Measuring and increasing the capacity of Natural HOG Statistics

Tinghui Zhou[1] Abhinav Shrivastava[1] Guillaume Obozinski[2] Abhinav Gupta[1] Alexei A. Efros[1]
[1]Carnegie Mellon University [2]Ecole des Ponts

Abstract

Just when the question of what is the best method for sliding window object detection was thought to be settled (linear SVM on groups of HOG features), several recent papers have suddenly made the story a lot murkier. In this unusual paper (more typical of the natural sciences than computer science), we explore one recent, surprising work – Hariharan et al. [7] – and try to understand it better experimentally. In particular we try to understand the relationship between the covariance matrix and the amount of data needed to fill the model's effective capacity. We find that the amount of data needed to saturate the performance of the system is surprisingly little – less than 20 images. Based on our findings, we propose two extensions that substantially improve the object detection performance on the PASCAL VOC dataset, pushing it ahead of Exemplar-SVM [10]. Since our extensions are relatively simple, and the performance is clearly better, we expect this work to be immediately useful in practice, whenever there is a need to compute distances in HOG space.

1. Introduction

For the last half a decade, it appeared that the story of object detection has been pretty much settled. The story that started with Dalal and Triggs [3], got solidified with Felzenszwalb et al [5] and is now taught in vision courses and appears in the textbooks is a story with three critical components: 1) sophisticated Histogram of Gradients (HOG) feature (to encode various visual invariances), 2) linear SVM classifier (to build a discriminative object class model that can generalize across the positive class instances), and 3) lots of negative data (to use in large-scale hard-negative mining). Each of these components has been thoroughly tested and reasonably well understood, and the various implementations of the framework have consistently outperformed other methods on a number of difficult datasets, including the highly regarded PASCAL VOC [4].

But recently, this nice and well-understood story has been getting a lot murkier... The first sign that things are not as they seem came two yeas ago from the work on Exemplar-SVMs [10]. It proposed a very simple experiment: to learn a separate classifier for each positive object instance in the dataset separately and then combine the results. This would seem like a silly thing to do since most researchers have attributed the success of SVM-based detection methods to their ability to generalize across the positive class instances. In the Exemplar-SVM setup, no generalization across the positives is possible by definition, yet, surprisingly, no drastic drop in detection performance was observed compared to the standard, category-based SVM approaches. The authors hypothesized that it was the pressure from the hard-mined negatives, not the generalization from the positives, that explained the good SVM performance. In their follow-up work [15], they showed that there was no need for any special "negative" data – any large-enough sample of the natural world (e.g. hundreds of thousands random, unlabeled images off the Internet) worked just as well. This observation was used to apply Exemplar-SVM to image retrieval, essentially using the SVM to compute a local distance metric for every query (using the "natural world dataset" as negatives) and then find the nearest neighbors in that new space. The explanation was that the negative support vectors chosen by the SVM represent a visually meaningful neighborhood around the query image and should provide a good local distance.

Then last year, the work by Hariharan *et al*. [7] gave an even bigger shock to the commonly accepted wisdom. Taking the Exemplar-SVM argument further, they showed that there was no need for the SVM at all! The idea is that since the negative set is simply capturing the natural image statistics and can be the same for every exemplar, it should be possible to model it once and for all. Indeed, modeling the negatives as a high dimensional Gaussian turns the SVM into an LDA (Linear Discriminant Analysis) which incurs negligible training cost. Even though LDA is a very coarse approximation of the linear SVM, surprisingly, the Exemplar-LDA results on PASCAL VOC are just a bit worse than Exemplar-SVM, at a fraction of the computational cost for training. Conceptually, Exemplar-LDA is equivalent to simply whitening the data and then finding the nearest neighbors in the decorrelated space! Suddenly, the magic of linear SVMs for object detection is not looking so

magical anymore.

Of course, we still have the "magic of data". After all, building the covariance matrix required to whiten the data still requires a very large "natural world dataset" (around $10,000$ images or over a million patches is used in Hariharan *et al*. [7]). This is not too surprising, since our visual world is so extremely rich and varied – it makes sense that the model would require a lot of data to capture enough natural image statistics to start being useful. *Except...* it turns out not to be the case.

Contributions: In this unusual paper (more typical of the natural sciences than computer science), we explore one existing approach – Hariharan *et al*. [7] – and try to understand it better experimentally. In particular we try to understand the relationship between the covariance matrix and the amount of data needed to fill the model's effective capacity. Based on our findings, we propose two extensions that substantially improve the object detection performance on the PASCAL VOC dataset, pushing it ahead of Exemplar-SVM [10]. Since our extensions are relatively simple, and the performance is clearly better, we expect this work to be immediately useful in practice, whenever there is a need to compute distances in HOG space. More importantly, we hope that this paper will restart a discussion on the role of natural image statistics for visual matching, yielding further follow-up work in this neglected area.

2. Motivation

In this section, we first briefly review the work by Hariharan *et al.* [7] on summarizing the natural image statistics in the HOG feature space with a Gaussian distribution, and then provide some experimental analysis which will motivate the rest of the paper.

2.1. Whitened Histograms of Orientations (WHO)

Hariharan *et al.* [7] present a method for estimating statistics of natural images in the HOG feature space, so that the entire set of natural images can be approximately represented by a high-dimensional Gaussian given by the mean μ_0 and covariance Σ. In this way, object detectors can be trained efficiently (without the infamously expensive hard-mining step) using Linear Discriminant Analysis [11]:

$$w = \Sigma^{-1}(\mu_1 - \mu_0)\ , \tag{1}$$

where w denotes the learned detector weights, and μ_1 is the input HOG descriptor. Once the mean and covariance are given, the original HOG descriptor can be whitened by $\hat{x} = \Sigma^{-1/2}(x - \mu_0)$, resulting in a new feature descriptor that has an isotropic covariance matrix. This new descriptor is referred as Whitened Histograms of Orientations (WHO) and it is shown to perfom better for pairwise comparisons and clustering than the original HOG descriptor. Throughout the rest of the paper, we'll use WHO to refer to their approach.

Mean estimation: The mean vector μ_0 for any $M \times N$ HOG template is obtained by simply replicating the mean for a single HOG cell μ by $N * M$ times, where μ is estimated by sampling a large set of HOG cells from an unlabeled image set and taking the mean.

Covariance estimation: The covariance Σ is modeled as a block matrix with blocks $\Sigma_{(ij),(lk)} = E[x_{ij}x_{lk}^\top]$, and each block is estimated by a *spatial autocorrelation function* [14] exploiting the property of translation invariance for natural images:

$$\Sigma_{(ij),(lk)} = \Gamma_{(i-l),(j-k)} = E[x_{uv}x_{(u+i-l),(v+j-k)}^\top]\ , \tag{2}$$

where the expectation is over cell locations (u, v) and HOG features x. This means that $\Sigma_{(ij),(lk)}$ only depends on the relative spatial offsets $(i-l)$ and $(j-k)$. Furthermore, since the statistics of smaller-size templates can be obtained by marginalizing out that of larger-size templates, the covariance for smaller-size templates can be generated by selecting blocks of Σ for larger templates. Thus, one only needs to estimate Σ for a large-enough template (*e.g.* 20×20) once for all. For regularization, a small constant λ is added to the diagonals of Σ when training with LDA.

2.2. How much data does WHO need?

At first glance, one might think that the amount of data required to estimate Σ would be very large (Hariharan *et al.* [7] report using $10,000$ images and even then Σ was not invertible and needed to be regularized). But what's the minimum number of images required for the same level of performance? To investigate this, we randomly sampled N (about 10^7) image patches of 160×160 pixels (*i.e.* 20×20 HOG windows) from PASCAL VOC 2010 dataset, which still only represents a very small fraction of the available visual data. We further divided the patches into subsets of different sizes (from 10^4 to 10^7), and estimated a covariance matrix using each subset independently.

One way to measure how different these covariances are in practice is to apply LDA to train object detectors using each of them independently, and see how the detection performance changes. Therefore, for each covariance matrix, we select 50 exemplars for each of the 5 categories (*horse, train, tvmonitor, bus* and *sheep*) from PASCAL VOC 2007 dataset [4], train a LDA model for each exemplar using Eq. 1, and evaluate on the full PASCAL test set (exemplar scores are calibrated using logistic regression as in [10]). The performance for each category using different covariances is shown in Fig. 1. Surprisingly, although the covariance is of size greater than 10000×10000, its performance saturates rapidly with respect to the amount of image data used for estimation. In most cases, the saturation

point roughly happens at $N = 10^5$ patches (less than 20 images!). Therefore, it appears WHO does not actually need much data at all, and that attempts to throw more data at it do not improve the performance.

We can also try to directly measure how different are the estimated covariance matrices using all vs. a part of the data using Affine Invariant Riemannian Metric (AIRM) [13], a widely used distance metric for covariance matrices:

$$D(\Sigma_1, \Sigma_2) = \| \log(\Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2}) \|_F, \qquad (3)$$

where $\log(\cdot)$ is the matrix logarithm, and $\| \cdot \|_F$ denotes the Frobenius norm. In our case, Σ_1 is the covariance estimated using all the image patches, and Σ_2 is the covariance estimated using only a subset of them. Fig. 1(f) shows how the distance varies as different number of image patches are used for covariance estimation.

2.3. Statistics for different scenes

Our previous experiments demonstrated how the estimation of natural image statistics using WHO saturates rapidly, with a large portion of image data being unused. We can think of two possible explanations for this: 1) The PASCAL dataset contains a rich combination of different types of visual scenes (indoor, outdoor, natural, man-made, etc) so perhaps capturing them by the same covariance matrix averages away valuable information; 2) The effective capacity of the WHO model is relatively low, and it is unable to fully capture the richness of the visual world.

To evaluate hypothesis 1), we conducted an experiment where we varied the subset of image data (background) on which we estimate the image statistics based on scene type. Specifically, we used three different annotated sets of images from the SUN database [1] to estimate three different covariance matrices: Indoor Images, Outdoor Man-made and Outdoor Natural. Our hypothesis is that since indoor images have significantly different visual properties from outdoor natural images, the estimated covariance matrices should be quite different. Hence, training LDA with these different covariances should result in very different object detectors. Figure 2 shows both the learned HOG weights and the top retrievals for five different exemplars using covariance matrices estimated from three different scenes. As one can see, the retrievals for all the three covariances look quite similar with only small changes in their ordering. For example, for the car exemplar, it retrieves exactly the same cars in all the three cases, and for the horse exemplar, even the false positives are almost the same across three cases. This experiment suggests that 1) is *not* the reason why statistics captured by WHO saturates so quickly. Therefore, we hypothesize that it is the low model capacity of WHO that causes rapid saturation.

3. Towards capturing richer image statistics with more data

As shown in our experiments from Section 2.1, the saturation of the the "natural HOG statistics" happens surprisingly fast, implying it has insufficient capacity for fully capturing properties of the visual world. In this section, we propose two ways of increasing the capacity of the WHO framework: 1) modeling statistics across multiple scales in the HOG space and 2) modeling HOG statistics in a new space induced by the intersection kernel via explicit feature mapping. Both have been shown to outperform WHO in our experiments.

3.1. Multiscale statistics

One intrinsic disadvantage of the WHO framework is that when estimating the covariance, it only looks at pairwise cell statistics within the *lcoal* scale, and any given cell has no access to information in other scales. This is not ideal as it has been shown that object detectors trained on multiple scales tend to have better performance [12, 9] than those trained on a single scale. Therefore, here we propose a multiscale statistical model that captures not only the cell statistics within the same scale (as in WHO) but also the more *global* statistics across multiple scales.

Denoting the image at the original scale as $I^{(1)}$ with m rows and n columns, we will demonstrate the proposed multiscale framework with three scales: $I^{(1)}$, $I^{(2)}$ (original image down-sampled by half with $m/2$ rows and $n/2$ columns), and $I^{(3)}$ (original image down-sampled by one-fourth with $m/4$ rows and $n/4$ columns). Therefore, a cell in $I^{(2)}$ ($I^{(3)}$) summarizes information for the corresponding 4 (16) cells of $I^{(1)}$ at the original scale. Alternatively, one can obtain the image representation at multiple scales by simply using different bin sizes for HOG feature extraction.

3.1.1 Mean estimation

Due to the property of scale invariance for natural images, we assume that the mean is the same across different scales, and use the same method as WHO to estimate the mean (see Section 2.1 for more details).

3.1.2 Covariance estimation

We propose multiscale covariance estimation that captures statistics not only for cells within the local scale but also across multiple scales. This is achieved by utilizing the autocorrelation function not only for spatial offsets but also for offsets in the scale space. Specifically, given two HOG cells at location (i, j) of $I^{(s)}$ and location (l, k) of $I^{(t)}$, re-

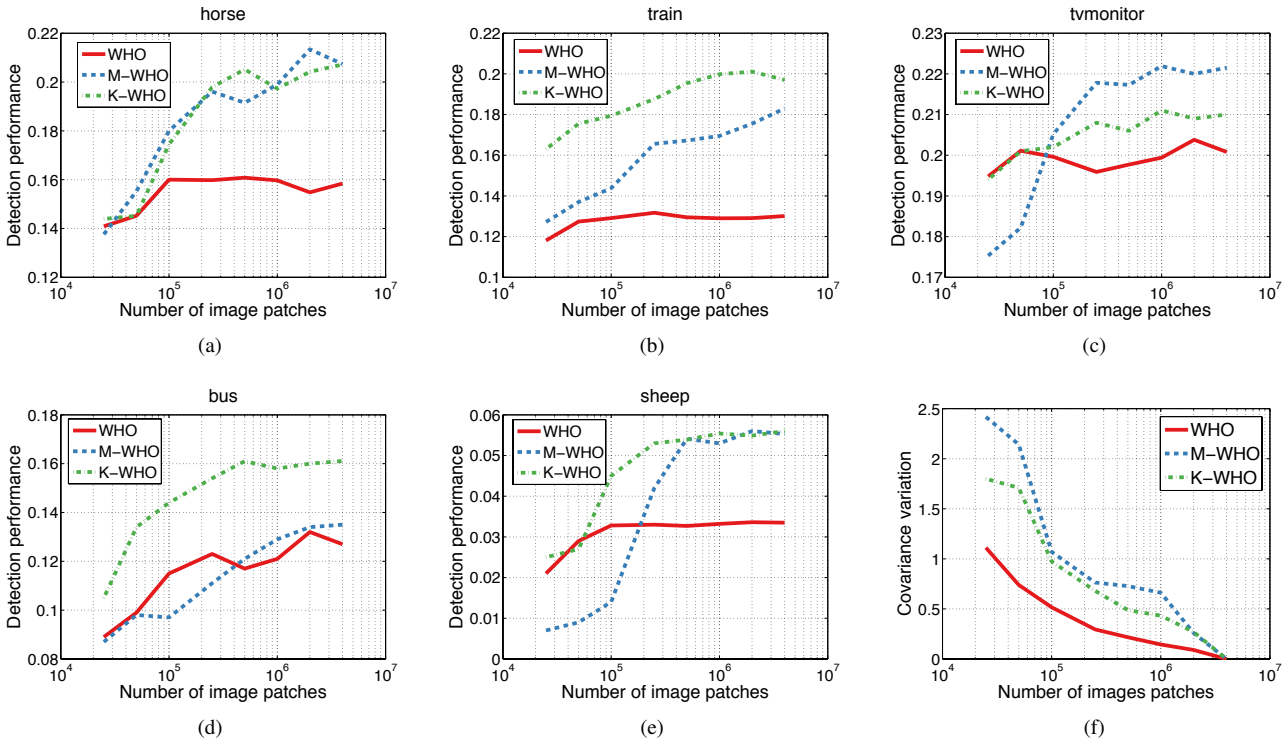

Figure 1. (a)–(e) Variation of the detection performance for different categories when different number of images are used to estimate the background statistics. Overall, WHO tends to saturate much faster than Multiscale-WHO and Kernelized-WHO, implying that its effective model capacity is lower. (f) Difference of covariances (measured by the AIRM metric) estimated using varying amount of image patches compared to the one estimated using all patches.

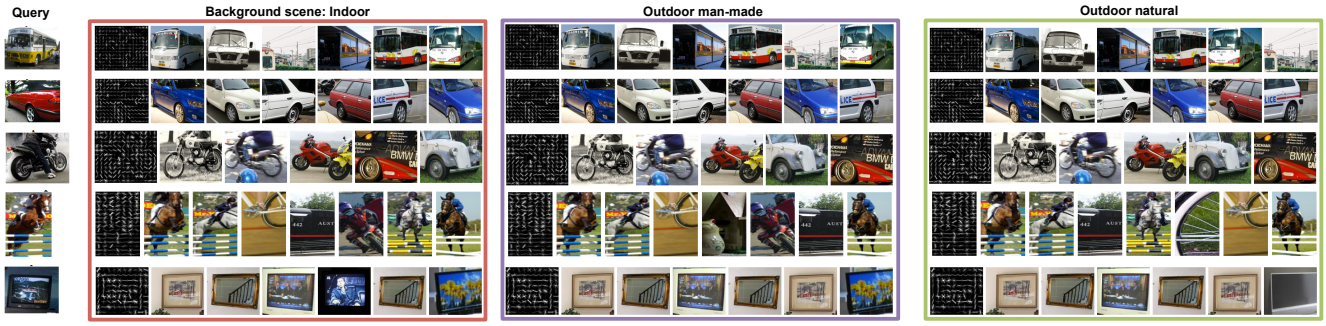

Figure 2. Top retrievals by the WHO model trained using background statistics estimated from different scene images. Surprisingly, despite high variation in the visual properties of these three different scenes, the statistics estimated using WHO appear to be almost the same – the top retrievals seem to make no difference across the three cases, and for some exemplars even the retrieved false positives are the same.

spectively, the covariance Σ is estimated by

$$\Sigma^{(s,t)}_{(ij),(lk)} = \Gamma^{s-t}_{(i-l),(j-k)} = E[(x^{w}_{uv})(x^{w+s-t}_{(u+i-l),(v+j-k)})^{\top}] \quad (4)$$

where s, t and w are scale indices, and x is the HOG descriptor for the corresponding cell. This means that the covariance $\Sigma^{(s,t)}_{(ij),(lk)}$ depends on the relative offset both in cell location and the scale space. If we are only concerned with the covariance across three scales, $I^{(1)}$, $I^{(2)}$ and $I^{(3)}$, then only cells with scale offset $s - t = 0, 1$ or 2 are of interest. (When $s - t = 0$, the estimated covariance reduces to WHO [7], *i.e.* within the local scale.)

3.1.3 LDA for multiscale input

Recall that the LDA weights are learned by $w = \Sigma^{-1}(\mu_1 - \mu_0)$. In the multiscale case, as shown in Fig. 3, μ_1 can be obtained by concatenating the HOG features extracted in multiple scales, μ_0 is the concatenation of background mean vectors, and Σ is the covariance matrix with diago-

nal blocks capturing the image statistics within the same scale, and off-diagonal blocks capturing the cross-scale image statistics (see Fig. 4).

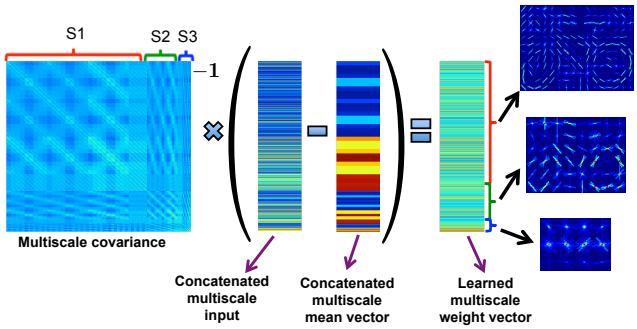

Figure 3. Training LDA for multiscale input. The input HOG features from multiple scales are first concatenated together as one long vector, then LDA with mean and covariance estimated for multiscale HOG statistics is applied to obtain the detector weights, which consist of weights for multiple scales.

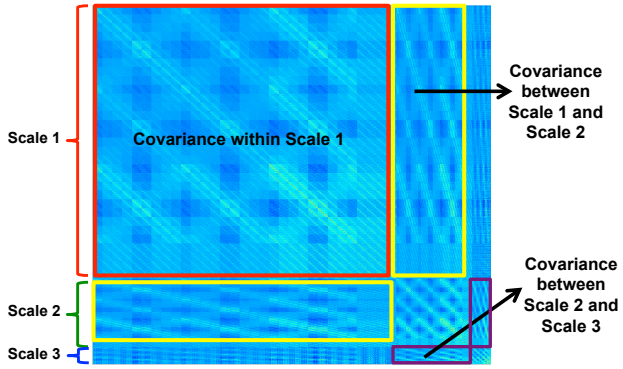

Figure 4. Structure of the multiscale covariance matrix. The diagonal blocks are covariances within the same scale, and the off-diagonal blocks are covariances between different scales.

3.1.4 Detection

At test time, detection is done using sliding window with the learned weights across multiple scales at the same time. Specifically, for each scale, we apply the standard sliding window approach with the weights for the corresponding scale to obtain a score map for all sub-windows. Then the score maps in higher scales are resized to align with that in the lowest scale using bilinear interpolation. The final detection score for each sub-window is obtained by simply adding up its scores from all the scales.

3.2. Statistics in the kernel induced feature space

For many computer vision feature representations, such as bag of visual words [2], spatial pyramids [8] and HOG, the intersection kernel [6] has been shown to outperform the linear kernel for detection and recognition tasks [9]. However, such performance improvement usually comes at great computational cost, since nonlinear kernels could induce much higher training and testing time for the SVM classifier.

In this section, we show that by utilizing an explicit feature mapping technique [16] one can directly estimate the statistics in the induced kernel space, and train object detectors efficiently using LDA in the feature space. This approach is equivalent to kernel LDA [11], except that we use an approximate finite dimensional feature mapping to work directly in feature space.

3.2.1 Explicit HOG feature mapping for the intersection kernel

The intersection kernel is part of a larger family called *additive kernels* [16]. For histogram-based feature vectors x, y (HOG descriptors in our case), an additive kernel is given by $K(x,y) = \sum_b k(x_b, y_b)$, where b is the bin index. The scalar kernel k is typically chosen to be a positive definite kernel, which implies that there exists a feature mapping $\Psi(x_b)$ such that $k(x_b, y_b) = \langle \Psi(x_b), \Psi(y_b) \rangle$, where $\Psi(x_b)$ is however possibly infinite dimensional. For homogeneous[1] positive definite kernel, it is also possible to construct an approximate feature map $\hat{\Psi}(x_b)$ of dimension only $2n+1$.

This kernel map has the generic form:

$$\frac{[\hat{\Psi}(x_b)]_j}{\sqrt{x_b L}} = \begin{cases} \sqrt{\kappa(0)}, & j = 0, \\ \sqrt{2\kappa(\frac{j+1}{2}L)\cos(\frac{j+1}{2}L\log x_b)} & j > 0 \text{ odd}, \\ \sqrt{2\kappa(\frac{j}{2}L)\sin(\frac{j}{2}L\log x_b)} & j > 0 \text{ even}, \end{cases} \tag{5}$$

where n and L are parameters controlling the approximation accuracy, and $\kappa(\lambda) = \frac{2}{\pi}\frac{1}{1+4\lambda^2}$ for the intersection kernel.

Combining this technique with the WHO pipeline, we can directly obtain natural image statistics in the new feature space induced by the intersection kernel, which are $\mathcal{O}(n^2)$ times as many as the original HOG statistics (in terms of the size of the covariance matrix). Specifically, this is achieved by adding an extra mapping step into the feature extraction process, where each 31-dimensional HOG cell x is mapped to the $(62n+31)$-dimensional space using Eq. 5. Given that a typical HOG template has more than 100 cells, the dimensionality will explode easily with a large n. Fortunately, as shown in [16], n needs not be a large value for the approximation to be reasonable.

During training, each cell of the input HOG descriptor is mapped using Eq. 5 to obtain its representation in the kernel induced space. Then LDA with mean and covariance

[1] $k(x,y)$ is homogeneous if $\forall c \geq 0, k(cx, cy) = ck(x,y)$, which holds true for the intersection kernel.

capturing the statistics within the new space is used to learn the detector weights w_K. At test time, all potential sub-windows generated by the sliding window technique are first mapped using Eq. 5, and the score for each sub-window is simply the inner product between the feature map $\hat{\Psi}(x)$ and the weights w_K.

3.3. Has the model capacity been increased?

For the two statistical models proposed in previous sections, which we call Multiscale-WHO and Kernelized-WHO , we conduct similar experimental evaluation as in Section 2.2 by varying the amount of image data used for covariance estimation. The results shown in Fig. 1 show that for both Multiscale-WHO and Kernelized-WHO, the performance saturation point is pushed much farther along the x-axis compared to WHO, implying that both are capable of utilizing much more data than WHO for estimating natural image statistics, and consequently lead to better detection performance. The improvement in detection performance is further verified in Section 4 by comprehensive evaluation on a standard benchmark dataset.

4. Experimental Results

We evaluate the two proposed approaches, Multiscale-WHO and Kernelized-WHO, respectively, on the standard object detection benchmark dataset, PASCAL VOC 2007. For baseline comparison, we also report the performance of WHO (Exemplar LDA + Calibration) and Exemplar-SVM on the same dataset.

4.1. Object detection with Multiscale-WHO

To estimate the mean and covariance for Multiscale-WHO, we use a large set of unlabeled natural images, the PASCAL VOC 2010 dataset. From our preliminary experiments, we found that the detection performance using two scales is indistinguishable with that using three scales, which might be because image gradients at the third scale (original image down sampled by a factor of 4) are much less correlated with gradients that are two scales lower (see Fig. 5). For simplicity, we thus decide to evaluate the performance of Multiscale-WHO with two scales for full PASCAL detection experiments.

After the multiscale background statistics are learned, we train an ensemble of multiscale exemplar detectors for each of the 20 object categories using LDA. At test time, each exemplar detector is used to create detection windows with the method described in Section 3.1.4 separately (this step can be dramatically expedited via parallelized implementation on a cluster of hundreds of nodes), and the results are pooled via a calibration step [10] to form final detection scores for each window. Finally, standard non-maximum suppression is applied to generate a sparse set of detections per image. Performance on the PASCAL test set for each

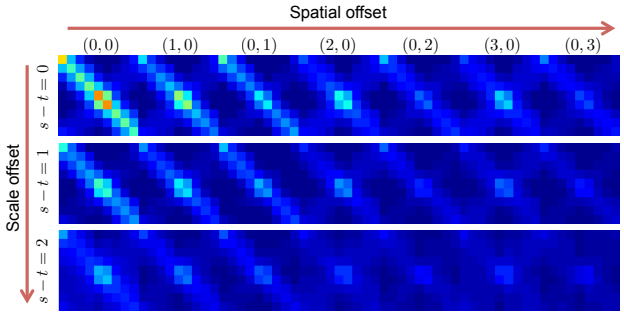

Figure 5. Visualization of covariances between the first 9 orientation features of HOG at different spatial and scale offsets. Lighter pixels have higher values. One can see that at scale offset two, the covariances start to appear not well-structured any more, and the intensity is significantly lower than smaller scale offsets. This indicates that image gradients at the original scale and ones at two scales higher are not well-correlated, and using the third scale for LDA training would not help much.

category is shown in Table 1. Overall, object detectors trained by Multiscale-WHO obtains a mean Average Precision (mAP) of .199, which not only performs better than WHO (.191), but does as well as ESVM (.198) that requires an expensive hard-mining step for training each exemplar.

4.2. Object detection with Kernelized-WHO

Same as for Multiscale-WHO, we estimate the mean and covariance of Kernelized-WHO on the PASCAL VOC2010 dataset. In our experiments, we set the kernel approximation parameter $n = 1$ such that a HOG cell of 31 dimensions is mapped to a 93-dimensional vector[2].

For evaluation, we also adopt the ensemble of exemplars pipeline for Kernelized-WHO. In particular, the weights for each exemplar detector are learned as described in Section 3.2.1, and at test time, the only difference is that an additional feature mapping step using Eq. 5 is applied to all sub-windows generated by the sliding window technique before multiplying the detector weights.

We also report the detection performance of Kernelized-WHO on the PASCAL test in Table 1. As shown, the average performance of Kernelized-WHO improves significantly compared to WHO (a relative 11% improvement), and it is even better than ESVM (a relative 8% improvement) without the need for a costly hard-mining step.

4.3. Qualitative results

In this section, we show qualitative results for further comparing Multiscale-WHO and Kernelized-WHO with the baseline WHO model, and hopefully shed some light on

[2]As shown in [16], a larger n typically does not lead to significant performance improvement, while the computational cost might be doubled or even quadrupled, and is thus unjustified.

what natural image properties are being captured from more extensive use of image data.

First, we would like to visually compare the weights learned by the three different models. Apparently, this is not straightforward for Kernelized-WHO as the learned weights are no longer in the HOG space. To resolve this, we apply least squares fitting to obtain an approximation of the kernelized weights in the HOG space. Specifically, we randomly sample 10,000 image patches as the training set for least squares fitting, and find the HOG weights that give best approximation to the kernelized weights in terms of their scores on the training set. The optimization objective is given by

$$\hat{w} = \arg\min_{w} \frac{1}{2} \sum_{x} \|\langle w, x\rangle - \langle w_K, \hat{\Psi}(x)\rangle\|^2 \qquad (6)$$

where w_K denotes the learned weights in the kernel induced space, x is the HOG descriptor for a random image patch, $\hat{\Psi(x)}$ is the feature mapping for x, and $\langle \cdot, \cdot \rangle$ denotes the inner product between two vectors. Obviously, this is not an ideal way to obtain the approximation, but it's sufficient to help us understand what Kernelized-WHO is capturing to some extent.

We visualize the learned weights using different models (WHO, Multiscale-WHO and Kernelized-WHO) for a variety of exemplars in Fig. 6. As highlighted in the figure, Multiscale-WHO tend to produce less noisy weights overall compared to WHO, which verifies our hypothesis that capturing image statistics more globally using multiple scales is indeed helpful. It's worth emphasizing that although for some cases the weights shown for Kernelized-WHO might appear more noisy than the other two, this does not necessarily mean Kernelized-WHO is inferior, since the weights shown are only an approximation of what is actually being learned in the kernel induced feature space.

We also show in Fig. 7 the top retrievals by different models using various query exemplars on the PASCAL test set. Not surprisingly, both Multiscale-WHO and Kernelized-WHO are able to retrieve detection windows that are more visually coherent with the query than WHO.

5. Discussion

At the start of this paper, we argued that the understanding of what works in object detection and why has gotten murkier. Has this manuscript been able to clear anything up? Hardly! In fact, we now have more questions than when we started (which is, alas, often the case in science). For example, how could it be that the original WHO system was almost beating the SVM (with its tens of thousands of negative images) after only 20 images worth of natural image statistics? How is it that even our two extensions saturate after about 200 images? Is it possible that there is little more in terms of image statistics that could still be fished out of the data? These are all intriguing questions that should be asked. We hope that this paper will be part of the conversation that tries to find answers to them.

References

[1] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, pages 129–136, 2010. 3

[2] G. Csurka, C. R. Dance, L. Dan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Stat. Learn. in Comp. VIsion*, 2004. 5

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1

[4] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 1, 2

[5] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010. 1

[6] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005. 5

[7] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *ECCV*, 2012. 1, 2, 4, 8

[8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 5

[9] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008. 3, 5

[10] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 1, 2, 6, 8

[11] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Mullers. Fisher discriminant analysis with kernels. In *IEEE Signal Processing Society Workshop*, pages 41–48, 1999. 2, 5

[12] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *ECCV*, 2010. 3

[13] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *IJCV*, 2006. 3

[14] H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005. 2

[15] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros. Data-driven visual similarity for cross-domain image matching. In *SIGGRAPH Asia*, 2011. 1

[16] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *CVPR*, 2010. 5, 6

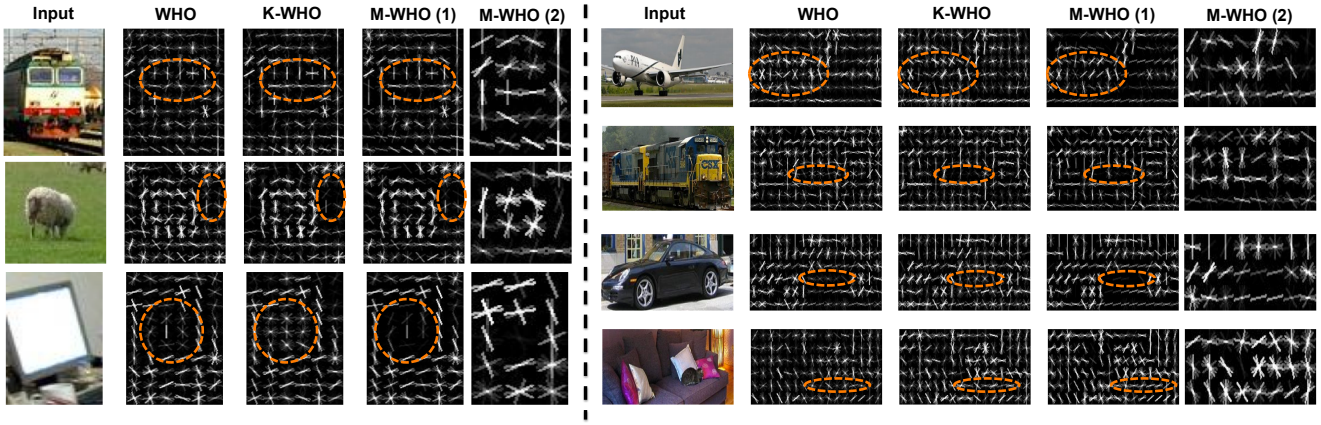

Figure 6. Visualization of HOG weights learned by different models. M-WHO(1) and M-WHO(2) are weights learned by Multiscale-WHO at two different scales. Note that HOG weights shown for Kernelized-WHO are only approximation of the actual learned weights in the induced kernel space. We highlight regions that appear differently across three models with dashed ellipses. Comparing M-WHO with WHO, we can see that weights learned by M-WHO tend to appear less noisy overall. Consequently, more weights are assigned to the discriminative parts of the exemplar, which supports our hypothesis that by looking at information from multiple scales one can capture the image statistics more globally.

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	mean AP
ESVM [10]	.204	.407	.093	.100	.103	.310	.401	.096	.104	.147	.023	.097	.384	.320	.192	.096	.167	.110	.291	.315	.198
WHO [7]	.184	.399	.096	.100	.113	.396	.421	.107	.061	.121	.030	.106	.381	.307	.182	.014	.122	.111	.276	.302	.191
M-WHO	.214	.403	.094	.096	.098	.329	.424	.107	.104	.155	.092	.103	.403	.306	.180	.009	.129	.123	.289	.326	.199
K-WHO	.232	.429	.092	.097	.109	.354	.433	.121	.092	.202	.098	.112	.422	.309	.190	.063	.155	.102	.342	.328	**.214**

Table 1. Detection results on the 20-category PASCAL VOC 2007 dataset. We compare the proposed approaches, Multiscale-WHO and Kernelized-WHO, with the Exemplar-SVM and WHO models. To make the comparison fair, we use the same ensemble of exemplars training/testing pipeline for all approaches (see Section 4.1 for more details on the setup). The performance numbers are in terms of Average Precision (AP). Overall, Multiscale-WHO is able to achieve a mean AP of .199, which is higher than WHO (.191) and comparable with Exemplar-SVM (.198). For Kernelized-WHO, the performance improvement is even more significant with a mean AP of .214, which is 11% and 8% better than WHO and Exemplar-SVM, respectively.

Figure 7. Top retrievals by different models for a variety of query exemplars. Overall, retrievals by Multiscale-WHO and Kernelized-WHO tend to look more visually coherent with the query than ones by WHO.