Looking outside of the Box: Object Detection and Localization with Multi-scale Patterns

Eshed Ohn-Bar, *Student Member, IEEE,* and Mohan Manubhai Trivedi, *Fellow, IEEE*

Abstract—Detection and localization of objects at multiple scales often involves sliding a single scale template in order to score windows at different scales independently. Nonetheless, multi-scale visual information at a given image location is highly correlated. This fundamental insight allows us to generalize the traditional multi-scale sliding window technique by jointly considering image features at *all scales* in order to detect and localize objects. Two max-margin approaches are studied for learning the multi-scale templates and leveraging the highly structured multi-scale information which would have been ignored if a single-scale template was used. The multi-scale formulation is shown to significantly improve general detection performance (measured on the PASCAL VOC dataset). The experimental analysis shows the method to be effective with different visual features, both HOG and CNN. Surprisingly, for a given window in a specific scale, visual information from windows at the same image location but other scales ('out-of-scale' information) contains most of the discriminative information for detection.

Index Terms—Object detection, localization, context modeling, multi-scale analysis, structured prediction.

1 Introduction

Object detection is commonly performed using models that are trained in a binary fashion, trained on information inside a local sliding window. This popular scheme is often applied over re-sampled versions of the original image in order to handle detection at multiple scales. This leads to evaluation of the trained model at each window and at each scale independently. Alternatively, a set of models corresponding to different scales may be used [1], [27]. A key disadvantage of such approaches is that it ignores the highly structured information over the differently scaled features or templates. Looking at Fig. 1, the information over scales is far from independent. Take for instance a car detection task, as shown in Fig. 1. Although only one of the scales best fits the car, different scales may contain cues important for the localization and detection of that vehicle in the best fitting scale, such as parts of the vehicle (the bumper, license plate, or tail lights) and contextual scene information (such as road cues). This is a strong motivation for 'looking outside of the box'-as classification can be hard even for humans when just looking at small cropped instances of objects. The multi-scale analysis can therefore provide additional information to the classifier. Consequently, we propose to formulate multi-scale detection as a multi-class problem, with a joint model reasoning over information in all the scales in order to produce a best fit for the detection. Our formulation generalizes the single-scale template case, since if information outside of the best-fit scale ('out-of-scale') is non-informative it can be simply ignored, thereby reducing to the traditional single-scale template case.

Contribution: This study presents the following,

1) Significant gains in detection performance can be obtained without altering the underlying descriptor but by replacing the traditional multi-scale pipeline with the proposed novel multi-scale approach.
2) To our knowledge, this work is the first to show that the majority of discriminative weight for detection at a given scale is found in features that are outside of the best fit scale. This is intuitive, as scale alignment cues can be found at all scales of the feature pyramid.
3) The improved model capacity results in several benefits not discussed in literature before. For instance, hard negatives mining with the proposed approach and histogram of orientated gradients (HOG) or convolutional neural network (CNN) features converges significantly faster. In contrast, other detectors employ many more rounds (over 10 rounds in the deformable part model (DPM) [10]).

2 Related Research

This work is related to techniques leveraging structural scene information and features or detector responses at multiple scales. Some related research studies are highlighted below. They can generally be categorized into two: those that consider visual descriptors extracted from multiple resolutions and those the employ score output of classifiers in order to better model contextual relationship.

Multi-scale feature reasoning: This includes the works of [7], [23], [24], which consider high and low resolution models for detection. Such techniques were used for better handling appearance variation due to scale. The feature pyramid is still converted to a scored feature pyramid which is resolved using Non-Maxima Suppression (NMS) [23], [24], or using a second independent layer of spatial/scale contextual reasoning over the output scores [7]. Here lies a main difference with our approach, which involves a single step for reducing the descriptor pyramid

• *E Ohn-Bar and M. M. Trivedi are with the Department of Electrical and Computer Engineering, University of California, San Diego, CA, 92093. E-mail: {eohnbar,mtrivedi}@ucsd.edu*

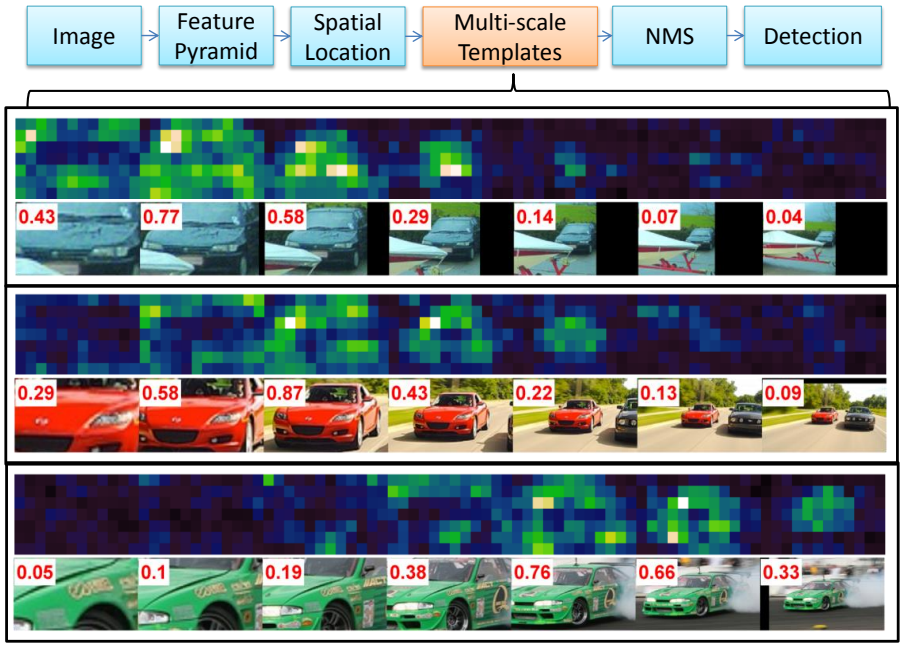


Fig. 1. Pipeline of the proposed, multi-scale structure **MSS**, approach. Traditional object detectors train and test models at a single scale, thereby ignoring information over scales at a specific spatial location. This paper studies methodologies for replacing this multi-scale mechanism by integrating features over all scales into a structural detection problem. Using the framework, windows can be scored over both detection and localization quality. The study is motivated by the fact that cues outside of the sliding box are essential for detection and localization of the object inside the box. Furthermore, multi-scale information is highly structured, as shown in the above example images of car, hence localization quality can benefit from multi-scale analysis. Cues at different scales, such as the road surface or car parts, provide detection and localization information (as shown in the example multi-scale templates above, brighter implies more discriminative value). Finally, as appearance varies with respect to scale, it is shown to be better modeled by a scale-specific model.

into a score map of a single scale and not a score pyramid. The proposed approach employs feature at all scales, remote and adjacent, as opposed to one selected additional higher or lower resolution. NMS is still used, but the scores already incorporate localization and contextual reasoning. Furthermore, the proposed approach explicitly accounts for variation in appearance due to scale by learning a template filter for each label class. The aforementioned studies generally do not alter or generalize the traditional (Fig. 2), sliding window at multiple scales technique, as we attempt to do in this work. The multi-resolution framework of [34] involves rejecting windows at low resolutions before the rest of the image pyramid is processed, thereby achieving speed gains. Recently, some studies proposed using a template pyramid approach instead of a feature pyramid [1], [27], which may also be able to capture variation in appearance due to scale. Nonetheless, the motivation here is for speed/memory consideration with little impact on performance as shown in these studies. The proposed approach can also be seen as a generalization of such approaches, as will be discussed in Section 3.

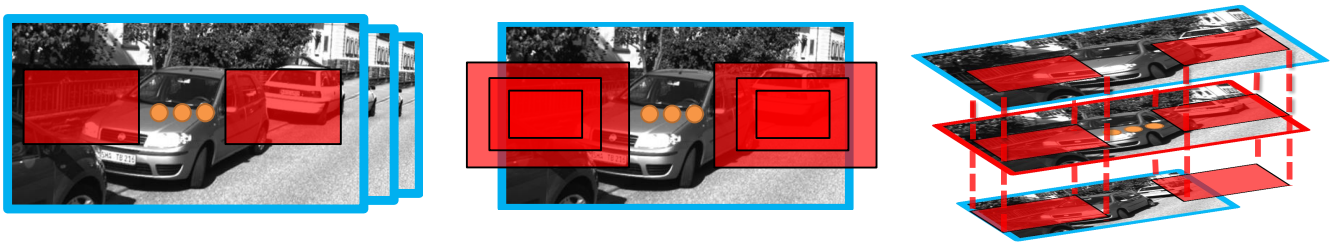
Contextual re-scoring: Another common approach is to an additional module in the detection framework which captures spatial and scale contextual interactions, applied over the score pyramid output of a traditional sliding window multi-scale detector. This re-scoring technique has been the general trend, as researchers tend to employ multiple, independent modules in detection and localization (as in [10], [13], [20]). In [5], a Multi-Order Contextual co-Occurrence (MOCO) framework was proposed, extending the Auto-Context idea [25], [30] for context modeling among boxes produced by traditional object detection schemes. Structured prediction has been used in [6], [26], [33] for capturing spatial arrangement statistics.

Localization: Structural max-margin classifiers were shown to be powerful tools for the object localization task. The studies of [2], [16] employ a structural SVM for the localization task. These may be limited in being able to detect only a single object instance in a given image, unlike the proposed approach, but are a noteworthy effort in attempting to incorporate the detection and localization tasks into a single model. This paper performs a comprehensive study in replacing one element of the well established multi-scale detection and localization pipeline, as a step towards training models which reason over the two tasks jointly.

Training with hard negatives mining: A large portion of existing detection schemes involve an iterative process for obtaining hard negative samples and re-training the classifier [10]. Latent linear discriminant analysis was proposed in [15] to significantly reduce time spent in the data mining. Nonetheless, mining is still required to obtain matching performance with the Latent SVM baseline. On the contrary, the proposed approach is shown to outperform the baseline even without data mining at all. Generally, only one iteration of hard mining is required for convergence.

3 The Multi-scale Structure (MSS) Approach

The outline of the proposed approach is shown in Fig. 3. The input to the proposed approach is the input to a traditional single scale template procedure with a main difference, which we aim to study as thoroughly in this paper. Instead of classifying a set of features at a single scale at a given image location, the proposed



(a) Traditional, single-scale approaches (image pyramid and template pyramid) (b) Our proposed approach

Fig. 2. Contrasting the proposed MSS approach with existing approaches. MSS generalizes traditional object detection techniques, as it learns a discriminative weight vector for each scale, and does so jointly over all possible scales. As an input, it takes image features at multiple scales, as opposed to the traditional, single-scale approaches shown in (a). Our analysis shows that for a detection at a given scale, image features at both adjacent and remote scales provide large discriminative support, resulting in improved detection performance. A main reason for this is due to contextual and localization information across scales.

approach classifies the features in all pyramid scales at a given location jointly.

3.1 Multi-scale detection with a single-scale template

The detection baseline employ a pyramid of features and fixed size template in order to handle detection at multiple scales (see Fig. 2). First, for efficiency, a feature pyramid is constructed over the entire image at each scale to avoid redundant computation for each striding window. Let $p_s = (x, y, s)$ be a window in the s-th level of a feature pyramid with S scales anchored in the x, y position. Generally, the feature pyramid is at a lower spatial resolution then that of the image of the same scale. Consequently, a zero-based index (x, y) in the feature map can be mapped to a pixel in the original image using a scale factor (cx, cy) based on the resolution of the feature map. Mapping locations over scales can be achieved by a multiplication by the scale factor as well.

Each window contains an array of feature values, $\phi(p_s) \in \mathbb{R}^d$, to be scored using a filter w learned by a discriminative classifier, in our case a support vector machine (SVM). The scoring is done using a dot product,

$$f(p_s) = w \cdot \phi(p_s) \tag{1}$$

Generally, the template size is defined as the smallest object size to be detected, and further reduction in template size results in degradation of the detection performance. Note that learning and classification only occurs over a local window. A similar pipeline can be described using a template pyramid as studied in [1], [24], [27] and was shown to improve results due to capturing finer features at different scales that would have been discarded by the down-sampling. In this approach, a set of templates are learned, $(w_1, \ldots, w_S)$, and output scores are calibrated (in [24], one model is learned over two resolutions so no calibration is required). In detection, the S templates are evaluated so that each location p in the original image scale is scored using the set of model templates

$$f(p) = \max_{s \in \{1,\ldots,S\}} w_s \cdot \phi(p) \tag{2}$$

where we drop s as only one scale of the image is considered. We emphasize that the model filters in this approach are also trained on locally windowed features only, but may capture different cues for each scale. In principle, this approach is similar to the baseline as it performs the convolutions at each scale independently of all other scales (unlike ours, as shown in Fig. 3).

3.2 Multi-scale detection with a multi-scale template

:

The feature pyramid computation and handling is mostly left unchanged in the proposed MSS approach. Spatial locations in the image space can be mapped across scales using a scale factor. As shown in Fig. 1, evaluations at the same spatial location occur repeatedly over scales. This mechanism is replaced by considering features from all scales at a given image location, i.e. $\psi(p) = (\phi(p_1), \ldots, \phi(p_S)) \in \mathbb{R}^{d \times S}$ descriptor.

3.2.1 Label space

Next the process of labeling positive samples is outlined. Each sample is assigned a label, $y = (y^l, y^b, y^s) \in \mathcal{Y}$ with y^l the object class (in this study only $y^l \in \{-1, 1\}$ is considered), $y^b \in \mathbb{R}^4$ is the object bounding box parameters, and y^s is a scale label. In our experiments, the model dimensions are obtained from the the average box size of all positive instances in the dataset. This choice is arbitrary, but it provides the general size of the object in all of the original image scales. The scale label is obtained by: 1) placing the average-sized model box centered on top of the positive sample bounding box at each scaled version of the image pyramid. 2) Computing the overlap of that placed box with all of the ground truth bounding boxes for each scale. Formally, a vector of overlaps F is constructed. If the image at s-th level contains $\hat{y}(s) = \{\hat{y}_1(s), \ldots, \hat{y}_N(s)\}$ ground truth boxes, the template box is centered on a positive sample at the s-th level (denoted as $B(s)$), so that entries of F are computed for each pyramid level,

$$F(s) = \max_{i \in \{1,\ldots,N\}} \text{ov}(B(s), \hat{y}_i^b(s)). \tag{3}$$

where ov(a,b) = area($a \cap b$)/area($a \cup b$) for two rectangles, a and b. F is shown for three examples in Fig. 1. Finally, peaks are found in $F(s)$, and entries that have a higher overlap than 0.6 are set to 1, while the rest are set to 0. For instance, for Fig. 1 first row, $y^s = (0100000)$. The purpose here is to maintain as much as generality as possible in the formulation. This process potentially allows for multiple labels over scales to be predicted jointly, i.e. two almost overlapping objects at different scales or two spatially adjacent instances as shown in Fig 4.

3.2.2 Learning

The problem can now be posed as a multi-class problem.

One-vs-All: There are well developed machine learning tools for dealing with a large-dimensional multi-class classification problem. A straightforward solution is with a one-vs-all (OVA)

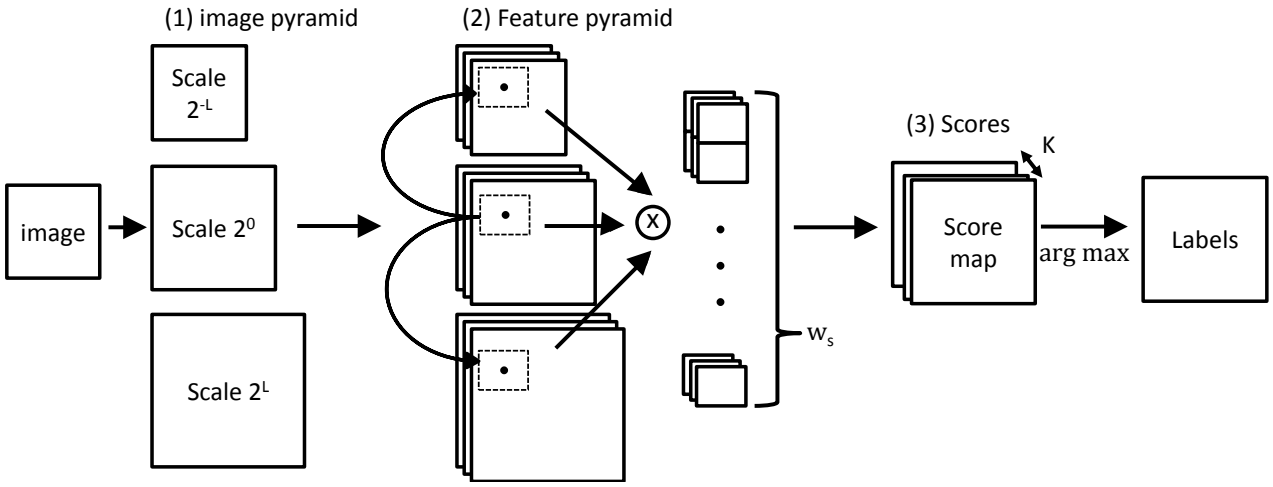
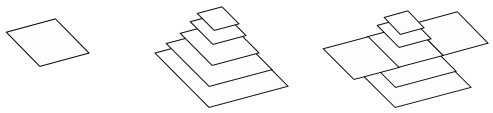


Fig. 3. Our proposed approach takes features at all scales for training template filters for a given object class. The template which best fits the image evidence provides the final scale alignment. Note that the feature maps are usually at a lower spatial resolution then the original images.



(a) Traditional (b) Scale (d) Scale+Spatial

Fig. 4. Options for the predicted label space allowing for reasoning over the scale and spatial configuration. (a) single scale (traditional detector) (b) scale labels for one or multiple objects in a single spatial location (c) a joint scale and spatial layout. In (c) adjacent objects also involvse a scale label prediction, although not shoswn for clarity.

SVM, which allows training the multi-class templates quickly and in parallel. Window scoring is done using

$$f(p) = \max_{s \in \{1,\ldots,K\}} w_s \cdot \psi(p) \tag{4}$$

The scale of the box is obtained with an $\arg\max$ in Eqn. 4. In order to learn the K linear classifiers parameterized by the weight vectors $w_s \in \mathbb{R}^{d \times S}$, the stochastic dual coordinate ascent solver of [31] with a hinge loss is used. The maximum number of iterations is fixed at 5×10^6 and the tolerance for the stopping criterion at 1×10^{-7} for all of the experiments. Training a single multi-scale template on a CPU on average takes less than a minute.

Structured SVM: A second common approach can be used in order to learn all of the multi-scale templates jointly. A feature map is constructed using the labels of each sample as following,

$$\Phi(p, y) = (\Psi_1(p, y), \ldots, \Psi_K(p, y)). \tag{5}$$

$$\Psi_k(p, y) = \begin{cases} \psi(p) & \text{if } y = k \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

This approach allows for learning a joint weight vector over all classes $w = (w_1, \ldots, w_K)$, such that

$$f(p) = \max_{s \in \{1,\ldots,K\}} w \cdot \Phi(p, s) \tag{7}$$

The formulation is kept general so that K may or may not equal S.

Given a set of image-label pairs, $\{x_i, y_i\}$ the model is trained using a cost-sensitive SVM objective function [3], [17], [19]

$$\begin{aligned} &\min_{w,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i \\ &\text{s.t. for } \forall i, \bar{y} \in \mathcal{Y} \setminus y_i \\ &w \cdot (\Phi(x_i, y_i) - \Phi(x_i, \bar{y})) \geq \bar{L}(y_i, \bar{y}) - \xi_i \end{aligned} \tag{8}$$

The loss function, $\bar{L}$, is chosen to favor large overlap with the ground truth. Since the formulation and label spaces (Section 3.2.1) consider all the ground truth boxes in the current image, $\hat{y} = \{\hat{y}_1, \ldots, \hat{y}_N\}$, the method supports prediction of multiple boxes for a given image location. First, we define a per-box loss, L, as following

$$Ł(y, \hat{y}) = \begin{cases} 0 & \text{if } y^l = \hat{y}^l = -1 \text{ or} \\ & \max_{i \in \{1,\ldots,N\}} \text{ov}(y^b, \hat{y}_i^b) < 0.6 \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

The training loss function is then taken as an average loss over each of the predicted boxes, $\bar{L} = \frac{1}{M}\sum_{j \in \{1,\ldots,M\}} L(y_j, \hat{y})$, for M predicted boxes.

3.2.3 *Generalization of the single-scale approach*

The main difference in the proposed MSS approach and the baseline is in the training of the detection model. Eqns. 4 and 7 employ features at all scales for a given spatial location. Such a formulation allows learning the class weights jointly, as in Eqn. 7. It can be shown that this is a generalization of the single-scale template baseline. For instance, if no discriminative value is added by adding features at different scales, then the corresponding weights w_s in Eqn. 4 will only select features in the single best-fit scale (i.e. a degenerate case). Therefore, for each level s in the pyramid, $w_s \cdot \psi(p)$ becomes identical to $w \cdot \phi(p_s)$ as in Eqn. 1. A similar argument demonstrates the same for Eqn. 7. Therefore, both of the studied multi-scale template learning approaches can benefit by having access to additional information not accessible to the single-scale template approaches which only employs local window features at one scale. Furthermore, by learning a separate weight for each class, the model can account for appearance variations at different resolutions [1].

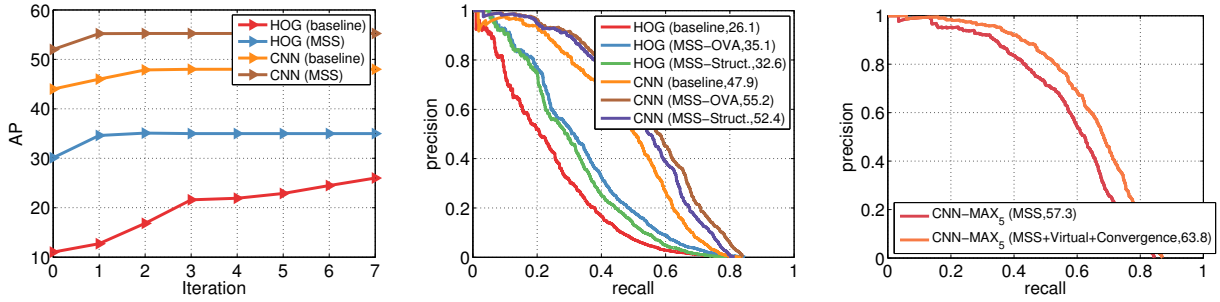


Fig. 5. (a) Analysis of the proposed MSS approach on the car category with the two types of features studied and the two techniques proposed, One-vs-All and Structurally trained multi-scale templates. Average Precision (AP) is shown in parenthesis. (b) AP throughout the iterative hard samples mining. Note how the MSS approach provides a greater modeling capacity, with high detection performance compares to the baseline even without hard negative mining but using an initial random set. (c) For the final results on the car category, results after applying a 3×3 max filter to the conv_5 features, additional positives from re-sized samples, and mining until convergence are shown.

3.2.4 *Challenges with the MSS approach*

The main issue is in the large dimensionality of the feature vector as each scale label is associated with a concatenation of the features at all scales. This comes with a benefit, as feature selection techniques can be used in order to choose features over all scales at once. Although training becomes more memory intensive, this is the reason for studying two approaches for multi-scale modeling. For instance, with the OVA approach training a MSS template is a matter of minutes.

4 Experimental Evaluation

Dataset: In this section the two studied multi-scale template approaches are compared against the baseline. Initially, a comprehensive analysis is performed on the car category of the PASCAL VOC 2007 dataset [9]. Generalization of the proposed approach will then be studied on other detection categories of the challenging dataset. Evaluation is done using a 50% overlap requirement.

Features: Two representative visual descriptors are employed, the HOG implementation of [11] which is still widely used and serves as a classical baseline, as well as newer, more rich CNN features [13], [14]. Several CNN-based detectors still employ a sliding window approach [28], [29]. In this paper, an off-the-self network [18] is employed without any fine tuning, trained on ILSVRC 2012. The fifth convolution layer output has 256 feature channels. The input to each convolutional or max pooling layer is zero-padded so that the features in a zero-based pixel location (x, y) in the feature space were generated by a receptive field centered at $(16x, 16y)$ in the image space (a stride of 16). As noted by [14], the CNN features already provide part and scale selective cues. This can be enhanced by applying a 3×3 max-pooling layer. For direct comparison with [14], the same feature extraction and pyramid pipeline was implemented. HOG is used with a cell size/stride of 8.

A single rigid template: The experimental analysis validates the significance of the proposed approach with a single aspect ratio rigid template. This is done in order to validate the benefit of using our approach against the traditional baseline. Although the performance improvement due to incorporation of parts and multiple aspect ratio components [10], [22] is noteworthy, a large number of studies still employ a single rigid template in detection [8], [21]. Parts and aspect ratio components are seen as orthogonal to our approach, which emphasizes contextual information.

Image Pyramid: The scale factor between levels is set to $2^{-1/2}$. The CNN feature pyramid spans three octaves with 7 levels. For HOG features, adding 3 more levels to the image pyramid for a total of 10 was shown to improve performance. Furthermore, for the baseline single-scale HOG detector, replication padding of the image was shown to work best. For the multi-scale template approach, the HOG feature pyramid is 0-padded which was experimentally shown to work better.

Hard negative mining: Training object detectors usually employs an iterative process by which hard negatives are collected for re-training. The process eventually converges, when the number of negative samples generated are below a certain threshold. For HOG, up to 5 iterations are allowed. For CNN-based features, most of the experiments employ a single mining iteration (also done in R-CNN [13]), unless otherwise stated. All of the experiments begin with a random set of 5000 negative samples. For a given object category, the initial negative samples are kept the same across techniques to allow direct comparison. In each iteration, up to 5000 additional negatives are collected. A mining threshold which provides a high recall of the positive samples was used, as in [10]. For mining, both images containing positive instances and negative images are used. A threshold of 0.3 overlap is used for mining negative samples from images with object instances. When mining with the proposed approach, only negative samples which are negative to all of the K MSS classes are collected.

4.1 PASCAL Car

Our experiments initially were on the car category, as it contains many samples with large variation in scale. A study of the generality of the MSS approach to other categories is done in Section 4.2.

A significant improvement in performance is shown for both HOG and CNN features due to the MSS approach. Generally, HOG+MSS showed greater gains than CNN+MSS as the latter features have been shown to be more scale-invariant in recent studies [12], [14]. Nonetheless the improvement is significant in both types of features. One reason for this could be the incorporation of contextual features with the MSS approach. Fig. 5 studies the different proposed techniques and HOG and conv_5 CNN features. For HOG features, with a fixed set of positive and negative samples, the HOG-MSS approach converges quickly and achieves high detection performance even before mining hard

negatives (random negatives at iteration 0). This is remarkable, as the baseline does not fully converge even after many rounds of mining. The plot clearly demonstrates the improved modeling capacity of the proposed approach. Employing the conv_5 CNN features results in fast convergence in both the baseline and the proposed approach, yet the proposed approach still converges faster, with significantly less added hard negatives per round.

Which modeling approach?: Fig. 5(b) details the analysis of different learning and features combinations on the 'car' category. Interestingly, for this object category the improvement when using either type of features is consistent with the proposed approach. Although the structural SVM allows an explicit incorporation of localization loss and joint learning of the templates, the One-vs-All solution is shown to perform well. As both approaches employ the same multi-scale features, the improvement in both is to be expected. As each multi-scale template is specific to a given scale (see the visualization in Fig. 9) little benefit came from learning several multi-scale templates for prediction at different scales jointly. Furthermore, due to the larger memory demands in training the structured SVM template, a smaller number of negatives can be kept in each mining iteration. Nonetheless, the structural SVM formulation is of interest as it may allow for more sophisticated bonding box regression [2], latent variable incorporation, parts integration, etc. Both approaches show promise when compared to the single-scale template baseline.

Label space choice: As discussed in Section 3.2.1, there are multiple choices for setting the output of the MSS approach. Because the approach allows for prediction of multiple boxes, it can be immediately generalizable joint detection of adjacent spatial locations. Two nearby cells, east and west of the sliding template and overlapping it at 50%, are added. Other than the fact that the multi-scale features in all of the three windows are now used, the framework remains unchanged. On the 'car' category, this additional complexity in the label space resulted in a small 1 AP improvement. The small gain can be explained by the fact that spatially adjacent objects within the same class are somewhat rare. Furthermore, the multi-scale templates alone already capture a great deal of spatial context.

Comparison with state-of-the-art: A noticeable gain was shown when employing a 3×3 max filter on top of the conv_5 features as in [14]. With these max_5 features, the performance with CNN+MSS reaches 57.3 AP. For comparison, the approach of [14] which employs the same features reports 56.5, yet two main differences are important to note. First, [14] employs a model with three aspect-ratio components whereas CNN+MSS employs a *single aspect-ratio component*. Furthermore, as already mentioned, the scale sensitive models of CNN+MSS are exposed to a fraction of the positive samples, as only positive instances falling within the label class are included. This is not the case in the baseline and comparison methods, where all positive images are re-sized to a fixed size and can be used for learning a car template. Initially, re-sizing the image for additional positive samples was not required to get good results on the car category as shown in Fig. 7, but was shown to be useful when other object categories were studied (see further discussion in Section 4.2). Addition of samples from re-sized versions of the original image results in slower mining convergence, yet the fully converged MSS model provides a final AP of 63.8, which improves over R-CNN's [13] 60.6 AP and by 7.3 points over [14]. As discussed in [14] the feature computation is different in nature for R-CNN and also not efficient (per region as opposed to per image), hence these results are encouraging. Our results nearly match the stronger fine-tuned version of R-CNN and pool_5 features of 66.9 AP [13].

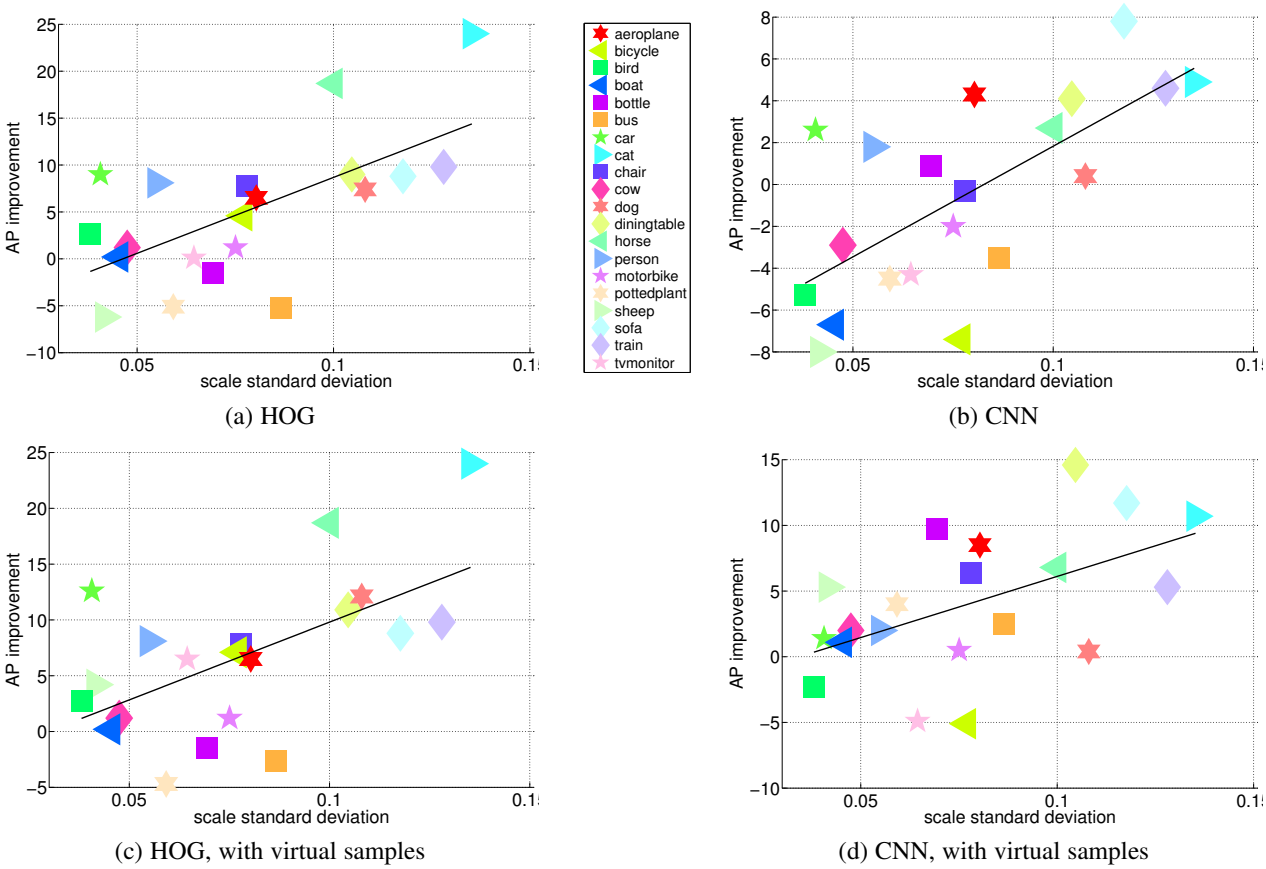
For HOG+MSS, the results are also promising with an increase in performance of 9 points without incorporation of parts, aspect-ratio components models, or re-sized versions of the original image. The latest version of HOG-DPM with 6 aspect ratio components (and template mirroring) achieves 46.3 and with 8 parts 58.2. HOG+MSS achieves 35.1 and 42.5 without and with addition of re-sized image samples, respectively. The addition due to parts and aspect ratio components is complementary to ours. This is proved by the fact that CNN features were shown to implicitly model part appearances (in [14], part addition over CNN features provided no improvement on the car category and a 1.1 AP improvement overall on PASCAL). Hence, the improvement due to MSS+CNN implies that the multi-scale models capture complementary cues to parts. Aspect-ratio components allow for better modeling variation in appearance due to orientation (as opposed to contextual information). A study of their incorporation into MSS is left for future work.

4.2 Results on All of PASCAL Object Categories

While most of our experiments were performed on the 'car' category, generalization to other object classes is also of interest and will be studied next. With the goal of revealing insights into the improvement provided by the MSS approach, several sets of experiments were devised. Throughout all of the experiments, HOG-based models use all of the training set (both images containing object instances and images without) and up to 5 hard mining iterations. For CNN, one mining iteration is commonly used (as in R-CNN [13]). For CNN, the mining is performed over a random set of 200 negative images (without object instances) together with all the positive images in each category. Since the max_5 features were shown to generally improve over the conv_5, these are used in all of the remaining experiments.

When does it work and when does it fail?: Our experiments showed the MSS method to significantly impact performance in some classes, yet fail to show a considerable improvement on others. A closer inspection of the scale distribution of the different classes reveals some insight, as shown in Fig. 6(a)-(b). First, CNN features are sensitive to scale which results in a smaller improvement. On the other hand large gains are consistently shown for the majority of the classes when using HOG features. Second, some classes exhibit smaller variation in scale which also translates to smaller AP improvement. Most importantly, the label space generation in Section 3.2.1 implies that many positive samples will be excluded in training the MSS templates. On the other hand the baseline method benefits being exposed to all positive samples at once when learning the appearance model of an object class. Hence, no AP improvement due to MSS can occur for several reasons, from not enough training data to small distribution of scale in test time. This is addressed in a second set of experiments, as detailed below.

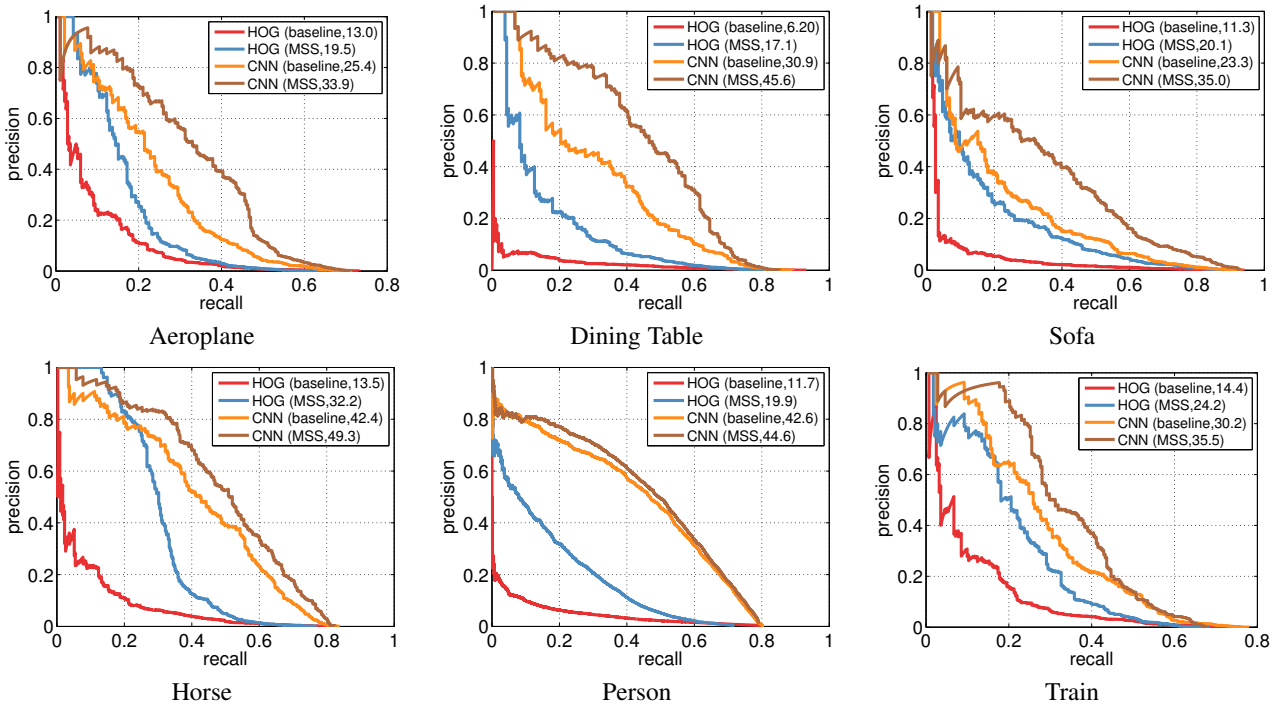
Virtual positive samples: As shown in Fig. 6, some scales are poorly represented in training and testing images for certain classes. In order to avoid a small number of positive samples for training a MSS template, positives are added at 5 additional scales. This adds virtual positive samples (re-sized versions of the original image), and emulates the re-sizing of all samples to a single scale in the traditional single-scale template approach. Addition of re-sized images resolved the issue for the most part,



(a) HOG (b) CNN

(c) HOG, with virtual samples (d) CNN, with virtual samples

Fig. 6. Relationship between the scale distribution of class samples in test time and the corresponding improvement in AP with the proposed MSS approach. As shown, our method shines when there is a large spread in the distribution over scales. Although some classes tend to appear in the PASCAL VOC dataset in a narrow scale distribution, this phenomenon is dataset and object specific. Therefore, if more instances at varying scales were to be added, the proposed approach would be better suited for such settings.



Aeroplane Dining Table Sofa

Horse Person Train

Fig. 7. Detection performance for selected classes of the 2007 PASCAL VOC dataset.

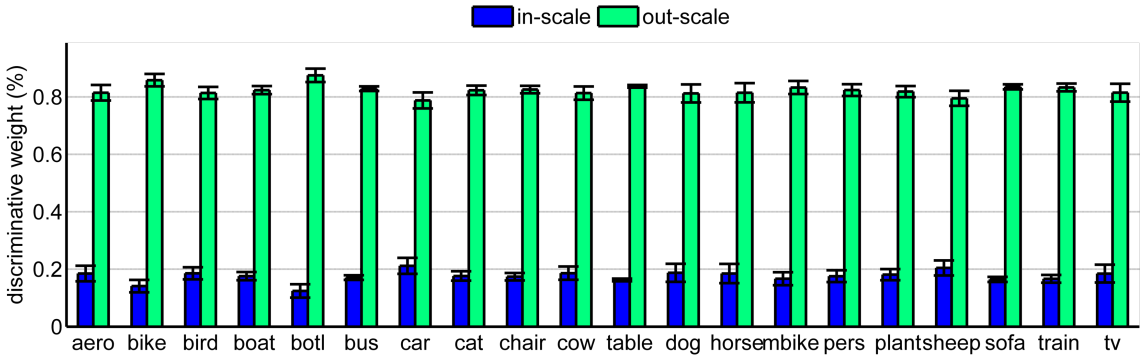


Fig. 8. For a detection at a given scale, how important are out-of-scale features? The model parameters were learned using CNN features for each object class. Next, the percentage of positive model weights that are within the best fit-scale are averaged over each of the in-class multi-scale templates and compared to the percentage of discriminative positive weights given to features which lie outside of the best-fit scale. The average and standard deviation for each are plotted. Remarkably, most of the discriminative value is found outside of the ground truth scale, motivating the use of a multi-scale filter.

as shown in Fig. 6(c)-(d), but came at a cost of longer convergence for the hard negative mining process. We note that the purpose in this set of experiments was to validate the MSS approach for different object categories, which is clearly shown. Although large gains can be made by adding virtual samples at all possible scales (which would require more memory in training) and mining until full convergence over the entire training set (see Fig. 7(c)), further tuning the MSS system is left of future work.

Data-driven confirmation of the proposed MSS approach: For each class, features were divided into two. First are those that are found in the best-fit scale correspond to the same features that would be employed if a single-scale template (referred to as 'in-scale' features). Second, 'out-of-scale' features which lie outside of the best-fit scale. This allows us to quantify the improvement in detecting and localizing due to incorporation of out-of-scale features. The learned parameters w can be decomposed to positive and negative entries as $w = w^+ + w^-$. Indices with higher absolute value correspond to locations in the feature space which provide large discriminative value. By studying the values in w^+ for MSS+CNN, Fig. 8 demonstrates the clear trend of choosing features that are placed outside of the ground truth scale in training. This is a data-driven affirmation of the proposed approach. Although only positive weights shown in Fig. 8, the trends are similar both over positive weights w^+ and negative weights w^-.

Summary of results on PASCAL VOC 2007: On the entire PASCAL dataset, incorporation of HOG+MSS results in a significant 6.6 mAP points improvement over the HOG baseline. The final mAP 19.9 using MSS-HOG with virtual samples and mining until convergence. For CNN+MSS, initially the overall improvement is small as some classes benefit greatly but others suffer due to having a small number of positive instances for training each of the MSS templates. With the addition of virtual samples, a significant 4 mAP point improvement is gained for an overall 43.7 mAP with a single aspect ratio component.

5 Concluding Remarks

This paper proposed a generalization of the traditional single-scale template detection approach. Training single-scale templates considers features only in a local box for a binary task of detection. Re-formulation of the problem as a multi-class classification problem, allowed the study of models which are trained to reason over both detection and localization cues. These significantly improved detection performance compared to their single-scale template counterparts. In the future, bounding box regression within the structural SVM could be studied [2], [13]. Training multiple aspect ratio components [32] could also provide further gains in detection performance. Feature selection over scales could significantly reduce the dimensionality of the problem and allow for faster detection. Finally, employing a fine-tuned or stronger CNN (VGG [4]) could also provide further improvements in detection.

6 Acknowledgments

The authors would like to thank Zhuowen Tu for helpful discussions.

References

[1] R. Benenson, M. Mathias, R. Timofte, and L. V. Gool. Pedestrian detection at 100 frames per second. In *CVPR*, 2012. 1, 2
[2] M. Blaschko and C. Lampert. Learning to localize objects with structured output regression. In *ECCV*, 2008. 2, 7
[3] S. Branson, O. Beijbom, and S. Belongie. Efficient large-scale structured learning. In *CVPR*, 2013. 4
[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional networks. In *BMVC*, 2014. 7
[5] G. Chen, Y. Ding, J. Xiao, and T. X. Han. Detection evolution with multi-order contextual co-occurrence. In *CVPR*, 2013. 1
[6] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 95(1):1–12, 2011. 2
[7] Y. Ding and J. Xiao. Contextual boost for pedestrian detection. In *CVPR*, 2012. 1
[8] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *PAMI*, 2014. 4
[9] 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. 4
[10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010. 1, 2, 4
[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Analysis and Machine Intelligence,*, 2010. 4
[12] R. Girshick. Fast r-cnn. *arXiv*, 2015. 4
[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1, 4, 6, 7
[14] R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. *CVPR*, 2015. 4, 5, 6
[15] R. Girshick and J. Malik. Training deformable part models with decorrelated features. In *ICCV*, 2013. 2
[16] M. Hoai, L. Torresani, F. D. la Torre, and C. Rother. Learning discriminative localization from weakly labeled data. *Pattern Recognition*, 47(3):1523–1534, 2014. 2

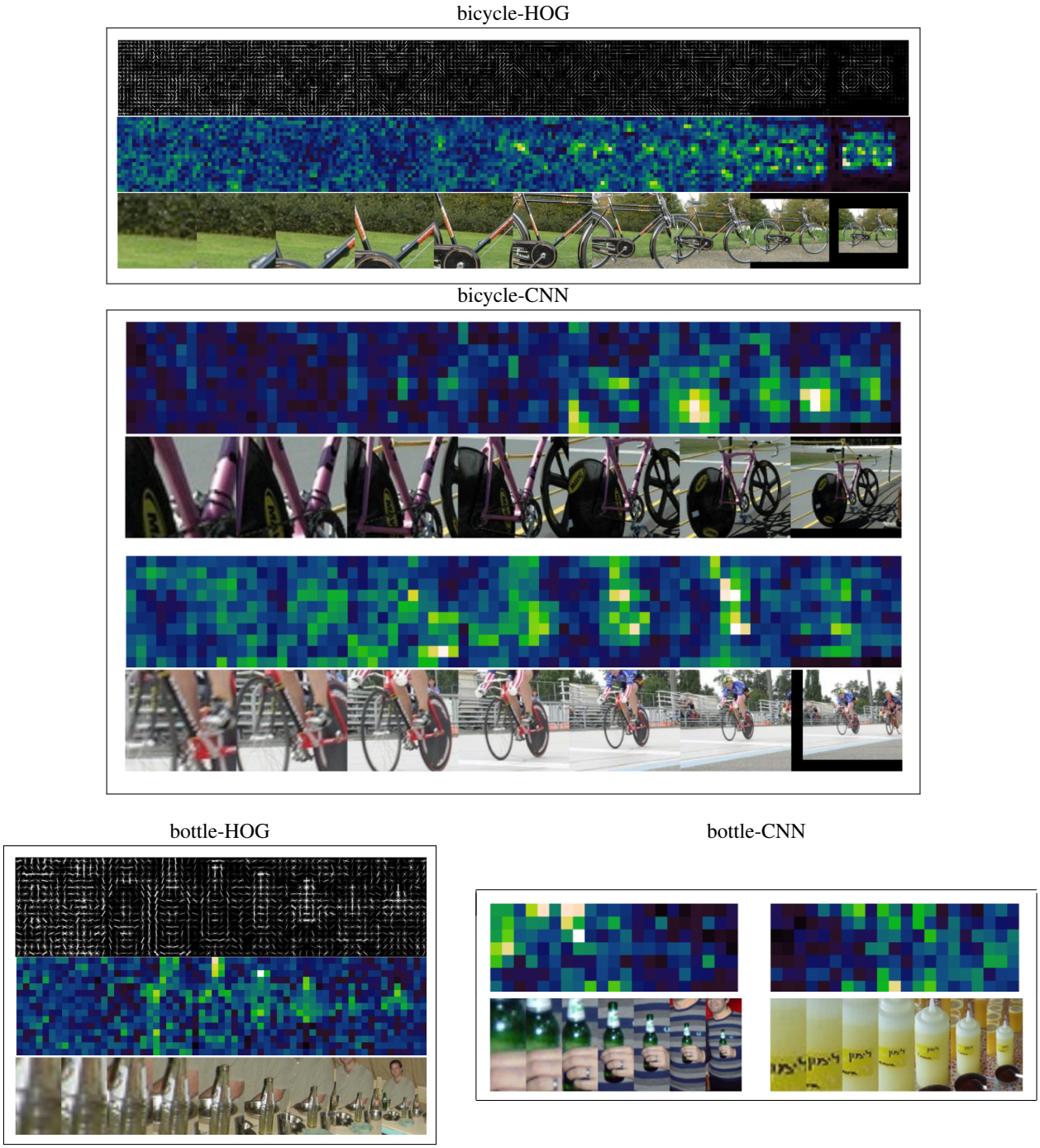
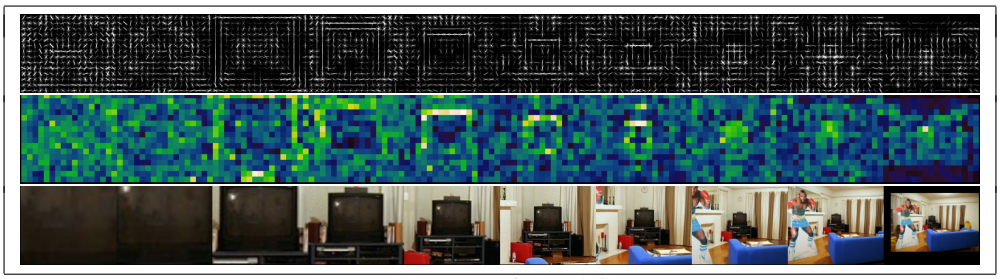
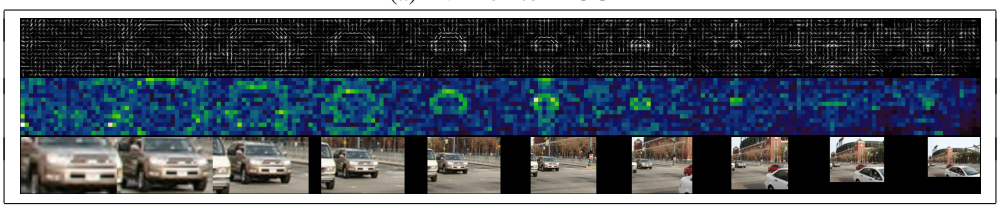


Fig. 9. Visualization of multi-scale HOG and CNN templates. For each model, For each spatial location the maximum positive SVM weight for each block is shown together with an example instance. Brighter colors imply higher discriminative value. Note the large amount of discriminative value is chosen in nearby and remote scales corresponding to contextual information. For instance, for the detection and localization of a bike at a given scale, there may be a vertical structure corresponding to the bike rider at another scale.

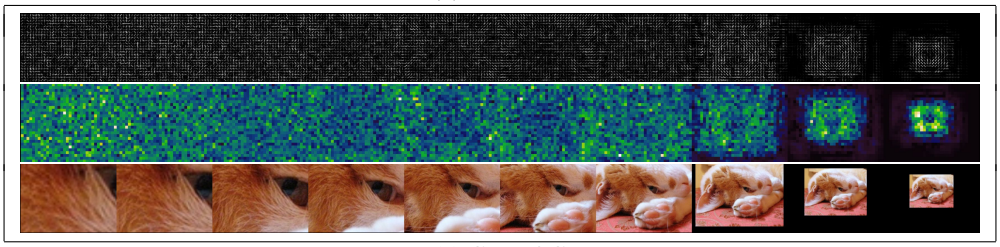
[17] T. Joachims, T. Finley, and C.-N. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59, 2009. 4
[18] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 4
[19] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *ICML*, 2013. 4
[20] C. Long, X. Wang, G. Hua, M. Yang, and Y. Lin. Accurate object detection with location relaxation and regionlets relocalization. In *ACCV*, 2014. 1
[21] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014. 4
[22] E. Ohn-Bar and M. M. Trivedi. Fast and robust object detection using visual subcategories. In *CVPRW*, 2014. 4
[23] E. Ohn-Bar and M. M. Trivedi. Learning to detect vehicles by clustering appearance patterns. *T-ITS*, 2015. 1
[24] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *ECCV*, 2010. 1, 2
[25] L. Quannan, J. Wang, Z. Tu, and D. P. Wipf. Fixed-point model for structured labeling. In *ICML*, 2013. 1
[26] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 2
[27] M. A. Sadeghi and D. Forsyth. 30Hz object detection with DPM V5. In *ECCV*, 2014. 1, 2
[28] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using



(a) TV Monitor-HOG



(b) Car-HOG



(c) Cat-HOG

Fig. 10. Visualization of additional multi-scale HOG templates.

convolutional networks. In *ICLR*, 2014. 4

[29] D. E. C. Szegedy and A. Toshev. Deep neural networks for object detection. In *NIPS*, 2013. 4

[30] Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *PAMI*, 32(10), 2010. 1

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

[32] Y. Xiang, W. Choi, Y. Lin, and S. Savarese. Data-driven 3D voxel patterns for object category recognition. In *CVPR*, 2015. 7

[33] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *IVC*, 32(10):790–799, 2014. 2

[34] W. Zhang, G. Zelinsky, and D. Samaras. Real-time accurate object detection using multiple resolutions. In *ICCV*, 2007.