restate Eq. (2) in the frequency domain using Parseval's theorem [3]

$$\arg\min_{\hat{\mathbf{f}}^1,\dots\hat{\mathbf{f}}^c} \sum_{i=1}^{N} \left\| \sum_{c=1}^{C} \hat{\mathbf{X}}_i^c \hat{\mathbf{f}}^c - \hat{\mathbf{r}}_i \right\|_2^2 + \lambda \sum_{c=1}^{C} \hat{\mathbf{f}}^{c\dagger} \hat{\mathbf{f}}^c. \tag{5}$$

Where we use the $\hat{}$ notation for the Fourier Transform of a vector, for example $\hat{\mathbf{r}}_i$ is the Fourier Transform of $\mathbf{r}_i$. $\hat{\mathbf{X}}_i^c$ is a diagonal matrix with the values of $\hat{\mathbf{x}}_i^c$ on its main diagonal. $\hat{\mathbf{f}}^{c\dagger}$ is the conjugate transpose of $\hat{\mathbf{f}}^c$. Let $\hat{\mathbf{f}}$ and $\hat{\mathbf{y}}$ be

$$\hat{\mathbf{f}} = \begin{bmatrix} \hat{\mathbf{f}}^1 \\ \vdots \\ \hat{\mathbf{f}}^C \end{bmatrix} \qquad \hat{\mathbf{y}} = \sum_{i=1}^{N} \hat{\mathbf{y}}_i = \sum_{i=1}^{N} \begin{bmatrix} \hat{\mathbf{X}}_i^1 \hat{\mathbf{r}}_i \\ \vdots \\ \hat{\mathbf{X}}_i^C \hat{\mathbf{r}}_i \end{bmatrix}, \tag{6}$$

Then Eq. (5) can be written in matrix form as

$$\arg\min_{\hat{\mathbf{f}}} \hat{\mathbf{f}}^\dagger \hat{\mathbf{S}} \hat{\mathbf{f}} - 2\hat{\mathbf{f}}^\dagger \hat{\mathbf{y}}, \tag{7}$$

where $\hat{\mathbf{S}} = \hat{\mathbf{D}} + \lambda \mathbf{I}$ and

$$\hat{\mathbf{D}} = \sum_{i=1}^{N} \hat{\mathbf{D}}_i = \sum_{i=1}^{N} \begin{bmatrix} \hat{\mathbf{X}}_i^{1\dagger} \hat{\mathbf{X}}_i^1 & \dots & \hat{\mathbf{X}}_i^{1\dagger} \hat{\mathbf{X}}_i^C \\ \vdots & \ddots & \vdots \\ \hat{\mathbf{X}}_i^{C\dagger} \hat{\mathbf{X}}_i^1 & \dots & \hat{\mathbf{X}}_i^{C\dagger} \hat{\mathbf{X}}_i^C \end{bmatrix}. \tag{8}$$

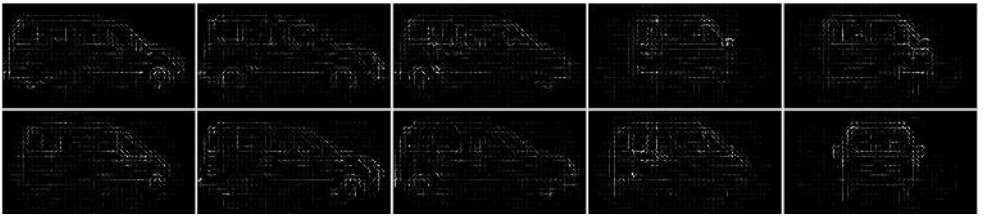

Figure 2: Some filters from a learned Ensemble of Exemplar Correlation Filter (EECF) of size 40. The filters are learned in HOG space. We use the standard HOG visualization in which light pixels show orientations with a large positive weight.

The matrix $\hat{\mathbf{S}}$ is block diagonal. $\hat{\mathbf{D}}$ is the interaction energy between different channels of the feature. For a HOG template of size m each block is of size $m \times m$, and the matrix is of size $mC \times mC$.

The desired filter is

$$\hat{\mathbf{f}} = \hat{\mathbf{S}}^{-1}\hat{\mathbf{y}}. \tag{9}$$

2.2 Learning the Basis Filters

The filter basis can be solved for analytically in the frequency domain by weighted averaging of the inputs. Define

$$\hat{\mathbf{f}} = \left[\hat{\mathbf{f}}_1^1 \ldots \hat{\mathbf{f}}_1^C \ldots \hat{\mathbf{f}}_K^1 \ldots \hat{\mathbf{f}}_K^C\right]^T. \tag{10}$$

The solution that minimizes Eq. (4) is

$$\hat{\mathbf{f}} = (\hat{\mathbf{D}} + \lambda_1 \mathbf{I})^{-1}\hat{\mathbf{y}}. \tag{11}$$

where

$$\hat{\mathbf{D}} = \sum_{i=1}^{N} \begin{bmatrix} \alpha_1^i \alpha_1^i \hat{\mathbf{D}}_i, \ldots \alpha_1^i \alpha_K^i \hat{\mathbf{D}}_i \\ \vdots \quad \ddots \quad \vdots \\ \alpha_K^i \alpha_1^i \hat{\mathbf{D}}_i, \ldots \alpha_K^i \alpha_K^i \hat{\mathbf{D}}_i \end{bmatrix}, \hat{\mathbf{y}} = \sum_{i=1}^{N} \begin{bmatrix} \alpha_1^i \hat{\mathbf{y}}_i \\ \vdots \\ \alpha_K^i \hat{\mathbf{y}}_i \end{bmatrix} \tag{12}$$

and $\hat{\mathbf{D}}_i$ and $\hat{\mathbf{y}}_i$ are as defined in Eq. (8) and Eq. (6)

The matrix $\hat{\mathbf{S}}$ is now of size $mKC \times mKC$ which can become large — typical values are $m = 450, K = 20, C = 32$. However this matrix is block diagonal and sparse. Each $\hat{\mathbf{X}}_i$ block is of size $m \times m$ but only has m non zero elements. The number of non zero elements in $\hat{\mathbf{S}}$ is $m(CK)^2$. The block structure of the matrix allows for efficient inversion using Schur complement matrix inversion.

We initialize the matrix of coefficients A using sparse coding on the HOG features of the training images. This provides a reasonable starting point for the optimization.

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.

Detection WCVP (mAP)		
Method	*KNOWN*	*UNKNOWN*
EECF, $K = 20$	0.54	0.52
EECF, $K = 40$	0.61	0.58
ECF full (360)	0.72	0.62

((a)) Detection Results: WCVP

Detection CMU-car (AP)	
Method	
EECF, $K = 20$	0.2
EECF, $K = 40$	0.36
ECF Full (360)	0.46

((b)) Detection Results: CMU-car

Table 1: (a) Detection results on the WCVP dataset. Reported numbers are the mean average precision (mAP) over all 22 car classes in the dataset. (b) Detection results on the CMU-car dataset. Reported numbers are average precision (AP).

3 Results

The main focus of our method is estimating the pose of an object. For completeness we also report detection results on two datasets: WCVP and CMU-car. Table 1 shows the average precision (AP) of our method using Ensembles of 20 and 40 correlation filters, and compared to using a non reduced set of 360 filters. We urge the reader to be careful when interpreting these reported numbers. Both datasets have images with multiple cars, but only a single instance is annotated. This can have unintuitive results, e.g. an improved detector may have lower AP. Figure 3 shows a number of instances in which the top detection is a car that is not annotated in the dataset.

Figure 3: The results above illustrate the problem with measuring detection rate on datasets used to evaluate pose estimation. In each image there only one car is annotated (yellow dotted box). In these examples the top scoring detection (red solid box) is on another vehicle.

References

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

[2] V.N. Boddeti. *Advances in correlation filters: vector features, structured prediction and shape alignment*. PhD thesis, Carnegie Mellon University, 2012.

[3] A.V. Oppenheim, A.S. Willsky, and S.H. Nawab. *Signals and systems*. Prentice-Hall Englewood Cliffs, NJ, 1983.