Multiresolution Models for Object Detection

Dennis Park, Deva Ramanan, and Charless Fowlkes

UC Irvine, Irvine CA 92697, USA
{iypark,dramanan,fowlkes}@ics.uci.edu

Abstract. Most current approaches to recognition aim to be scale-invariant. However, the cues available for recognizing a 300 pixel tall object are qualitatively different from those for recognizing a 3 pixel tall object. We argue that for sensors with finite resolution, one should instead use scale-variant, or multiresolution representations that adapt in complexity to the size of a putative detection window. We describe a multiresolution model that acts as a deformable part-based model when scoring large instances and a rigid template with scoring small instances. We also examine the interplay of resolution and context, and demonstrate that context is most helpful for detecting low-resolution instances when local models are limited in discriminative power. We demonstrate impressive results on the Caltech Pedestrian benchmark, which contains object instances at a wide range of scales. Whereas recent state-of-the-art methods demonstrate missed detection rates of 86%-37% at 1 false-positive-per-image, our multiresolution model reduces the rate to 29%.

1 Introduction

Objects appear at a continuous range of scales in unconstrained photographs of the world. This constitutes a significant mode of intra-class variability in detection problems. The dominant perspective in the recognition community is that one should strive for scale-invariant representations, e.g., by computing features with respect to an appropriately adapted coordinate frame, as in SIFT or scanning window detectors. While this is conceptually elegant, it ignores the fact that finite sensor resolution poses an undeniable limit to scale-invariance. Recognizing a 3-pixel tall object is fundamentally harder than recognizing a 300-pixel object or a 3000-pixel object.

This is perhaps most readily apparent in common demonstrations of the importance of context in recognition (e.g., [1]). For example, the same local patch of pixels may be identified as a car or phone depending on whether the surroundings look like a street scene or a person in an office. However, such demonstrations always involve a low-resolution, heavily-blurred image of the object in question. Given enough resolution, one *should* be able to recognize a toy-car held up to someone's ear despite the improbable context. This suggests that scene context itself should also be entered into detection in a scale-*variant* fashion with contextual cues only being used to increase the accuracy of recognizing small instances, where local image evidence is uninformative.

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part IV, LNCS 6314, pp. 241–254, 2010.

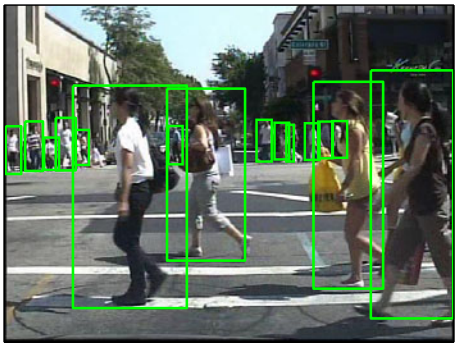

Fig. 1. An example test image in Caltech Pedestrian dataset and its ground truth annotations. The detection results of baselines and our algorithm on this image are shown in Fig 3. Note that people appear at a wide range of scales.

In this paper we propose that models for object detection should have a multiresolution structure which utilizes features ranging from detailed high-resolution parts, to whole object templates, to scene context cues. Furthermore, we treat these features in a scale dependent manner, so that high-resolution features are not used when detecting low-resolution instances.

We examine the interplay of resolution and context in the domain of pedestrian detection for autonomous vehicle navigation. Much of the recent successful work is based on template detection. We begin by asking a simple question - *what should the size of the template be*? On one hand, we want a small template that can detect small people, important for providing time for a vehicle to react. On the other hand, we want a large template than can exploit detailed features (of say, faces) to increase accuracy. Such questions are complicated by the fact a simple rigid template is not likely to accurately model both extremes, and that contextual cues should perhaps be overridden by high-confidence, large-scale detections. Using a well-known pedestrian benchmark [2], we demonstrate that contextual multiresolution models provide a significant improvement over the collective recent history of pedestrian models (as surveyed in [2]).

2 Related Work

There is storied tradition of advocating scale-invariance in visual recognition, from scale-invariant feature detectors [3,4,5] to scale-invariant object representations [6,7]. Unfortunately, such scale-invariant representations don't leverage additional pixel resolution for detecting large-scale instances.

Another family of representations deal with *multiscale* models that compute features at multiple scales. Such models are typically not multiresolution in that they do not adapt in complexity to the size of a putative detection. Examples include multiscale edge models [8] and object representations based on multi-scale wavelets [9,10]. Our approach is most similar to the multiscale part model

of [11] that defines both a low-resolution root template and high-resolution part filters. We extend the publically-available code to encode adaptive multiresolution models that act as rigid templates when scoring small-scale instances and flexible part-based models when scoring large-scale instances.

There is a quite large literature on pedestrian detection, dating back to the early scanning-window classifers of [9,12,13]. We refer the reader to the recent surveys [2,14] for an overview of contemporary approaches. Recent work has focused on models for handling pose variation [15,11,16,17,18], reducing complexity of learning [19,20]), and multicue combination [21,22]. To the best of our knowledge, there has been no past work on multiresolution representations of pedestrians.

3 Multiresolution Models

We will describe a family of multiresolution template models of increasing complexity. To establish notation, we begin with a description of a simple fixed-resolution template.

3.1 Fixed-Resolution Models

Let x denote an image window and $\Phi(x)$ denote its extracted features - say, histogram of oriented gradient (HOG) features [13]. Following an established line of work on scanning-window linear classifiers [23,13], we label x as a pedestrian if

$$f(x) > 0 \qquad \text{where} \qquad f(x) = w \cdot \Phi(x) \tag{1}$$

Such representations are trained with positive and negative examples of pedestrian windows – formally, a set of pairs (x_i, y_i) where $y_i \in \{-1, 1\}$. Popular training algorithms include SVMs [23,13] and boosting [24,25]. In our work, we will train w using a linear SVM:

$$w^* = \underset{w}{\operatorname{argmin}} \frac{1}{2} w \cdot w + C \sum_i \max(0, 1 - y_i w \cdot \Phi(x_i)) \tag{2}$$

One hidden assumption in such formalisms is that both the training and test data x_i is assumed to be scaled to a canonical size. For example, in Dalal and Triggs' [13] well-known detector, all training and test windows are scaled to be of size 128×64 pixels. The detector is used to find larger instances of pedestrians by scaling down in the image, implemented through an image pyramid. Formally speaking, the detector cannot be used to find instances smaller than 128×64. In practice, a common heuristic is to upsample smaller windows via interpolation, but this introduces artifacts which hurt performance [11,2].

In this paper, we define a feature representation $\Phi(x)$ that directly processes windows of varying size, allowing one to extract additional features (and hence build a more accurate model) when x is a large-size window.

3.2 Multiple Fixed-Resolution Models

Arguably the simplest method of dealing with windows of varying sizes is to build a separate model for each size. Assume that every window x arrives with a bit s that specifies whether it is "small" or "large". One can still write two templates as a single classifier $f(x, s) = w \cdot \Phi(x, s)$ where:

$$\Phi(x,s) = \begin{bmatrix} \phi_0(x) \\ 1 \\ 0 \\ 0 \end{bmatrix} \text{if} \quad s = 0 \qquad \text{and} \qquad \Phi(x,s) = \begin{bmatrix} 0 \\ 0 \\ \phi_1(x) \\ 1 \end{bmatrix} \text{if} \quad s = 1 \tag{3}$$

Here, $\phi_0(x)$ and $\phi_1(x)$ represent two different feature representations extracted at two different scale windows - say for example, 50-pixel and 100-pixel tall people. Given training data triples (x_i, s_i, y_i) one could learn a single w that minimizes training error in (2) where $\Phi(x_i)$ is replaced by $\Phi(x_i, s_i)$.

It is straightforward to show that (2) reduces to *independent* SVM problems given the above multiresolution feature. It is equivalent to partitioning the dataset into small and large instances and training on each independently. This poses a problem since the detector scores for small and large detections need to be comparable. For example, one might expect that small-scale instances are harder to detect, and so such scores would generally be weaker than their large-scale counterparts. Comparable scores are essential to allow for proper non-max suppression between scales, contextual reasoning [26] and for ROC benchmark evaluation.

3.3 Multiscale Multiresolution Models

One mechanism of integrating two fixed-scale models is to also compute $\phi_0(x)$ for windows with $s = 1$. In other words, we can always resize a 100-pixel windows to 50-pixels and compute the resulting small-scale feature. This allows the large-resolution model to be *multiscale* in that features are computed multiple resolutions:

$$\Phi(x,s) = \begin{bmatrix} \phi_0(x) \\ 1 \\ 0 \\ 0 \end{bmatrix} \text{if} \quad s = 0 \qquad \text{and} \qquad \Phi(x,s) = \begin{bmatrix} \phi_0(x) \\ 0 \\ \phi_1(x) \\ 1 \end{bmatrix} \text{if} \quad s = 1 \tag{4}$$

Note that because the coarse-scale features $\phi_0(x)$ are shared across both representations, the training problem no longer reduces to learning separate SVMs. In this case, distinct bias terms make scores for large and small instances comparable.

3.4 Multiresolution Part Models

One limitation of the above approach is that both small and large-scale models are encoded with a rigid template. Low-level descriptors such as HOG are

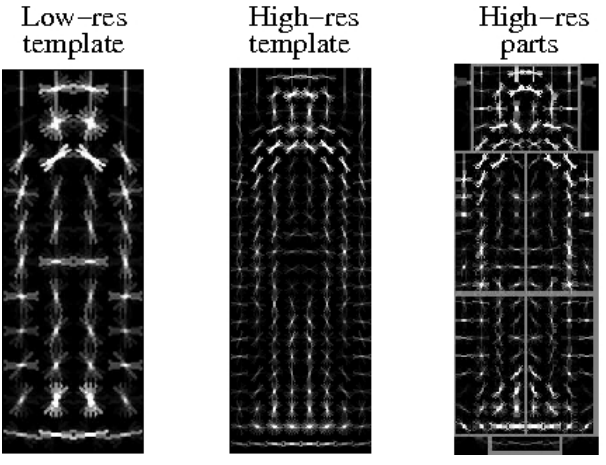

Fig. 2. Finding large-scale instances. One might use a low-resolution template (shown on the **left**). Alternatively, to exploit the extra resolution of large-scale instances, one might define a high-resolution template (**middle**). Edges capturing the boundary of the body and head are blurred out due to variation in the postures of pedestrians in the training data. A more successful approach is to explicitly model the deformation with a part model (shown on the **right**), which learns sharper part templates.

invariant to small scale image deformation due to the local spatial binning of gradient values. However, this binning occurs at a fixed-size neighborhood (in our case, a neighborhood of 4×4 pixels). On the other hand, object deformations (such as the articulation of a pedestrian) occur at a scale relative to the *size of the instance*. This means that a HOG descriptor is likely invariant to the pose deformations of a 50-pixel pedestrian, but not a 100-pixel tall pedestrian.

To model pose variations at larger scales, we augment our large-scale model with a latent parameter capturing pose variation. Following the work of [11], we add a latent parameter z that specifies the location of a collection of parts. Given the z, we define $\phi_1(x, z)$ to be a vector of vectorized-HOG features extracted at the given part locations, appended with the part offsets themselves. This allows the corresponding parameters from w to encode part templates and part deformation parameters that penalize/favor certain part deformations over others.

$$\Phi(x, s, z) = \begin{bmatrix} \phi_0(x) \\ 1 \\ 0 \\ 0 \end{bmatrix} \text{ if } \quad s = 0 \qquad \text{and} \qquad \Phi(x, s, z) = \begin{bmatrix} \phi_0(x) \\ 0 \\ \phi_1(x, z) \\ 1 \end{bmatrix} \text{ if } \quad s = 1 \quad (5)$$

The final classifier searches over latent values $f(x, s) = \max_z w \cdot \Phi(x, s, z)$:

$$f(x, s) = \begin{cases} w_0 \cdot \phi_0(x) + b_0 & \text{if} \quad s = 0 \\ w_0 \cdot \phi_o(x) + \max_z w_1 \cdot \phi_1(x, z) + b_1 & \text{if} \quad s = 1 \end{cases} \quad (6)$$

When scoring small instances, the above reduces to a standard linear template. When scoring large instances, the above requires a search over all part deformations, for the configuration that yields the maximum score. As in [11], we assume parts are independently positioned given a root location, equivalent to the standard "star" model assumptions in part-based models. This allows us to use dynamic programming to efficiently compute the max:

$$\max_z w_1 \cdot \phi_1(x, z) = \max_z \sum_j w_j \cdot \phi(x, z_j) + \sum_{j,k \in E} w_{jk} \cdot \phi(z_j, z_k) \tag{7}$$

where z_j is the location of part j, w_j is the template for part j, w_{jk} is a deformation model (spring) between part j and k, and E defines the edge structure in the star graph. We write $\phi(x, z_j)$ for the HOG feature extracted from location z_j and $\phi(z_j, z_k)$ for the squared relative offset between part j and k. Given training data triples (x_i, s_i, y_i), w can be trained with a latent SVM using the coordinate descent procedure outlined in [11] or the convex-concave procedure described in [27]. We use the publically-available coordinate descent code [28].

3.5 Latent Multiresolution Part Models

One limitation of the above model is that training data is still specified in terms of a fixed, discrete size s_i - all instances are either 50 or 100 pixels tall. Given a training window of arbitrary height x_i, one might resize it to 50 or 100 pixels by quantization. The correct quantization may be ambiguous for datasets such as PASCAL where many images of people are truncated to head and shoulder shots [29] – here a small bounding box may be better described with a truncated, high-resolution model. When the training data x_i is given as set of bounding box coordinates, [11] shows that one can significantly improve performance by estimating a latent location and scale of a "hidden" bounding box that sufficiently overlaps the given ground-truth bounding box.

We augment this procedure to also estimate the "hidden resolution" s_i of a training instance x_i. Training examples that are large will not have any low-resolution (e.g., 50-pixel tall) bounding boxes that overlap the given ground-truth coordinates. In these cases, the resolution is fixed to $s_i = 1$ and is no longer latent. Similarly, training instances that are very small will not have any high-resolution bounding boxes with sufficient overlap. However, there will be a collection of training instances of "intermediate" size that could be processed as low or high-resolution instances. The values of s_i will be treated as latent and estimated through the latent SVM framework: starting with a random initialization of latent s_i and z_i values, (1) a model/weight-vector w is trained through convex optimization, and (2) the model is used to relabel an example x_i with a latent resolution state s_i and part location z_i that produces the best score.

Relationship to mixture models: It is relevant to compare our model to the mixture models described in [23]. One might view our multiresolution model as a mixture of two models. However, there are a number of important differences from [23]. Firstly, our components share many parameters, while those in [23]

do not share any. For example, we use both low and high resolution instances to learn a low-res "root" template, while [23] only uses high-resolution instances. Secondly, the mixture component variable s_i is treated differently in our framework. At test time, this variable is *not* latent because we know the size of a putative window that is being scored. At train time, the variable is treated as latent for a subset of training instances whose resolution is ambiguous.

Extensions: Though we have described two-layer multi-resolution models, extensions to hierarchical models of three or more layers in straightforward. For example, the head part of a pedestrian may be composed of an eye, nose, and mouth parts. One would expect such a model to be even more accurate. Note that such a model is still efficient to score because the edge structure E is now a tree rather than a star model, which is still amenable to dynamic programming. Training a single resolution hierarchical part model poses a difficulty since it cannot exploit the many training and testing instances where the details, e.g., of the eyes and nose, are *not* resolvable. Our multiresolution formalism provides a framework to manage this complexity during both training and testing.

4 Multiresolution Contextual Models

We now augment our analysis of resolution to consider the effects of contextual reasoning. Our hypothesis, to be borne out by experiment, is that context plays a stronger role in detecting small-scale instances. Toward that end, we add a simple but effective contextual feature for pedestrian detection - ground plane estimation. Hoeim et. al. [1] clearly espouse the benefit of ground plane estimation for validating the observed locations and scales of putative detections. One approach would be to treat the ground plane as a latent variable to be estimated for each frame or video. We take a simpler approach and assume that the training and test data are collected in similar conditions, and so apply a ground-plane model learned from the training data at test time. We begin with the following assumptions:

1. The camera is aligned with the ground plane
2. Pedestrians have roughly the same height
3. Pedestrians are supported by a ground plane

Given the above and a standard perspective projection model, it is straightforward to show that there exists a linear relationship between the projected height of a detection (h) and the y-location of the lower edge of its bounding box in the image (y):

$$h = ay + b \tag{8}$$

Features: One reasonable contextual feature is to penalize the score of a detection in proportion to the squared deviation from the model:

$$(h - (ay + b))^2 = w_p \cdot \phi_p(x) \quad \text{where} \quad \phi_p(x) = \begin{bmatrix} h^2 & y^2 & hy & h & y & 1 \end{bmatrix}^T \tag{9}$$

where we have assumed the image features x include the location and height of the image window, and where model parameters w_p implicitly encode both the parameters of the ground plane and the amount to penalize detections which deviate from the ground plane model.

Our intuition says that low-resolution models should strongly penalize deviations because the local template will generate false positives due to its limited resolution. Alternately, the high-resolution model should not strongly penalize deviations because the local template is more accurate and the assumptions do not always hold (people are not all the same height). We investigate these possibilities experimentally using different encodings of our contextual features, including augmenting $\Phi(x, z, s)$ with a single set of perspective features $\phi_p(x)$ used across both low and high resolution models, or a separate set of features for each resolution ($\phi_p^0(x)$ and $\phi_p^1(x)$).

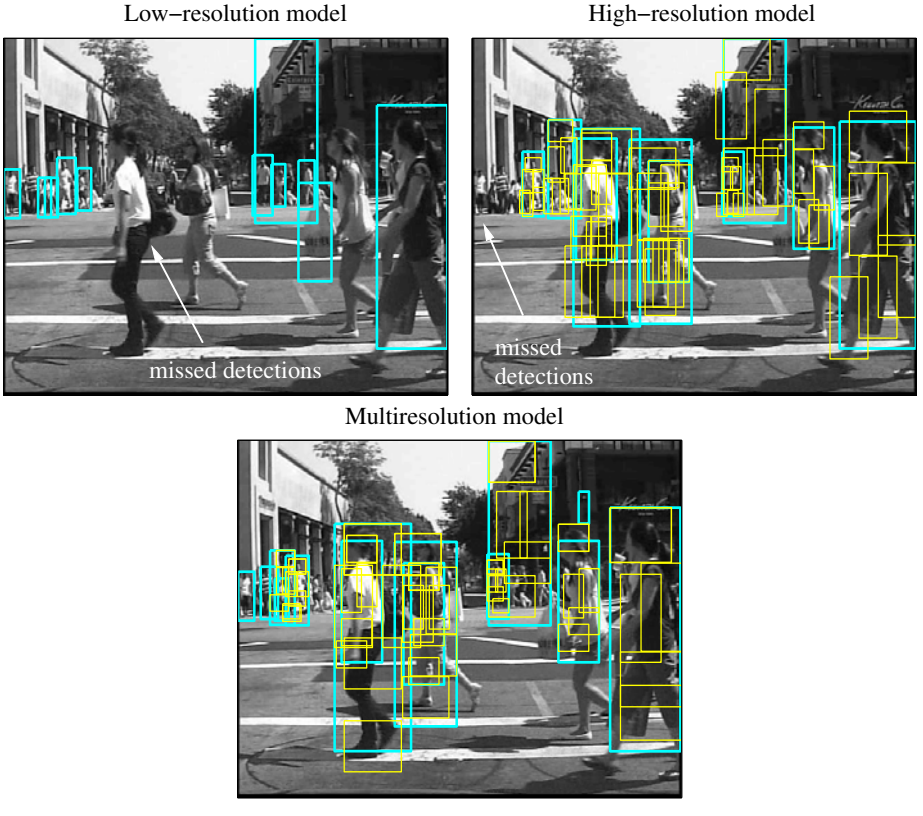

Fig. 3. On the **left**, we show the result of our low-resolution rigid-template baseline. One can see it fails to detect large instances. On the **right**, we show detections of our high-resolution, part-based baseline, which fails to find small instances. On the **bottom**, we show detections of our multiresolution model that is able to detect both large and small instances. The threshold of each model is set to yield the same rate of FPPI of 1.2.

5 Experimental Results

Implementation: We implemented our final context-augmented multiresolution model through fairly straightforward modification to the online multiscale part-based code [28]. In both the benchmark and diagnostic evaluation, we compare to the original code as a baseline. The contextual model decribed in the following results use scale-specific contextual features ($\phi_p^0(x)$ and $\phi_p^1(x)$), which we found slightly outperformed a single-scale contextual feature (though this is examined further in Sec.5.2).

5.1 Benchmark Results

We submitted our system for evaluation on the Caltech Pedestrian Benchmark [2]. The benchmark curators scored our system using a battery of 11 experiments

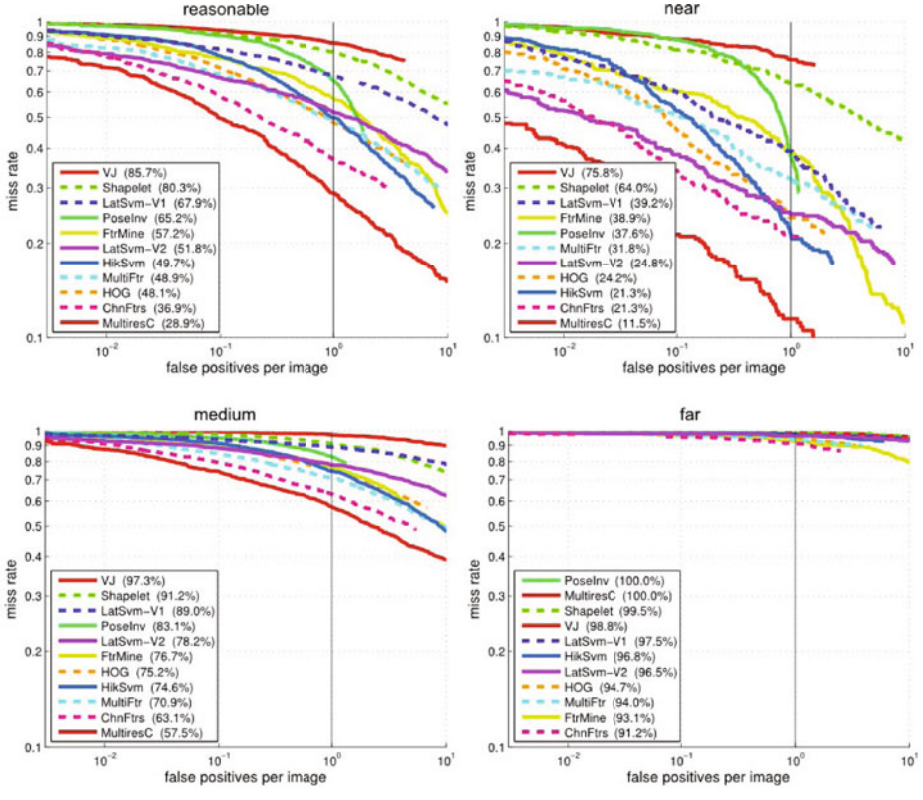

Fig. 4. Benchmark results. From the upper left graph in **clockwise** direction, we show the results for *reasonable*, *near*, *far* and *medium* experiments, evaluated on test instances with various heights ($h > 30$, $h > 80$, $h < 30$, and $30 < h < 80$ and $h < 30$, respectively). Our context-augmented multiresolution model, labeled as **MultiresC**, significantly outperforms all previous systems in 10 out of the 11 benchmark experiments (all but the 'far' experiment').

on a held-out testset, designed to analyze performance in different regimes depending on object scales, aspect ratios, and levels of occlusion (Fig. 4). The results are impressive - *our system outperforms all previously-reported methods*, across the entire range of FPPI (false positives per image) rates, in 10 out of 11 experiments. The sole experiment for which we do not win is the far-scale experiment, in which all detectors essentially fail. Even given our multiresolution model, finding extremely small objects is a fundamentally difficult problem because there is little information that can be extracted in such instances.

Our results are particularly impressive for the near-scale experiment, where we *halve* the previous-best miss rate at 1 FPPI [30]. Previous approaches, including the multiscale part-based model of [23], use fixed-resolution detectors that tend to be tuned for the small-scale regime so as to correctly fire on the set of small instances in this dataset. Our multiresolution model leverages the additional pixels available in large instances to significantly boost performance.

5.2 Diagnostic Experiments

To further analyze the performance of our system, we construct a set of diagnostic experiments by splitting up the publically-available Caltech Pedestrian training data into a disjoint set of training and validation videos. We defined this split pseudo-randomly, ensuring that similar numbers of people appeared in both sets. We compare to a high-resolution baseline (equivalent to the original part-based code [28]) and a low-resolution baseline (equivalent to a root-only model [13]), and a version of our multiresolution model without context. We

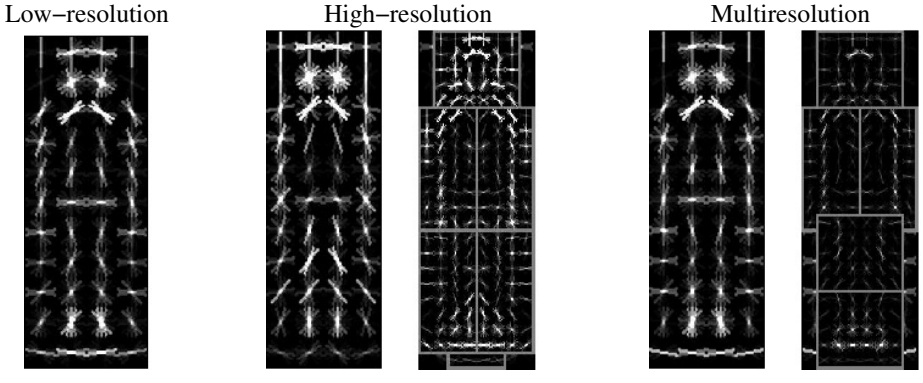

Fig. 5. On the **left**, we visualize our low-resolution rigid-template. In the **middle**, we visualize the high-resolution part-based template of [11] trained on Caltech pedestrians. Note the root templates look different, as only a small portion of the training data (of high enough resolution) is used to train the part-model. On the **right**, we visualize the multiresolution model. Note that the root component looks similar to the low-resolution model. Also note that the parts overall have weaker weights. This suggests that much of the overall score of the multiresolution model is given by the root score. However, it is still able to detect both small and large instances as shown in our results.

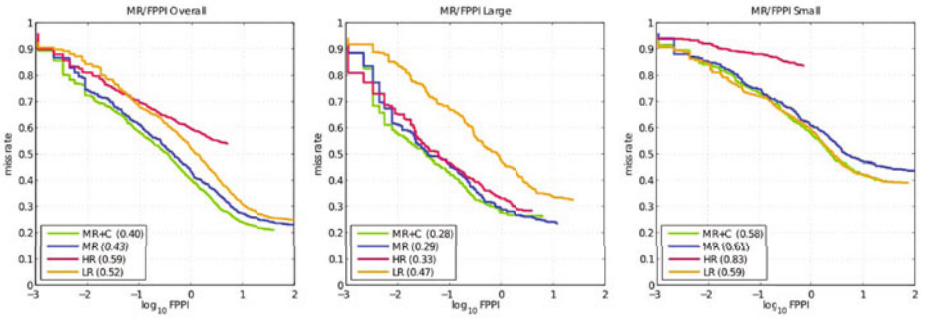

Fig. 6. Results of diagnostic experiments. We compare results to fixed resolution baselines, where "LR" is a low-resolution rigid template and "HR" is a high-resolution part-based model. On the **left**, we show results evaluated on the full set of test instances from validation data. In the **middle**, we show results for large-instance (> 90 pixels). On the **right**, we show the results on small-instances (< 90 pixels). The "LR" template performs well on small instances, while the "HR" template performs well on large instances. Our multiresolution "MR" model exploits the best of both, in the appropriate regimes. Our context-augmented model "MR+C" provides a small improvement overall, but a noticable improvement when detecting small instances at a higher FPPI rate.

visualize our baseline models in Fig. 5. All methods are trained and evaluated on the exact same data. To better interpret results, we threw out instances that were very small (< 30 pixels in height) or abnormal in aspect ratio (i.e. $h/w > 5$), as we view the latter as an artifact of annotating video by interpolation.

Overall: Overall, our multiresolution model outperforms baseline models. Our contextual model provides a small but noticeable improvement, reducing the missed detection rate from 43% to 40%. We shall see that the majority of this improvement comes from detecting small-scale instances. Somewhat surprisingly, we see that a simple rigid template outperform a more sophisticated part model - 52% MD compared to 59%. One can attribute this to the fact that the part-based model has a fixed resolution of 88 pixels (selected through cross-validation), and so cannot detect any instances which are smaller. This significantly hurts performance as more than 80% of instances fall in this small category. However, one may suspect that the part-model should perform better when evaluating results on test instances that are 88 pixels or taller.

Detecting large instances: When evaluating on large instances (> 90 pixels in height), our multiresolution model performs similarly to the high-resolution part-based model. Both of these models provide a stark improvement over a low-resolution rigid template. We also see that perspective context provides no observable improvement. One might argue that this is due to a weak contextual feature, but we next show that it does provide a strong improvement for small scale detections.

Detecting small instances: When evaluating on small instances (< 90 pixels in height), we see that the part-based model performs quite poorly, as it is

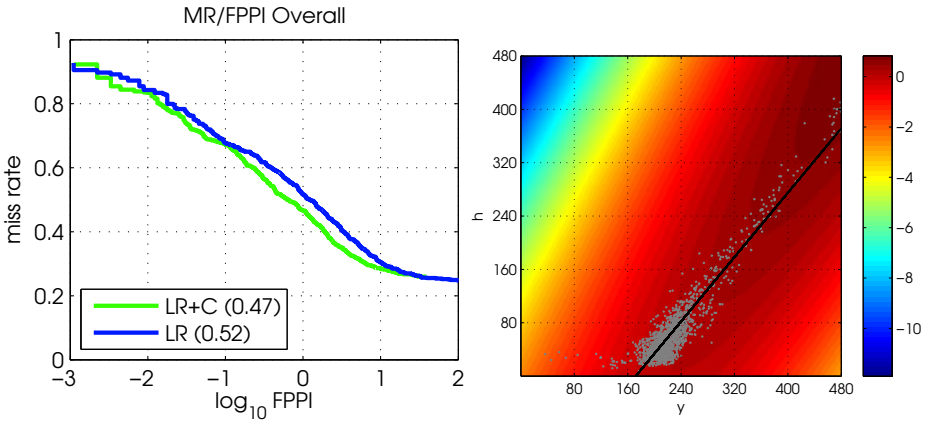

Fig. 7. We show the effectiveness of our perspective features on low-resolution models. Overall performance increases from 51% MD to 46% MD. We visualize our perspective features on the **right**. We plot the distribution of h and y (bounding box height and image-y locations) in the ground truth data, and plot the score $w_p \cdot \phi_p(x)$ as a function h and y. We also display the distribution of ground truth (visualized with a point cloud) along with its linear fit. We see that the learned contextual features penalize detections whose heights and image-y locations are not consistent with the ground plane.

unable to detect the majority of test instances which are small. Our multiresolution model performs slightly worse than a low-resolution model (61% compared to 59%). Perspective features provide a noticeable improvement for our multiresolution model, increasing performance from 61% MD to 58%.

Context features: To verify that our contextual features are indeed reasonable, we analyze the benefit of our contextual features on a low-resolution model. We see a noticeable reduction in the MD rate from 51% to 46%, suggesting our contextual features are indeed fairly effective. Their effect is diminished in our multiresolution model because the part-based model is able to better score large-scale instances, reducing the need for score adjustment using context.

6 Conclusion

We describe a simple but effective framework for merging different object representations, tuned for different scale-regimes, into a single coherent multiresolution model. Our model exploits the intuition that large instances should be easier to score, implying that one should adapt representations at the instance-level. We also demonstrate that context should be similarly adapted at the instance-level. Smaller objects are more difficult to recognize, and it is under this regime that one should expect to see the largest gains from contextual reasoning. We demonstrate impressive results on the difficult but practical problem of finding large and small pedestrians from a moving vehicle.

Acknowledgements

Funding for this research was provided by NSF grants 0954083 and 0812428, and a UC Labs research program grant.

References

1. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. International Journal of Computer Vision 80(1), 3–15 (2008)
2. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (2009)
3. Lindeberg, T.: Scale-space theory in computer vision. Springer, Heidelberg (1994)
4. Lowe, D.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60, 91–110 (2004)
5. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International Journal of Computer Vision 60, 63–86 (2004)
6. Fergus, R., Perona, P., Zisserman, A., et al: Object class recognition by unsupervised scale-invariant learning. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2 (2003) Citeseer
7. Dorko, G., Schmid, C.: Selection of scale-invariant parts for object class recognition. In: ICCV 2003 (2003) Citeseer
8. Mallat, S., Zhong, S.: Characterization of signals from multiscale edges. IEEE Transactions on pattern analysis and machine intelligence 14, 710–732 (1992)
9. Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., Poggio, T.: Pedestrian detection using wavelet templates. In: IEEE CVPR, pp. 193–199 (1997)
10. Schneiderman, H., Kanade, T.: A statistical method for 3D object detection applied to faces and cars. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, IEEE Computer Society, Los Alamitos (1999/2000)
11. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Computer Vision and Pattern Recognition, Anchorage, USA (June 2008)
12. Gavrila, D.: Pedestrian detection from a moving vehicle. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 37–49. Springer, Heidelberg (2000)
13. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. I886–I893 (2005)
14. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: Survey and experiments. IEEE PAMI 31, 2179–2195 (2009)
15. Mohan, A., Papageorgiou, C., Poggio, T.: Example-based object detection in images by components. IEEE PAMI 23, 349 (2001)
16. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemannian manifolds. IEEE PAMI 30, 1713–1727 (2008)
17. Wang, X., Han, T.X., Yan, S.: An HOG-LBP human detector with partial occlusion handling. International Journal of Computer Vision (2009)
18. Lin, Z., Hua, G., Davis, L.S.: Multiple instance feature for robust part-based object detection. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 405–412 (2009)
19. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel SVMs is efficient. In: IEEE Conf. on Computer Vision and Pattern Recognition (2008)

20. Schwartz, W., Kembhavi, A., Harwood, D., Davis, L.: Human detection using partial least squares analysis. International Journal of Computer Vision (2009)
21. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: IEEE Conf. on Computer Vision and Pattern Recognition (2009)
22. Gavrila, D.M., Munder, S.: Multi-cue pedestrian detection and tracking from a moving vehicle. International journal of computer vision 73, 41–59 (2007)
23. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE PAMI 99(5555)
24. Dollar, P., Babenko, B., Belongie, S., Perona, P., Tu, Z.: Multiple component learning for object detection. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 211–224. Springer, Heidelberg (2008)
25. Sabzmeydani, P., Mori, G.: Detecting pedestrians by learning shapelet features. In: Proc. CVPR, pp. 1–8 (2007)
26. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models of multi-class object layout. In: ICCV (2009)
27. Yu, C., Joachims, T.: Learning structural SVMs with latent variables. In: Proceedings of the 26th Annual International Conference on Machine Learning. ACM, New York (2009)
28. `http://people.cs.uchicago.edu/~pff/latent`
29. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results (2007), `http://www.pascal-network.org/challenges/VOC/voc2007/workshop`
30. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: BMVC (2009)